The role and prosodic characteristics of hesitation lengthening and filled pauses in speech planning in Maltese
Sarah Agius, Flavia Chetcuti, Alexandra Vella
University of Malta
agius.sarah@gmail.com, fchetcuti@hotmail.com, alexandra.vella@um.edu.mt

Abstract
This paper examines two phenomena which occur in speech, rather than during periods of silence: hesitation lengthening of lexical elements, and filled pauses involving non-lexical elements. An interesting characteristic of hesitation lengthening of lexical elements is that there do not seem to be any constraints on what element within the segmental and syllabic stream gets lengthened: although there is a tendency for the final (often unstressed) syllable to be lengthened, it seems possible for lengthening to affect any element within the segmental stream. The two phenomena share a number of features, the most prominent of which are marked lengthening and level F0. More significantly, both phenomena have a clear turn-holding effect: change of speaker rarely occurs when hesitation lengthening takes place. What is also interesting is that filled pauses come in two forms, one of which exhibits some sort of heightening of the usual characteristics for this sort of element. The discourse function of turn-holding which occurs in all cases of hesitation lengthening on lexical elements occurs also in the case of filled pauses but only when these also involve lengthening.

Keywords: hesitation lengthening, filled pauses, prosodic characteristics, turn-taking and turn-holding

1. Introduction
Spontaneous speech, unlike read speech, is marked by a variety of features, collectively often referred to as “normal disfluencies” (Shriberg 1994). Dialogue data is no different to spontaneous speech data of other sorts. Speakers in a dialogue need to keep the information flow going. They use a variety of strategies to keep the discourse moving forward and in doing so, may need to negotiate the floor. As part of this negotiation, it may sometimes be necessary for the current speaker to signal to her/his interlocutor that s/he is still thinking about what, and how best, to say what s/he has to say, and that therefore s/he is not yet ready to relinquish the floor.

These “talk-in-interaction” type requirements have some interesting effects on elements which have been observed to be found in dialogue data. Analysis of new spontaneous (but significantly non-Map Task) data from Maltese has brought to light a feature which appears to occur relatively frequently in these data. This involves lengthening which might not otherwise occur at a “natural” prosodic (or other type of) boundary and/or whose role is not one involving either simple demarcation of intonation phrases or straightforward hesitation. Such lengthening has been noted to be a distinct phenomenon which plays a role in planning similar to that noted to be involved in the use of both filled pauses (FPs) and, to a lesser extent, also of unfilled pauses in different languages (e.g. Cutler and Pearson 1986, Grosz and Hirschberg 1992, Swerts 1998, Clark and Fox Tree 2002, Campione and Véronis 2005), Maltese included (Vella et al., 2011 and 2014).

This study examines the role, as well as the phonetic, particularly prosodic, characteristics of hesitation lengthening, in two each of the Map and Conversation Task data available as part of the MalToBI corpus (Vella and Farrugia 2006) of spoken Maltese. It does so in the first instance by comparing hesitation lengthening to a phenomenon we believe to be closely related to it, that involving FPs. The study seeks to establish whether the functional and prosodic characteristics associated with these two – on the surface distinct – phenomena, are in fact one and the same thing.

The questions we examine are: What are the functional and prosodic characteristics of hesitation lengthening? And are the defining characteristics of hesitation lengthening the same or different to those of FPs?

2. Silence and speech
The stream of speech is broken up by silent intervals of different sorts. Silent intervals, or unfilled pauses, can perform a variety of functions in speech. Couper-Kuhlen (1986:75) provides a succinct summary of this variety of functions as resulting from “a performance-related origin – a pause for breath, a pause to search for a word or to plan”. As hinted at earlier, the “pause to search for a word or to plan” is rendered that much trickier in conversation. If the interlocutor does not recognize the silence for what it is, there is the risk that s/he, mistakenly assuming that the floor is being relinquished, takes the floor. So what strategies are available to speakers engaged in conversation to minimize the chances of a breakdown in appropriate turn-taking?

To start with, an attempt is made to map out some of the distinctions we make in this study, in the light of the literature on this topic. Figure 1 below provides a schematisation.

Figure 1: Schematisation of distinctions

One distinction which is often made is that between unfilled and filled pauses (Cruttenden 1997). However, such a distinction is far too simplistic. There are a number of ways in which speakers can slow things down, and these different
means can be used separately, or they can interact in interesting ways. The details of this interaction will be the subject of the analysis of data carried out and reported in Sections 4 and 5. Each distinction shown in the top three layers of the schematization is described in Figure 1 above.

Silence, when it occurs in speech often involves unfilled pauses. Pauses or breaks can occur for purely articulatory reasons such as when a speaker stops to breathe. They can also occur for junctural reasons involving a speaker organising text into phonological phrases of different types. One side-effect of unfilled pauses is that the stretch of speech immediately preceding such a pause is subject to the phenomenon often referred to as pre-boundary lengthening regardless of whether hesitation is involved. Unfilled pauses can however also arise in contexts of hesitation.

Speech can contain lexical as well as non-lexical elements. Lexical elements, although not normally associated with the phenomena of interest here (labelled as “default” in Figure 1) are interesting in the context of this study in their manifestation as elements involving “hesitation lengthening” (see also Figure 1). Amongst the non-lexical elements of interest to us are the so-called FPs, examples of which include [e] in Scottish English and [n] in Russian (Cruttenden 1997:174) and at least [e], [m], [em] and [ʔ] in Maltese (Vella et al. 2011 and personal communication). Other non-lexical elements also occur in speech. These include backchannels such as ehe and mham (orthographic uh and um for English), and vocalisations such as tr (orthographic tsk in English).

3. Data and methodology

The data analyzed forms part of the MalToBl corpus consisting of 8 Map Tasks and 8 Conversation Tasks (see Vella and Farrugia 2006).

3.1 The Tasks

The Map Task involves quasi-naturalistic conversational data. The dialogue generated is oriented towards task completion, something which is not necessarily a prerequisite of spontaneous speech. The Maltese Map Task was designed with a view to collecting data involving the use of specific target items, items having different syllable structures and accent placement (final, penultimate or antepenultimate) but composed of all-sonorant material to better facilitate analysis in terms of F0.

The Conversation Task by contrast provided speakers with a specific scenario which they had to use as the basis for the conversation with their interlocutor. One of the speakers was asked to pretend that s/he knew someone who was seeking to fill a vacant post in the company s/he worked for. The task involved this speaker talking to the other speaker and trying to gather information which could be relayed on, regarding the suitability or otherwise of this person as a possible candidate for the job. Speakers changed roles when they felt they had no more to say.

3.2 Material used, and adaptations to annotation

Two of the orthographically annotated files from each task were used in this study. All data had been annotated using the guidelines and conventions developed in the course of the projects SPAN and ISMA post-SPAN (Vella et al. 2010). The annotation of each file was checked and revised by a second transcriber. The analysis of the Conversation Tasks brought to light a feature which was relatively less present, and which had somehow not been noted when the annotation of the Map Tasks first took place. A specific annotation marker (+) was introduced within the orthographic tier. The SPAN guidelines were updated accordingly. The (+) is placed after the lengthened element (initially the lengthened element was assumed to be the syllable, but it was later observed to be possible for practically any segment within an element to be lengthened).

4. Hesitation lengthening

Our working definition of hesitation lengthening is the lengthening of segmental material in a lexical element for reasons other than those normally associated with a prosodic boundary of some sort. In what follows we look more closely at hesitation lengthening of lexical material which has undergone what transcribers intuitively marked as hesitation lengthening. Any differences in the occurrence of this phenomenon as a function of task type will not be considered in the analysis which follows.

4.1 Hesitation lengthening and preceding or following silence

The data analysed shows that hesitation lengthening can occur with silence preceding and/or following it, as well as with speech segments on both sides of it. Moreover, this phenomenon rarely occurs with silence on both sides of it.

There was silence on both sides of a lengthened element in only 16% of all the instances of lengthening in the data analysed. Lengthened elements mid-speech, i.e. with speech on both sides of the lengthened element, account for 27% of the instances analysed. Silence occurred on either the left or the right of a lengthened element in 57% of the instances analysed, although silence to the right of this phenomenon is more frequent, as compared to silence to the left.

4.2 Hesitation lengthening and change of speaker

Of the instances of hesitation lengthening in the data analysed, only 10% were followed by a change of speaker. In the overwhelming majority of cases, 90%, the speaker employing the lengthening continued speaking after the particular instance of lengthening. This clearly suggests that this type of lengthening serves as a cue to the listener that the speaker intends to hold her/his turn. The use of hesitation lengthening appears to be a clear signal to the interlocutor that the speaker needs to buy time for some difficult retrieval process or simply to process and formulate her/his thoughts.

4.3 Hesitation lengthening and F0

The data analysed shows that the lengthening phenomenon is linked to a level F0 which stays steady in a sort of “hold” for a while. The F0 before and after the hold seems to flow, in a way, as though there had been no interruption.

Figure 2: Excerpt (436.5 to 438.9s) from MG_SP_C1

Figure 2 above shows three instances of lengthening, one in which lengthening occurs at the beginning of a word and two in which lengthening occurs at the end of a word. The former
case, of lengthening on n+igu ‘we come’, occurs in the middle of a stretch of speech with no surrounding silence and illustrates how F0 levels out on the segment which is being lengthened (in this case, the initial [n]). In the latter two instances, where lengthening occurs at the end of the word, one (hdejn+ ‘near’) occurs in a stretch of speech and one (Żona+ ‘zone/area’) occurs preceding a break. What is interesting is that, in the case of hdejn+ ‘near’, where there is no surrounding silence on either side, F0 is level not only on the segment which is lengthened, but throughout the word, suggesting that here, it is this level F0 on its own that is serving as a turn-holding cue. This is different to what happens in the case of Żona+ where the lengthening is followed by a break and where the pitch contour flattens out towards the end of the word and not throughout the word. Furthermore, the last of these instances in particular also illustrates how F0 is reset after each instance of lengthening, i.e. at the beginning of Żona+.

Figure 3: Excerpt (166.5 to 168.4s) from MG_SP_C1

Figure 3 illustrates two examples of lengthening of the lexical item ghal ‘towards’, namely ghal+ ‘towards’ and ghall-genb ‘towards the side’ (the apparent difference is purely orthographic), which both have silence in their surroundings. The first example has silence on both sides and interestingly F0 is level throughout. This parallels what we saw in Figure 2 above, where lengthening on a word, albeit one surrounded by speech rather than by silence, (in the above-mentioned case, hdejn+), had level F0 throughout and not simply on the lengthened segment. The second instance of lengthening illustrated in Figure 3 also shows level F0 on the lengthened segment and a clear reset of the pitch by the speaker to continue speaking. These examples thus show that this lengthening does not disturb the flow of speech at all and seems to serve a turn-holding function.

4.4 Hesitation lengthening and location in the segmental stretch

The phenomenon in question has been found to occur at a variety of locations in the segmental stretch, both on “complete” stretches as well as on incomplete, abandoned ones of different types such as ghaim+... and m+. The lengthening of stretches of speech without disfluencies in them involves instances of lengthening at the beginning of a word, in the middle of a word, as well as at the end of the word. It is to be noted that “word” refers to some kind of phonological word. Examples of this include the following: L+ej[ε], Lejn+n and Lejn+ which one particular speaker uses in a specific conversation task analysed, together with other non-lengthened instances of the same lexical item, lejn ‘towards’. This example illustrates perfectly the flexibility of this phenomenon, occurring as it does at all possible locations of a particular lexical item. Moreover, the phenomenon does not seem to be constrained by the phonological form of the segment in question. Thus, for example, in the case of word-initial lengthening, the analysis showed instances of lengthening at the beginning of a word where the word started with a sequence of a vowel, or a liquid, or a nasal, but also in stretches starting with obstructors or obstructant sequences. One example is lengthening in S+gaaq ‘alley’. Other examples are the lengthening at the beginning of f-ilghodu ‘morning’ and h+oteq ‘hotel’.

In the case of lengthening in the middle of a word, once again, the phenomenon seems to occur irrespective of syllabic length or phonological constraints. Thus there were instances of lengthening in the middle of monosyllabic words, like the above-mentioned example of lejn+n ‘towards’ and xogho+l ‘work’. Other examples included lengthening of the initial, penultimate and ultimate syllables in bisyllabic and polysyllabic words, such as ghall+-genb ‘towards the side’, mill+-Bajja ‘from the bay’, il+-lemi ‘the right-hand side’, l-imna+rija – name of a specific Maltese feast and /Institu+te/ ‘institute’. Although the phenomenon seems to occur more frequently on non-stressed syllables, it also can occur on stressed syllables as illustrated by the above-mentioned example of l-imna+rija – name of a specific Maltese feast. (The stressed syllable in the preceding examples is shown in **bold**.)

To sum up, word-final, post-stress location seems to be the preferred location for this phenomenon to occur, at least in the data analysed. It is clear however that hesitation lengthening in this position is not constrained by the phonetic nature of the lengthened element itself or by the nature of the segmental material in its immediate surroundings. A few examples of the many instances of lengthening noted in this context are the following: lejn+ ‘towards’, cirku+ ‘circle’, nibew+ ‘we start’, il-Mara+ ‘the wife’ and c-certifikati+ ‘the certificates’.

5. Filled pauses revisited

Vella et al. (2011) carried out an analysis of all the FPs in the Map Task data. Although no basis for a clear durational distinction, for example, for forms labelled as m, mm and mmm was found, it was nevertheless observed that there were cases where FPs themselves appeared to be “enhanced” by means of something which could well be hesitation lengthening. Comparison of the phenomenon of hesitation lengthening with the hesitation lengthening of FPs seems worth carrying out at this point.

Vella et al. (2011) have shown that FPs, by their very nature, vary in their duration. Therefore, in order to account for the observed phenomenon of lengthening while keeping in mind this inherent variability in duration, a decision was taken to mark FPs, which seemed to be proportionately longer than the “default” equivalent filled pause, using the same annotation marker (+) of lexical elements used in the Speaker tiers.

Analysis of these FPs showed that a distinction between lengthened and non-lengthened FPs is actually made by speakers, with lengthened FPs having an average duration of 0.50 seconds and non-lengthened ones having an average of 0.22 seconds in the data analysed. Furthermore, not only can one make this distinction for the FPs analysed in the data, but analysis showed that a small majority of them (67 out of a total of 121, or 55% of all FPs analysed) are actually lengthened.

Figure 4 below illustrates the difference in duration of lengthened (the first ee) as compared to non-lengthened (the second ee) FPs.
5.1 FPs and preceding or following silence

Both lengthened and non-lengthened FPs seem to occur with silence in their surroundings (Vella et al. 2011). However, it seems that lengthened FPs tend to occur with silence on both sides, not just preceding or following them. This might be due to the nature of the lengthening in itself, which seems to favour contexts of following silence (cf. Section 4.1 above). Out of the lengthened FPs, only 7% are found in a stretch of speech with no silence in their vicinity. However, in the case of non-lengthened FPs, a much higher proportion, 22%, are found with no break or pause in their vicinity. This strengthens the argument that non-lengthened FPs, unlike their lengthened counterparts, do not serve the function of turn-holding.

5.2 FPs and discourse function

FPs serve a turn-taking as well as a turn-holding function. This is illustrated in Figure 5 below, which includes two instances of non-lengthened FPs. As we can see from this Figure, Speaker 2 uses the first FP, ee (which is 0.20 seconds long) to take the turn from Speaker 1. It is clear here that the speaker, although taking the turn, is still planning her utterance, and the use of the FP actually shows this. The same speaker then uses the same filled pause ee, a second time, with a shorter duration (0.10 seconds) to show that she is still planning her utterance.

6. Conclusion

The data we report on reveal that hesitation lengthening and (both types of) FPs are different in some respects. First, the former tends to affect the final, rather than the stressed syllable, suggesting that this type of lengthening is an edge phenomenon of some sort; this is clearly not the case for FP because of their very structure (monosyllabic elements). Secondly, hesitation lengthening can, but needn’t be followed, and is not usually preceded, by pause; FPs, by contrast, usually occur with a pause to their left or to their right or on both sides (Vella et al. 2011).

Nevertheless, hesitation lengthening and (lengthened) FPs are similar in that they enable the speaker to pause without relinquishing the floor, and both are possible at a wide range of locations, not always ones at which one would expect a boundary to be placed. All of these phenomena are characterised by level F0. In the case of FPs this is always present. In the case of hesitation lengthening on lexical items, this is always present on the lengthened element and could also stretch out over the surrounding elements.

Whereas lengthened FPs can be treated as having the same function as hesitation lengthening, non-lengthened FPs might serve a different function, and do not necessarily serve to hold the turn in the same way as hesitation lengthening does. This distinction between long FPs and short FPs however needs to be further analysed in order for any conclusions to be reached.

Given that it has been noted that FPs, but more especially, hesitation lengthening, occur more frequently in the Conversation Task-type data, it would be interesting to investigate this issue further.

7. Acknowledgements

We would like to thank the University of Malta for funding SPAN and ISMA post-SSPAN.

8. References


A Multi-Sensor Helmet to Capture Rare Singing, an Intangible Cultural Heritage Study

S. K. Al Kork¹,², A. Jaumard-Hakoun¹,², M. Adda-Decker³, A. Amelot³, L. Buchman⁴, P. Chawah³, G. Dreyfus², T. Fux³, C. Pillot-Loiseau³, P. Rousse⁵, M. Stone⁴, K. Xu¹,², B. Denby¹,²

¹Université Pierre Marie Curie, Paris, France,
²Signal Processing and Machine Learning Lab, ESPCI Paris-Tech, Paris, France,
³Phonetics and Phonology Laboratory, LPP-CNRS, UMR7018, University Paris3 Sorbonne Nouvelle
⁴Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, USA
⁵Aalto University, Espoo, Finland

Abstract

A portable helmet based system has been developed to capture motor behavior during singing and other oral-motor functions in a non-laboratory experimental environment. The system, based on vocal tract sensing methods developed for speech production and recognition, consists of a lightweight “hyper-helmet” containing an ultrasonic (US) transducer to capture tongue movement, a video camera for the lips, and a microphone, coupled with a further sensor suite including an electroglottograph (EGG), nose-mounted accelerometer, and respiration sensor. The system has been tested on two rare, endangered singing musical styles, Corsican “Cantu in Paghjella”, and Byzantine hymns from Mount Athos, Greece. The versatility of the approach is furthermore demonstrated by capturing a contemporary singing style known as “Human Beat Box.”

Keywords: portable speech collection system, ultrasound, EGG, accelerometer, data acquisition, intangible cultural heritage, i-Treasures project.

1. Introduction

A major objective of the i-Treasures project is to provide students with innovative multi-media feedback to train specific articulatory strategies for different type of rare singing, considered an endangered Intangible Cultural Heritage. To this end, i-Treasures will carry out vocal tract capture during rare singing performances to enable study of production mechanisms, and to define reliable features for a subsequent animation of these articulatory movements for use in educational scenarios and automatic classification tasks. To accomplish this, it is necessary to build a system that can record the configuration of the vocal tract – including tongue lips, vocal folds and soft palate – in real time, and with sufficient accuracy to establish a link between image features and actual, physiological elements of the vocal tract.

Ultrasound, US, is a popular non-invasive technique for real time imaging of the vocal tract. Examples of portable devices that acquire US images of the tongue and video of the lips, for applications in speech synthesis and silent speech interfaces (Denby and Stone 2004) (Cai, et al. 2011), have been described in the literature. Ultrasound also requires no external magnetic field, and can thus also be readily complemented with other sensors. Here, we present a system based on a helmet containing a US probe, lip camera, and microphone, coupled with a suite of other sensors including electroglottograph (EGG), to measure and record vocal fold contact movement during speech; a piezoelectric accelerometer, for detecting the nasal resonance of speech sounds (Stevens, Kalikow and Willemain 1990); and a respiration sensor belt to determine breathing modalities (Tsui and Hsiao 2013). The proposed system is advantageous in that 1) it is lighter and easier to wear for long periods than other solutions proposed in the literature (Wrench, Scobie and Linden 2007); and 2) the combination with other sensors has the potential to greatly enhance our knowledge of rare singing techniques, and allow the extraction of sensorimotor features in order to drive a 3D avatar for learning scenarios.

2. Methods

Figure 1 presents a schematic overview of the modules contained in the capture system. Each sensor first requires specific gain tuning and/or zero calibration protocols, before the streams are simultaneously and synchronously recorded with the RTMaps toolkit, which will be described in section 2.2.

2.1. Helmet design and sensor setup

The helmet allows simultaneous collection of vocal tract and audio signals. As shown in Figure 2, it includes an adjustable platform to hold the US probe in contact with the skin beneath the chin. The probe used is a microconvex 128 element model with handle removed to reduce its size and weight, which captures a 140° image to allow full visualization of tongue movement. The US machine chosen is the Terason T3000, a system which is lightweight and portable yet retains high image quality, and allows data to be directly exported to a PC via Firewire. A video camera (from The Imaging Source) is positioned facing the lips (Figure 2). Since differences in background lighting can affect computer recognition of lip motion, the camera is equipped with a visible-blocking filter and infrared LED ring, as is frequently done for lip image analysis. Finally, a commercial lapel microphone (Audio-Technica Pro 70) is also affixed to the helmet to record sound.

The three non-helmet sensors are directly attached to the body of the singer as indicated in Figure 3. An accelerometer attached with adhesive tape to the nasal bridge of the singer captures nasal bone vibration related to nasal tract airway resistance, which is indicative of nasal resonance during vocal production. Nasality is an important acoustic feature in voice timbre. An EGG (Model EG2-PCX2, Glottal Enterprises Inc.) is strapped to the singer’s neck to record a time dependent signal whose peaks are reliable indicators of glottal opening and closing instances (Henrich, et al. 2004). Finally, on the singer’s chest, a
respiration sensor or “breathing belt” is affixed to measure breathing modalities during singing.

**Figure 1 Overview of the vocal tract capture system for the rare singing sub-use cases. Six sensors (left), capture motion of five physical events: (1) tongue, (2) lips, (3) the acoustic speech wave, (4) nasal resonance, (5) vocal folds, and (6) respiratory muscle. Each instrument is processed (middle) and digitally recorded (right).**

### 2.2. Data Acquisition and system design

#### 2.2.1. System architecture

The data acquisition system must be able to synchronously record US and video data at sufficiently high frame rates to correctly characterize the movements of the tongue and lips, as well as the acoustic speech signals, the EGG, the accelerometer and the respiratory waveforms. The acquisition platform was developed using the Real-Time, Multi-sensor, Advanced Prototyping Software (RTMaps®, Intempora Inc, Paris FR).  

#### 2.2.2. RTMaps real time user interface

The data acquisition platform has the ability both to record and display data in real time, and the acquired data can be stored locally or transferred over a network. Figure 4 displays a screen shot from the platform. Ultrasound and video images are streamed at a rate of 60 frames per second, then stored in either .bmp or jpeg format. Image size for US and camera are 320 by 240 pixels and 640 by 480 pixels respectively. The EGG, the microphone, the piezoelectric accelerometer and respiration belt are interfaced to a four-input USB sound card (AudiBox44VSL) whose output interfaces to the acquisition system. These four analog input signals are sampled at 44100 Hz with a 16 bit encoding. The sampled analog signals are saved to a .wav format.

**Figure 2 Multi-sensor Hyper-Helmet: 1) Adjustable headband, 2) Probe height adjustment strut, 3) Adjustable US probe platform, 4) Lip camera with proximity and orientation adjustment, 5) Microphone.**

**Figure 3 Schematic of the placement of non-helmet sensors, including the (1) accelerometer, (2) EGG, and (3) respiration belt**

#### 2.2.3. Definition of recording material

In order to study a variety of singing techniques, and to extract features for automatic classification and pedagogical scenarios, we will collect material of varying degrees of complexity: isolated vowels (/i/, /u/, /e/, /o/, /a/), CV syllables (/papapapa/, /tatatatata/, /kakakakaka/...), sung phrases and entire pieces, where the material is to be produced both in spoken and singing modes. Byzantine chant, Corsican Paghjella, and the contemporary singing style known as “Human Beat Box”, HBB, have been chosen for study. For Byzantine chant, different styles (Mount Athos vs Ecumenical Patriarchate of Constantinople styles for example) have been selected. For Corsican Paghjella, we propose to study versus (melodies) from three different regions known for their traditional singing styles: Rusio, Sermanu and Tagliu -Isolacciu. The protocol to be used for Sardinian Canto a Tenore is still under discussion. For HBB,
basic material will be recorded as defined in (Proctor, et al. 2013), as well as short HBB phrases and longer performances in different styles, with details still to be defined.

Figure 4 RTMaps Data Acquisition User Interface, showing simultaneous recording and visualization of ultrasound tongue images, lip video, and the four analog sensors.

3. Preliminary experimental results

3.1. Sensor synchronization

Sensor synchronization is crucial in our study since extracted sensor features are ultimately to be used to drive a 3D tongue avatar. To check synchrony, an event common to all sensor channels is created using the procedure illustrated in Figure 5. A syringe containing ultrasound gel is struck by a spring-loaded weight, ejecting a gel droplet that strikes the head of the ultrasound probe and shorts together the two sides of the EGG sensor. Vibration induced in the syringe body is detected in the nasality accelerometer, and the acoustic signal of the weight hitting the syringe is recorded by the microphone. Finally, the video camera normally used for the lips is positioned so that it can also capture the droplet as it is ejected. The resulting signals obtained from the 5 sensors are shown in Figure 6.

We have calculated the timestamps at which the gel droplet occurred in all sensors. The time stamp at which the droplet was triggered for the microphone, piezo and EGG was 4.060s, 4.070s and 4.070s respectively. The droplet was captured by the camera and Ultrasound at image number 244 in sequence and 4.066s and 4.070s respectively. The time stamp at which the droplet was obtained from the 5 sensors can be extracted using the procedure THISSP·ISSP (Figure 7a). A specific problem for HBB is the difficulty of stabilizing of the ultrasound probe in view of the large range delays in sensor setup and ease of use. Figure 6 illustrates the versatility of the helmet, which can be used with any head size and shape and does not impede the singing function.

The voice capture system has been tested with one expert singer each for the Human Beat Box, HBB (Da Vox) (Figure 7a), the Paghjella (B. Sarocchi) Figure 7b, and the Byzantine (D. Manousis) (Figure 7c) musical styles. Each singer participated in a recording session to validate the helmet with respect to his or her style, and to assist us in specifying an appropriate data collection protocol. A recording session consists of three phases: 1) singer preparation (wearing of the helmet and body sensors), 2) sensor calibration and, 3) data collection proper. The three phases need to be optimized with respect to time delays in sensor setup and ease of use. Figure 6 illustrates the versatility of the helmet, which can be used with any head size and shape and does not impede the singing function.

The Byzantine expert singer (D. Manousis) produced vowels in both singing and speaking mode, before singing a dozen segments of Byzantine chants in both Mount Athos and Ecumenical Patriarchate of Constantinople styles. The Corsican Paghjella singer (B. Sarocchi) first produced spoken and singing voice using isolated vowels and connected CV syllables with major Corsican vowels and consonants, and then performed three Paghjella songs. For the HBB case, we undertook several testing and recording sessions with our expert, (Da Vox). A specific problem for HBB is the difficulty of stabilizing of the ultrasound probe in view of the large range of motion of the jaw in this singing style, as compared to the

Figure 5 Sensor synchronization experimental setup test. 1) Syringe, 2) Piezo, 3) EGG sensor, 4) US transducer, 5) Camera, 6) Microphone

Figure 6 Top left, arrow indicates gel droplet ejected from tip of syringe; top right, arrow indicates arrival point of droplet on ultrasound probe; bottom left and right, time waveforms of EGG, Piezo, and microphone, signals, see discussion in text.

3.2. Rare-singing data collection

Our data collection activities are focused on preparatory steps dealing with technical, operational and functional requirements. These steps are listed here below and some details are provided for the major steps.

3.2.1. Assessment phase of the hyper-helmet for the different singing types:

The voice capture system has been tested with one expert singer each for the Human Beat Box, HBB (Da Vox) (Figure 7a), the Paghjella (B. Sarocchi) Figure 7b, and the Byzantine (D. Manousis) (Figure 7c) musical styles. Each singer participated in a recording session to validate the helmet with respect to his or her style, and to assist us in specifying an appropriate data collection protocol. A recording session consists of three phases: 1) singer preparation (wearing of the helmet and body sensors), 2) sensor calibration and, 3) data collection proper. The three phases need to be optimized with respect to time delays in sensor setup and ease of use. Figure 6 illustrates the versatility of the helmet, which can be used with any head size and shape and does not impede the singing function.

The Byzantine expert singer (D. Manousis) produced vowels in both singing and speaking mode, before singing a dozen segments of Byzantine chants in both Mount Athos and Ecumenical Patriarchate of Constantinople styles. The Corsican Paghjella singer (B. Sarocchi) first produced spoken and singing voice using isolated vowels and connected CV syllables with major Corsican vowels and consonants, and then performed three Paghjella songs. For the HBB case, we undertook several testing and recording sessions with our expert, (Da Vox). A specific problem for HBB is the difficulty of stabilizing of the ultrasound probe in view of the large range of motion of the jaw in this singing style, as compared to the
other styles. Each of the three singing styles produced about 30 minutes of singing material, which are being used to develop and assess the next steps to be undertaken in the continuing development of our synchronous data collection platform, as well as our data calibration, data display and analysis modules.

Figure 7 a) HBB expert singer (Da Vox), b) Paghjella expert singer (B. Sarocchi), and c) Byzantine expert singer (D. Manousis)

3.3. Pilot data

In this section, some samples of the pilot data are presented. Figure 8 shows similar vocalic /o/ -like sounds by three singers specialized in different singing styles: Byzantine chant (left), Cantu in Paghjella singer (secunda voice, middle) and HBB singer (right). Initial data of ultrasound tongue images and video camera lips image are displayed for all the above singers as shown Figure 9, Figure 10 and Figure 11.

Figure 8 Three vocalic /o/ samples of different style singing voice (Byzantine, Cantu in Paghjella, HBB). Spectrograms (10kHz; band in black on the left) and f0 curves (in blue on the right) are shown in the upper panel, corresponding acoustic waveforms in the lower one.

Figure 9 Byzantine Singing Case: Left) Ultrasound Tongue image. Right) Video camera lip image

Figure 10 HBB singing case: Left) Ultrasound Tongue image. Right) Video camera lip image

Figure 11 Corsican singing case: Left) Ultrasound Tongue image. Right) Video camera lip image

4. Discussion and conclusion

The vocal tract acquisition system for the preservation of rare singing techniques has evaluated on initial data taken on all three singing cases. Results are promising with respect to system reliability, portability and quality of the recorded data. The system still needs to be tested on a larger number of users, and the helmet better adapted to the HBB case due to rapid jaw movements encountered in this singing style. The next steps will involve integrating a feature extraction processing block for all the sensors into the RTMaps platform, as well as provide an easy user friendly interface to drive and control a 3D vocal tract avatar to be developed in future.

5. Acknowledgements

This work is funded by the European Commission via the i- Treasures (FP7-ICT-2011-9-600676-i-Treasures).

6. References


Investigating the domain-specificity of phonetic training for second-language learning: Comparing the effects of production and perception training on the acquisition of English vowels by Arabic learners of English.

Wafaa Alshangiti, Bronwen G Evans

Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK.

w.alshangiti@ucl.ac.uk, bronwen.evans@ucl.ac.uk

Abstract

This study aimed to understand the specificity of second-language (L2) phonetic training by investigating changes in the speech production and perception in forty-six Arabic learners of English assigned to one of three vowel training programs: Production Training (PT), High Variability Phonetic Training (HVPT), and a Hybrid Training Program (HTP: production and perception training). Pre- and post-tests (vowel identification, category discrimination, speech recognition in noise and vowel production) confirmed that training does appear to be largely domain-specific: HVPT learners improved in vowel identification but not vowel production, whilst PT learners showed only small improvements in performance on perceptual tasks, but much greater improvement in production. Additionally, only a small amount of production training appears to be needed to yield improvements: HTP learners improved largely as much as PT learners in production. Lastly, the results indicate that initial L2 proficiency influences learning: high proficiency (HP) learners improved more than low proficiency (LP) learners in their perception of speech in noise, but LP learners showed greater improvement in vowel identification.

Keywords: production training, speech plasticity, the relationship between speech perception and speech production

1. Introduction

The relationship between speech perception and production has been a long-standing focus in speech science. Several theories of speech perception have suggested strong links between speech perception and production (e.g. Liberman et al, 1985), arguing that both processes share common underlying representations - a view supported by brain imaging studies (e.g., Wilson et al., 2004) which show that areas of the brain involved in speech production are activated during listening.

However, despite such links between production and perception, studies of L2 learning have not consistently demonstrated that perceptual learning leads to improvements in production and vice versa. Previous studies have shown that HVPT is beneficial for improving the perception of difficult L2 phonemic contrasts (e.g., Logan et al, 1991), and some have found that this training generalizes to production, at least for some learners. For example, Bradlow et al (1997) showed that after intensive perceptual training for the /r/-/l/ contrast (45 sessions over 3-4 weeks), Japanese speakers improved in their perception and were also able to transfer this learning to the production domain. Similar effects have also been found for vowel perception and production (see Lambacher et al., 2005).

By contrast, others have found little or no relationship between perceptual learning and production, suggesting that perception and production operate somewhat independently.

For example, Hattori (2009) trained Japanese speakers on English /l/-/r/ production over 10 one-to-one sessions using a multi-faceted approach that used explicit feedback from the instructor, real-time spectrograms, and feedback with synthesized versions of their own productions. Hattori found that after intensive production training, Japanese speakers improved their production to become more native-like, but that training did not improve their perception of English /l/-/r/ at all.

The present study aimed to examine whether the type of training affected learning of English vowels by Arabic learners of English. Learners were assigned to one of three vowel training programs: PT, HVPT and a hybrid training program (HTP) which included both production and perception training. Each training program aimed to give learners the same amount of training, but differed in its focus. A battery of pre- and post-tests assessed improvements in production and perception.

2. Method

2.1. Participants

Forty-six native Arabic participants (38 Saudi Arabian, 2 Egyptian, 2 Syrian, 1 Omani, 1 Jordanian, and 2 Kuwaiti) with no reported speaking/hearing problems were tested in London, UK. All participants had the standard Arabic six-vowel system, and were aged 18-39 years old (median 27yrs), had 3-69 months (median 48mths) experience of living in an English-speaking country, and had begun to learn English when they were 5-35 years old (median 13yrs). Participants were randomly assigned to one of the three training programs: PT (16 participants), HVPT (15 participants) and HTP (15 participants). All participants completed the written grammar section of the Oxford placement test (Allan, 1992) to evaluate their English proficiency. This test was later used to categorize learners as low (LP) or high proficiency (HP).

2.2. Apparatus and stimuli

All training, pre-tests and post-tests were conducted in a quiet room with stimuli played over headphones at a user-controlled comfortable level.

Production training was delivered by an instructor (first author) using a custom-made computer program, CALVin (Computer Assisted Learning for Vowels Interface). The stimuli consisted of /h/-/V/-/d/ keywords plus two example words, and the isolated vowels. The keywords were arranged into groups of minimal pairs selected by dividing 14 British English vowels into five clusters; High/front: /i/, /ɪ/ (e.g., heed, hid, head); Open: /æ/, /ʌ/, /ə/ (e.g., had, had, hoard); Central/low back: /ɑ/, /ɔ/ (e.g., heard, hard, hoard); Back: /u/, /uː/, /ʊ/ (e.g., who’d, how’d, hoard); and Diphthongs: /ei/, /ai/ (e.g., hayed, hide). The clusters were selected by conducting hierarchical cluster analysis on English vowel
identification data from Arabic learners of English conducted as part of a previous experiment. All stimuli were recorded by a monolingual male speaker of Standard Southern British English.

The training stimuli for the HVPT were the same as in Iverson and Evans (2009). They consisted of sets of minimal pairs recorded by five speakers of British English (three female and two male). The vowels were divided into four clusters: /ɪ/, /aɪ/, /u/, /ʌ/ (e.g., pet, part, pat, and putty); /i/, /v/, /aɪ/, /æ/ (e.g., feel, fill, file, fail); /v/, /aʊ/, /ɜ/ (e.g., was, woes, wars); and /u/, /aʊ/, /ð/ (e.g., shoot, shout, and shirt). The clusters were based on a hierarchical cluster analysis using English vowel identification data from L2 English speakers.

The pre- and post-test stimuli for vowel identification and category discrimination were the same as those used in Iverson et al. (2012). They comprised recordings of English /b/-V-/t/ words [English vowels that created non-words (e.g., /ɔː/) were not included], and were recorded by 10 speakers of British English (5 male, and 5 female); none of these words or speakers were used in the training, ensuring that all pre- and post-tests measured generalization to new stimuli.

The stimuli for speech in noise were recordings of IEEE sentences. There were 72 lists of 10 sentences, and each sentence contains 5 key words that identified by the listener, e.g., “Glue the sheet to the dark blue background”. The sentence lists were recorded by a male SSBE speaker and were taken from existing recordings at University College London. All the recordings were made in sound treated room. The speech was mixed with white noise; the noise level was fixed to 71dBA, and the level of the speech was varied adaptively.

2.3. Procedure

2.3.1. Training

Production training consisted of five 40 minute sessions of computer-based articulatory training completed over 1-2 weeks with no more than one session per day. The computer program played the 14 keywords (representing the 14 English vowels) each with 2 example words and displayed an animation showing the positions of the tongue, jaw and lips for each stimulus. Each training session started and finished with a 10 minute phase on all 5 vowel clusters (high/front, open, central/low back, back and diphthongs, with the order reversed in the last 10 minute phase) focussing on more difficult vowels. The 20 minutes in between consisted of training on a specific cluster (so all clusters were used over the 5 training session). Prior to each training session, participants completed a 10 minute familiarisation with the software and were shown the relationship between the different positions of their tongue, jaw and lips and resulting vowel sound. That was done by asking them to produce back, front, open, and closed vowels and explaining to them the position of the tongue in each (e.g., heed, had, who’d), and how the position of the moving articulators changes the resulting sound, while looking at a hand mirror.

For each vowel, subjects heard a keyword (e.g., heed) and then the vowel (e.g., /ɪ/) in isolation. They then viewed an animation of its articulation (in midsagittal section), and were guided through a function that described the principal articulatory positions. They were asked to produce the isolated vowel first, then the keyword and example words (e.g., heat, feet). Then they were asked to record themselves producing the isolated vowel, keyword, and the example words, play it back and compare with the native speaker while getting feedback from the instructor. All training was completed in English.

HVPT training consisted of five 45 minute sessions using the UCL vowel trainer (Iverson & Evans, 2009). Each session consisted of 225 trials of vowel identification with feedback using a different speaker in each session (as is typical of HVPT programs).

The hybrid programme consisted of one session of production training followed by four sessions of HVPT. The production training was identical to the training described above except that all 5 vowel clusters were used during the 20 minute phase.

2.3.2. Pre/post tests

There were four pre- and post-tests, (i) vowel identification, (ii) category discrimination, (iii) speech recognition in noise, and (iv) English vowel production. The vowel identification test consisted of 84 trials of a closed-set identification task consisting of /b/-V-/t/ words, randomly selected on each trial. The category discrimination task involved 66 trials, each consisting of three English /b/-V-/t/ words spoken by three different speakers. The three words contrasted different vowel pairs where two words were the same (i.e., the same vowel) and participants had to identify the one that was different. The vowel pairings were /ɪ/-/ɪ/, /aɪ/-/aɪ/, /u/-/u/, /æ/-/æ/; /ɜ/-/ɜ/, /i/-/i/, /əʊ/-/əʊ/, /i/-/i/, and were selected based on previous experiment on vowel perception in Arabic learners of English. The most confusable vowel pairs were selected in descending order until each of the 14 stimulus vowels appeared at least once.

For the speech recognition in noise test, participants listened to IEEE sentences in noise. They were asked to verbally repeat what they had heard and the number of correctly identified keywords was recorded. An adaptive Levitt-procedure (Baker & Rosen, 2001) varied the signal-to-noise ratio to find the Speech Reception Threshold (SRT). Participants identified two blocks of 20 sentences at the pre- and post-test. Each sentence was presented only once. Finally, participants recorded 3 repetitions of each of the /b/-V-/t/ words that they identified in the vowel identification task. The recordings of all participants in the pre- and post-test were analysed acoustically.

3. Results

3.1. Vowel identification task

Figure 1 shows the pre- and post-test vowel identification accuracy for the three different training types according to L2 proficiency. Increases in pre-post accuracy across all training programs regardless of L2 proficiency can be seen. A logistic mixed effects model in vowel identification performance from pre to post-test were highly significant: $\chi^2 (1) = 35.685$, $p<0.0001$. The orthogonal planned contrasts indicated a significant improvement in vowel identification at the post test (bs=0.3112, SE=0.05082, z=-6.125, $p<0.0001$). However the effect of training group was not significant, $\chi^2 (2) = 1.888$, $p>0.05$, indicating that training type did not affect learning. There was, however, a significant effect of proficiency: $\chi^2 (1) = 5.406$, $p<0.05$. The orthogonal planned contrasts showed that LP learners improved in their vowel identification significantly more than did HP learners ($b= 0.25251$, SE=0.08948, $z=2.822$, $p<0.05$). This may be because HP learners had less room to improve than did LP learners (see also Iverson & Evans, 2009).
Although there was no significant effect of training group, there was a significant interaction between group and time \( \chi^2 (2) = 13.78, p<.05 \), demonstrating that some groups improved more than others. The orthogonal planned contrasts showed that HVPT yielded more improvement in vowel identification than did the production training, \( b = -0.2123, SE = 0.05734, z = -3.704, p<.0001 \). However, the orthogonal planned contrasts showed no significant difference between the HVPT and the hybrid training programme. Additionally, there was a significant two-way interaction between group and proficiency, \( \chi^2 (2) = 7.819, p<.05 \). The orthogonal planned contrasts showed a significant difference between HP and LP participants in the HTP group compared with HP and LP participants in the two other training groups (i.e., PT, and HVPT). That is, in the HVPT and PT programs there was an improvement in the LP group performance from pre- to post-test, while all participants, both HP & LP in the HTP improved in their performance, \( b = -0.17220, SE = 0.06758, z = -2.548, p<.05 \).

### 3.2. Category discrimination

A linear mixed model was built for category discrimination data. The best fit-model included time (pre-post), as a fixed factor, and participant and word pair as random factors. The main effect of time was significant \( \chi^2 (1) = 27.99, p<.001 \), which suggest a change from pre to post-test. The orthogonal planned contrasts showed a significant improvement at the post-test, \( b = 0.045, SE = 0.00862, pMCMC<.001 \). The best model excluded training group, and the interaction between time and proficiency indicating no significant effect of these factors. Inspection of the data however, revealed that changes were small and that improvement was largely driven by changes in *bet-bit*.

### 3.3. Speech recognition in noise

After training, all HP listeners improved in their ability to process speech in noise; all HP listeners had lower SRTs after training (see Fig. 2). In contrast, LP listeners did not appear to improve after training. A linear mixed effects model confirmed significant effects of time, \( \chi^2 (1) = 17.48, p<.001 \), and L2 proficiency \( \chi^2 (1) = 5.708, p<.05 \). Additionally, orthogonal planned contrasts indicated that the HP listeners improved their speech recognition in noise at the post test more compared to those of the low LP listeners, \( b = -1.997, SE = 0.835, pMCMC<.05 \). There was no significant effect of training group, \( p > 0.05 \), indicating that training type did not affect performance in this task.

### 3.4. English vowel production

#### 3.4.1. Acoustic analysis

Monophthongs were divided into three groups: M1 (*beat, bit, bet, birt*), M2 (*bat, bat, bart*), and M3 (*boot, bought, bot*). Changes in F1 and F2 were analysed using linear mixed effects model. For M1 vowels, the main effect of training group was significant, \( \chi^2 (2) = 5.956, p=0.05 \), which suggests that changes in F1 values were differed between training groups. Orthogonal planned contrasts indicated a significant change in F1 values in the vowels produced by speakers in the PT group and HTP groups compared with those produced by the HVPT group, \( b = -0.0161, SE = 0.0288, pMCMC<.05 \), this was particularly noticeable for the /æ/ - /ε/ vowel contrast, where at the pre-test learners in the PT & HTP groups produced /æ/ with a lower F1 (more closed), and /ε/ with high F1 (more open), and after training they altered their values. No other significant effects were found for F1 and there were no significant effects for F2.

For M2 vowels, there were no significant changes in F1, but there were significant changes in F2 due to training group \( \chi^2 (2) = 7.5499, p<.05 \), and time \( \chi^2 (1) = 10.069, p<.05 \). Orthogonal planned contrasts showed a significant change in F2 values in the vowels produced by speakers in PT group at the post-test (especially *bart* vowel which they produce with a lower F2 at the post test) compared to those produced by
speakers in the HVPT, and in the HTP group, (b=0.078, SE=0.0299, pMCMC<.05), indicating that training type affected production of this group of vowels.

For M3 vowels there were no significant changes in F1 or F2. In brief, these results suggest the vowel space changed slightly (at least for the M1 and M2 vowels) but that these changes were limited to those learners who completed the HPT and PT training programs, i.e., their training included explicit training in production.

4. Discussion and conclusion

The present study examined whether phonetic training for second language learning is domain specific. The effect of 3 different training programmes, PT (production-based), HVPT (perception-based) and HTP (production and perception) on L2 production and perception was investigated using a battery of pre- and post-tests. The results demonstrated that different types of training affected performance in production and perception tasks differently. After training, learners who had completed perception-based training programmes (HVPT and HTP) improved more in their vowel identification than those who completed PT. However, those who received production training (PT and HTP) improved more in production that those who received only perception training (HVPT). Additionally, the results demonstrated that initial proficiency in the L2 affected learning in some tasks, in particular, speech in noise.

Overall, these findings indicate that training is domain-specific: that is, production trains production and perception trains perception. Previous research has shown that HVPT is particularly effective in improving pronunciation (e.g., Iverson & Evans, 2009) and learners in our HVPT and HPT training conditions also improved significantly more in their vowel identification than did those who received production training. However, in contrast to previous research (e.g., Bradlow et al., 1997) improvements in perception as a result of training did not generalize to production. Learners in the HVPT condition did not improve in production and only those who received production-based training (PT & Hybrid condition) produced more native-like vowels after training. Learners in both the PT and HTP conditions were able to generalize production training to different set of words (i.e., not included in the training), and adjusted their production of some confusable vowels, in particular bit and bet. Before training learners produced /e/ as a more closed vowel (i.e., more similar to /i/), and /i/ as a more open vowel (i.e., more similar to /e/), but after training they produced these vowels more like native speakers, such that /i/ was produced with a higher F1 than /e/. However, as in previous studies of production training (e.g., Hattori, 2009) learners who received only production training did not improve in their perception.

What is being learned? Interestingly, all participants improved in their performance on a speech recognition in noise task, regardless of training type. This suggests that the finding that learning is domain-specific may in part be task driven. In HVPT programmes like that used in the HVPT and HPT training conditions, learners identify different sets of minimal pairs, the same skill tested in the vowel identification task. Given that there was very little evidence for changes in low-level speech perception (i.e., few changes in performance on a Category Discrimination task), this indicates that learners were not making changes to underlying representations as a result of training. Instead, it is possible that HVPT enabled learners to become better and more efficient at mapping their native categories onto the L2 sounds they heard (see also Iverson & Evans, 2009). Similarly, production training may have enabled learners to develop more native-like motor patterns for particular vowels that they were better able to map onto their existing underlying representations, but these underlying representations themselves did not change.

In contrast, improvement in a more real-world task of speech perception, sentence recognition in noise, appeared to rely on initial proficiency with the L2 rather than training type. Although all learners improved in their performance in speech in noise, HPT learners improved more than LP learners regardless of training type. In our study, proficiency was determined by performance on a written comprehension test that tested grammatical and lexical knowledge. One possibility then is that a certain level of knowledge is necessary to apply learning on isolated sounds and words to real-world contexts. In brief, the results from this study confirm that phonetic training is largely domain specific and additionally indicate that adjustments to phonetic processing might be lexically driven. Perceptual training led predominantly to improvements in speech perception, whilst production training, even only a small amount, led to changes in production. However, performance on a speech in noise task was affected predominantly by proficiency rather than by training type. This implies that whilst perception and production may share the same underlying representations, the way in which they are mapped to tasks of perception and production in an L2 might differ.

Acknowledgement

We are grateful to Mike Coleman for his efforts in putting the computer aided programme (CALVin) together, and to Mohammed Alqarni for help with subject recruitment.

5. References


Spectral similarity and listener judgments of phonetic accommodation

Molly Babel¹, Michael McAuliffe¹, Grant McGuire²

¹Department of Linguistics, University of British Columbia, Canada
²Department of Linguistics, University of California, Santa Cruz, USA

molly.babel@ubc.ca, michael.mcauliffe@alumni.ubc.ca, gmcguil@ucsc.edu

Abstract

A fundamental issue in studies of phonetic accommodation is knowing what to measure to properly assess the degree of convergence between voices. Hand-selecting acoustic parameters (e.g., F1, F2, duration) may miss the relevant parameters. Listener judgments are also relevant: if accommodation is a strategy used in real-world communication, the role of listener judgments of accommodation is critical. Thus, many researchers rely on listener judgments of similarity. Listener-based experiments, however, have their own drawbacks. There is a strong desire in the field to have a reliable global acoustic measure of convergence. While there has been some use of more general acoustic measures of similarity no study has compared a global measure of spectral similarity to listener judgments, nor has a study compared different measures of spectral similarity. In this paper we compare measures of spectral similarity and their relationship to listener judgments.

Keywords: phonetic convergence, spectral similarity, speech perception

1. Introduction

Phonetic accommodation is the process whereby interacting talkers come to be more similar in some way. What is the best way of assessing this increase in similarity? There are three basic approaches to this question: (1) use listener assessment of similarity, (2) deploy targeted acoustic measurements, or (3) use a global acoustic measure of similarity. The purpose of this paper is to discuss these approaches and assess the relationship amongst these different metrics.

There are several reasons why we might want to consider listener judgments of convergence as the gold standard. First, accommodation is a behaviour rooted in perceptual processing: perceivers modify their speech patterns based on perceptual input which may be audio-only, visual-only, or audio-visual in nature [13]. Convergence has a positive effect on how others view an individual in an interaction [16] – such a consequence of accommodation is only possible if it is perceivable by a listener or perceiver. Phonetic accommodation has also been suggested as a means for initiating sound change via the change-by-accommodation model [1, 20, 19]. In order for phonetic accommodation to provide a mechanism towards sound change, those changes will need to be perceptible. Initial work on accommodation used listeners’ Gestalt judgments of similarity to assess dialect convergence [9]. Since then, listener judgements have frequently been used in seminal work on convergence [10, 17].

Phonetic accommodation also sparks interest in the nature of cognitive representations of spoken language, and this, along with an interest in the what of the process, has prompted a number of researchers to focus on targeted acoustic measurements. To this end, essentially any measure has been shown to be available for accommodation: VOT [15, 18], vowel formants [2, 3], fundamental frequency [6], amplitude [14]. This approach contributes, however, to a potential file-drawer problem in the field, as there is no reason to assume that all available phonetic characteristics will be accommodated in every potential instance of accommodation or by every listener-turned-talker.

There has also been interest in applying more general acoustic measures of similarity. For example, Delvaux and Soquet (2007) based their analysis on mel frequency cepstral coefficients (MFCCs), finding evidence for convergence. Lewandowski (2012) used a measure of spectral similarity that averaged cross-correlation values from amplitude envelopes in four logarithmically spaced frequency bands in a range of 80-7800Hz. This method also proved fruitful in her analysis, showing that participants with increased “phonetic talent” accommodated their interlocutor more.

The goal of this paper is begin an initial comparison of these diverse measures. Given space restrictions we limit our discussion here to comparing listeners’ quantification of convergence to a set of global acoustic measures of spectral similarity. No study has compared a global measure of spectral similarity to listener judgments, nor has a study compared different measures of spectral similarity. In this paper we compare three types of measures of spectral similarity and their relationship to listener judgments across two data sets of single word productions from auditory naming tasks. To assess phonetic distance we used a dynamic time warping algorithm us-
ing (1) spectrograms and (2) 12 MFCCs derived from the spectra of those spectrograms. Both of these are asymmetric distance functions, meaning that we compare shadower-to-model distance and baseline-to-model distance. We also used (3) the amplitude envelope similarity measure from Lewandowski using 4 and 8 frequency bands.

2. Methods

2.1. Data

We use two corpora of shadowing data to compare listeners’ judgments of similarity to a set of acoustic measures of similarity. We offer a brief description of both data sets.

Babel et al. [4] examined the accommodation patterns of New Zealand shadowers who were presented with an Australian model — the AXB listeners in this study spoke Canadian English. For this data set we also have unpublished AXB data from New Zealand English listeners (n=93); this offers the crucial comparison of whether the spectral measures of similarity better align with the NZ listeners or the Canadian listeners. Given that the spectral measures are linguistically naive and assess auditory-acoustic similarity, we predict that the spectral similarity measures will correlate more strongly with the responses from the Canadian listeners who are also linguistically naive with respect to what constitutes similarity in Antipodean dialects.

Our other data set, Babel et al. [5], reports on single word accommodation in North American English (NAE) speaking shadowers and NAE-speaking AXB listeners. The experiment used a set of 8 model talker voices which were selected to test hypotheses related to whether spontaneous accommodation patterns are affected by the novelty of a voice or listeners’ preferences with respect to voice types. Given this we refer to this data set as Novelty and Preference in Accommodation (henceforth, NPA).

2.2. Acoustic measures of similarity

2.2.1. Spectrogram similarity

Spectrogram similarity was computed as the inverse log distance between two spectrograms using the dynamic time warping algorithm implemented in Praat [7]. The acoustic tokens were first converted to spectrograms, with a window length of 0.005 seconds, maximum frequency of 7800 Hz, time step of 0.002 seconds, frequency step of 20 Hz, and a Gaussian window. Once both acoustic tokens were converted, we used Praat’s dynamic time warping algorithm with no slope restriction but matched beginning and end positions, to give a distance measure between the two spectrograms. As the distance returned by the algorithm was in excess of three for all tokens, the inverse log of the distance returned a value between 0 and 1.

2.2.2. Mel frequency cepstral coefficient (MFCC) similarity

The mel frequency cepstral coefficient (MFCC) similarity was computed in a similar fashion to the spectrogram similarity, but with a matrix of MFCCs rather than spectrograms as the input to the dynamic time warping algorithm, similar to the acoustic distance measure used in [12], but using 20 coefficients rather than 12 following [8]. For each acoustic token, the first 20 MFCCs were extracted using a window length of 0.015, a time step of 0.005, and a maximum frequency in mels corresponding to 7800 Hz. We once again used Praat’s dynamic time warping algorithm on the output of the conversion, with no slope restriction but matched beginning and end positions, to give a distance measure. As the distance returned by the algorithm was in excess of three for all tokens, the inverse logarithm of the distance returned a value between 0 and 1.

2.2.3. Band-based amplitude envelope similarity

Following Lewandowski [11], we implemented an algorithm that matches normalized amplitude envelopes across four and eight bands logarithmically spaced from 80 Hz to 7800 Hz. For two acoustic tokens of the same lexical item, we can cross-correlate the sampled waveform for each band to get cross-correlation values by alignment. For each possible alignment, we averaged the cross-correlation values across the bands and then took the maximum averaged value as the best possible alignment of the two acoustic tokens, which can be interpreted as a similarity value ranging from 0 (complete distinct) to 1 (identical waveforms).

3. Analysis and Results

Table 1 summarizes correlations between NZ and Canadian listeners’ judgments of similarity and the three global measures of similarity: 8-band amplitude envelope similarity, spectrogram matching, and MFCCs. The 8-band envelope measure offered consistently higher correlations with the listener judgments, so we focus on the 8-band instead of the 4-band analysis used in Lewandowski [11].

There are several observations we can make from this table. First, the amplitude envelope measure is the only global measure of acoustic similarity which is well correlated with listeners’ quantification of convergence. Comparing the correlation for New Zealand listeners (t(2198) = 9.4, p < 0.001, r = 0.2) and Canadian listeners (t(2198) = 12.4, p < 0.001, r = 0.26) reveals that the relationship between the acoustic measure of similarity is stronger for the Canadian listeners [z = -2.1, p < 0.05]. These patterns are shown in Figures 1 and 2 for the New Zealand and Canadian listeners, respectively.
Of the spectral measures themselves, the MFCC and spectrogram matching measures are well correlated with each other, but not with listener judgments or the envelope measure. In fact, in some instances they are negatively correlated with listeners’ judgments.

Table 1: Correlations between listener judgments and acoustic similarity measures for NZ diphthongs data. NZ = New Zealand listener judgments of similarity, CA = Canadian listeners judgments of similarity, Envelope = Amplitude envelope band similarity, Sp’gram = Dynamic time warping spectrogram similarity, MFCC = Mel frequency cepstral coefficients.

<table>
<thead>
<tr>
<th></th>
<th>NZ</th>
<th>CA</th>
<th>Envelope</th>
<th>Sp’gram</th>
<th>MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>1</td>
<td>0.41</td>
<td>0.2</td>
<td>-0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>CA</td>
<td>0.41</td>
<td>1</td>
<td>0.26</td>
<td>-0.06</td>
<td>-0.03</td>
</tr>
<tr>
<td>Envelope</td>
<td>0.20</td>
<td>0.26</td>
<td>1</td>
<td>-0.13</td>
<td>-0.09</td>
</tr>
<tr>
<td>Sp’gram</td>
<td>-0.02</td>
<td>-0.06</td>
<td>-0.13</td>
<td>1</td>
<td>0.55</td>
</tr>
<tr>
<td>MFCC</td>
<td>0.01</td>
<td>-0.03</td>
<td>-0.09</td>
<td>0.55</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1: New Zealand listeners’ quantification of convergence by amplitude envelope similarity metric.

For the second data set, summarized in Table 2, the amplitude envelope metric of global acoustic similarity also provides the strongest correlation with respect to listeners’ judgments of similarity. For this data set, however, the MFCCs are also significantly correlated with listeners’ responses.

Table 2: Correlations between listener judgments and acoustic similarity measures for NPA data. Listeners = listener judgments of similarity, Envelope = Amplitude envelope band similarity, Sp’gram = Dynamic time warping spectrogram similarity, MFCC = Mel frequency cepstral coefficients.

<table>
<thead>
<tr>
<th></th>
<th>Listeners</th>
<th>Envelope</th>
<th>Sp’gram</th>
<th>MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listeners</td>
<td>1</td>
<td>0.22</td>
<td>0.04</td>
<td>0.11</td>
</tr>
<tr>
<td>Envelope</td>
<td>0.22</td>
<td>1</td>
<td>-0.02</td>
<td>-0.002</td>
</tr>
<tr>
<td>Sp’gram</td>
<td>0.04</td>
<td>-0.02</td>
<td>1</td>
<td>0.72</td>
</tr>
<tr>
<td>MFCC</td>
<td>0.11</td>
<td>-0.002</td>
<td>0.72</td>
<td>1</td>
</tr>
</tbody>
</table>

[11] provides the strongest match to listener judgments. Crucially, we also find an 8-band analysis is an improvement over the 4-band analysis described in Lewandowski’s dissertation.

In the New Zealand data set the amplitude envelope similarity measure was a better fit for Canadian listener judgments than for the New Zealand listener judgments. Recall that the stimuli were New Zealand and Australian voices. This improvement in fit for Canadian listeners is likely due to their being linguistically naive with respect to what constitutes similarity in Antipodean varieties of English, just as the spectral measures are linguistically naive. This highlights a potential hiccup in the exclusive use of listener judgments for convergence measures: the micro-changes that take place in speech behaviour in the course of accommodation may be below a listener’s consciousness despite begin readily observable. It is this micro-changes that seed short-, medium-, and long-term...
language change despite their subtlety. While we did find significant correlations between the amplitude envelope similarity measure and listener judgments, the correlations are far from impressive. This perhaps lacklustre conclusion serves to illustrate that one’s selection of measure of convergence should be related to the hypothesis being tested. Our result also underscores the non-linearities in listeners’ assessment of speech patterns.

5. Acknowledgements
Thanks to Christine Mooshammer and Mark Tiede for organizing the ISSP satellite workshop on Interpersonal coordination and phonetic convergence. Data collection at various stages of this project were facilitated by Paul Warren, Jen Hay, Jamie Russell, Sophie Walters, Graham Haber, and Alice Nicholls: thank you.

References
Comparison of articulatory strategies for a bilingual speaker: Preliminary data and models

Pierre Badin\textsuperscript{1}, Thomas R. Sawallis\textsuperscript{2}, Simon Crépel\textsuperscript{1}, Laurent Lamalle\textsuperscript{3}

\textsuperscript{1}GIPSA-Lab / DPC, UMR 5216, CNRS – Grenoble Alpes University, France
\textsuperscript{2}New College, University of Alabama, Tuscaloosa, AL 35487, USA
\textsuperscript{3}Inserm US 17 — UMS IRMaGe – Grenoble Alpes University, France

pierre.badin@gipsa-lab.grenoble-inp.fr, tom.sawallis@ua.edu

Abstract

This study serves the goal of comparing the articulatory strategies used by bilinguals, to better understand both bilingualism and the differences between languages. We have therefore obtained mid-sagittal MRI scans of a bilingual talker producing all the phonemes in both his native American English (AE) and in French (FR), a second language in which he is highly fluent. We compared his articulatory contours in the two languages through: 1) direct comparison of the contours of analogous phonemes, 2) comparison of articulatory nomograms, and 3) crossed reconstructions of one language’s articulations by a model based on the other language. Various interesting observations were made, including that labial protrusion in vowels is greater in FR than in AE, and that tongue body movements were greater and more oriented toward fronting in AE than in FR.

Keywords: Bilingualism, French, American English, articulation, MRI, articulatory measurement, articulatory model.

1. Introduction

Greater familiarity with the comparative phonetic behaviors of bilinguals in their various languages is worthwhile on its own merits, but also with regard to applications in the field of second language (L2) pedagogy. The great majority of the literature linked to these questions focuses on either one or the other of bilingualism or articulation, but the studies focusing crucially on their intersection seem scarce, and they generally aim at two specific themes.

The first theme is articulatory setting (AS), defined by Homikman (1964) (cited by Mennen et al., 2010) as the language-specific habitual configuration toward which the vocal apparatus tends to return. Mennen et al. (2010) discuss the difficulty of separating the influence of the AS from other aspects of the language, such as phonemic inventory, and note that the inter-speaker variability which affects most measurements can mask the AS difference between languages. Wilson (2006) used Optotrak and ultrasound to characterize respectively mandible position and the form of the tongue during inter-speech posture (ISP) as an estimate of AS. He found that ISP distinguishes between the Francophone and Anglophone groups in his Canadian subjects, the Anglophones having on average a higher tongue tip, more lip protrusion, and a narrower horizontal lip aperture. Moreover, he found that the most successful bilinguals (those judged native by monolinguals of both languages) do have different ISPs for each language, which close to those of monolinguals. He advises teaching of AS to L2 learners, but we consider that premature until a demonstration that AS is the source of foreign accent or specific segmental errors.

The second theme, classic in bilingualism research, is the study of transfer, which is the appropriation of aspects of the talker’s native language for use in a target language, aspects which for our purposes would be phonological or articulatory. Tepperman et al. (2009) try to detect traces of phonological transfer with articulatory effects. Using MRI video, they compare English /v-w/ and /d-b/ contrasts between 3 Americans and 3 German speakers in their L2 English. The Germans show definite differences between the phonemes, but notably less than the American native speakers. Moreover, where the differences are lacking, the substitutions are not differentiated from their native German, which implies the expected transfer from their L1 to their L2.

Oh (2002) used ultrasound to study articulatory timing differences between /l/ productions by monolingual and bilingual English and Korean children. Findings showed that bilinguals blend the gestures specific to the two languages.

Zerling (1992) study of lip opening dimensions for vowels used French data he collected himself and American English data from studies by Fromkin (1964) and Linker (1982). His analysis contrasts not only the languages, but also the two studies of English, and he finds the dimension of phonological labialization inadequate to explain the differences. He adds the dimensions of type, degree, and strategy of labialization to fully account for and distinguish the languages. However, these studies do not address bilinguals or their L2, so we here learn nothing about their articulatory behavior.

Montagu (2002) compares lip shapes and acoustics for oral, nasal, and nasalized vowels of French in natives and in American English natives speaking L2 French. She finds that the French speakers distinguish the back nasals /\textipa{ã}/ from their oral analogs not only by phonetic nasality, but also by rounding (i.e., reduced lip opening) and by protrusion. For /\textipa{ã}/, bilingual Americans reproduce the French native articulation fairly well. However, they fail to reproduce the French hyper-rounding of /\textipa{ã}/, a nasal analog of an oral vowel that is already rounded. This seems to indicate that neither AS nor phonological transfer is adequate to explain bilingual articulation in a global fashion, and instead that both are required, plus a specific analysis of each phoneme, in order to fully address bilingual articulatory strategies.

This limited literature drew us to investigate and model the articulation of an American English native speaker who, despite a late start (during adolescence) attained a very high skill level in French. The study aims to generally document both similarities and differences in articulatory strategy between the two languages, which is thus a larger focus than simply AS or transfer. The main steps in the study are: 1) acquisition and treatment of MRI data, 2) direct comparison of analogous phonemes, 3) development of articulatory models, and 4) using those models for global comparison of the two articulatory systems.
2. Articulatory data acquisition and processing

The speaker was recorded in two sessions six weeks apart, one in AE and one in FR, following a standard protocol (cf. Badin & Serrurier, 2006, for details). For each session, care was taken that only the language of interest was spoken for the interactions with the operators, in order to keep the speaker in the right language mode.

The MRIs were recorded at Grenoble’s MRI facility, IRMaGE, with a Philips Achieva 3T.0 TX scanner in TSE mode, with a neurovascular SENSE coil, in turbo spin echo mode (TSE, TE = 10.7 ms, TR = 426 ms). Acquiring the midsagittal slice (4 mm thick, 1 mm/pixel resolution, field of view 256x256 mm²) took 8.1 seconds, during which the speaker had to sustain a fixed articulatory configuration.

The AE corpus contained 201 items: vowels /i ɪ e ɛ a ɔ o ʊ u ʌ ǝ ɹ̩/ and diphthongs /au ɔi ai/ under four conditions (alone, in four CVC contexts /m ɹ ɻ ɫ/, in a list of 17 natural words, and followed by /ɹ/ in a list of 14 natural words), and the consonants /p t k f s ʃ m n ŋ ɹ ɻ ɫ j w/ in six vowel contexts /i e æ a o u/. The French corpus contained 144 items: the oral /i e æ a o u/ and nasal /ã ɛ̃ ɔ̃/ vowels, and consonants /p t k s f j m n ʁ l/ in the thirteen vowel contexts. Note that only voiceless counterparts were used for stops and fricatives (voicing being impossible to maintain for the 8.1 seconds of MRI acquisition).

Using the procedure in Badin & Serrurier (2006), the contours of the various deformable organs were manually created and adjusted from the midsagittal image, while the rigid contours were manually positioned by translation and rotation, and all articulations were ultimately aligned on the same position of the skull’s rigid structures.

3. Pairwise comparison of analogous phonemes

As a first step we did pairwise comparisons of articulations that are analogous across the two languages. For the isolated vowels /i e æ a o u/, the main differences are located at the lips, more protruded for FR /u/ and more closed for FR /o/; we observe also that the tongue is further back for FR /u/ than for AE /u/, that it is lower and more bunched for FR /o/, and lower for the FR /a/ (see examples in Figure 1).

Comparing the consonants /p t k f s j m n/ analogous in the two languages, in the five analogous vowel contexts /i e a o u/, shows again that the main differences are in the lips in /u/ context for all consonants, as well as in /o/ context for /t k s m n/, the protrusion being stronger in French than in English. Lingual differences are also found: the tongue is further back in French than in English for all consonants in /u/ context, except for /n/. Also, small differences in the lingual articulation can be observed for labial consonants /p f m/, which is logical since tongue position is not specified for those consonants. As expected, the comparisons of AE-FR pairs /ɹ–ʁ/ and /ɫ–l/ in the five vowel contexts /i e a o u/ reveals major articulatory differences, despite their analogous phonological status. In particular, we found greater tongue body lowering for the dark velarized AE /ɫ/ than for the light laminal FR /l/ (see examples in Figure 2).

![Figure 1: Superposition of isolated vowels /i a u o/ in AE (red) and FR (blue).](image1)

![Figure 2: Superposition of consonants /ɫ–l/ (top) and /ɹ–ʁ/ (bottom) in AE (red) and in FR (blue) in /i a u o/ contexts.](image2)

In this preliminary work, we consider that the AE and FR /a/ are analogous. It would however seem that the speaker realizes it as [ɛ] in his mother tongue AE, more central than the FR [a] that he has learned to produce. This point will be investigated elsewhere.
4. Articulatory models

The next step compared the articulatory spaces of the two languages at a more general level, using articulatory modeling. From different subsets of the contour database, we built a series of articulatory models using linear decomposition methods described in Badin et al. (2013). For the tongue, the \( JH \), \( TB \), \( TD \), and \( TTV \) components correspond respectively to the influence of the jaw position, to the displacement of the tongue body, to the bunched or flat shape of the tongue dorsum, and to the vertical apex movement. For the lips, the measured protrusions (\( UL_{\text{pro}} \) and \( LL_{\text{pro}} \)) and heights (\( UL_{\text{hei}} \) and \( LL_{\text{hei}} \)) of each lip were used. We built these models from six different corpuses: oral vowels for FR (10) and AE (13), analogous vowels and consonants for FR and AE (45), and full corpus for FR (144) and AE (201). For the full corpuses the RMS reconstruction errors span from 0.5 mm for the upper lip to 1.7 mm for the tongue for English, and from 0.5 mm to 1.4 mm for French. These results are quite comparable with those documented in other similar studies in the literature (see references in Valdés Vargas et al., 2012).

To illustrate the general behavior of these models, we generated nomograms, i.e. displays of the articulators’ shapes across the core range of the model’s control parameters. Each of the following figures shows the variations of one or two parameters.

Figure 3 illustrates that the lip protrusion for vowels has a much larger range of movement in FR than in AE, which generalizes the observations in section 3. The difference was less striking when masked by consonantal coarticulation (hence our choice of the V-only subset from the corpus). Figure 3 also illustrates the Tongue Body component differences between FR and AE: the vertical movement is clearly greater and more oblique towards front in AE than in FR, especially when considering full corpuses. This corresponds in particular to the /s–x/ articulatory differences shown in Figure 2.

Figure 4 shows that vertical movement of the lower lip is slightly stronger for French than English. The tongue dorsum component \( TD \) for both English vowels and the full English corpus displays a clearly greater capability to bunch towards the velum than is shown in French. This is likely related to the specifically bunched shape of the English /ɹ/.

Figure 5 seems to show that the tongue tip movements are more localized towards the apex in FR than in AE. This might be due – among other reasons – to the velarization associated with the dark AE /ɫ/, and we can observe its trace on the tongue body movements associated with the \( TTV \) component for English (but not French), shown in the right-hand column of Figure 5.

5. Evaluation of cross-language capabilities of the models

In order to evaluate the overlap of the articulatory spaces for AE and FR, we followed the methodology proposed by Serrurier et al. (2012) to compare speech and feeding articulations: we built one model based on AE Vowels and Consonants and another based on analogous articulations of FR. We then reconstructed the FR articulations with the AE model and vice versa. These simulations did not lead yet to significantly different cross-reconstructions: it is possible that, though their components may have different characteristics, both models could represent equally well French and English.
6. Conclusions and perspectives

We have acquired a large set of articulatory contours produced by a bilingual speaker in both his native AE and his L2 FR. We compared his AE and FR articulations through: (1) direct pairwise comparison of contours, (2) comparison of articulatory nomograms, and (3) cross-language articulatory reconstruction of articulations. Despite the constraints imposed on the speaker by the limitations of static MRI, this study has generated a variety of interesting results that should benefit research at the crossroads between speech and bilingualism. In future studies, we foresee use of real time MRI, as well as simultaneous recording of speech sounds associated with the articulations, for later analysis. We will also have to assess more precisely the degree of bilingualism of the speakers. Over the longer term, we envisage recording more speakers and language pairs, with speakers of varying degrees of bilingualism. Applications in the domain of Computer Aided Pronunciation Training (CAPT) will also be explored.

7. Acknowledgements

This work was partly funded by the French program “Investissement d’Avenir” run by the “Agence Nationale pour la Recherche”, grant “Infrastructure d’avenir en Biologie Santé - ANR-11-INBS-0006”.

8. References


The Importance of Tonal Cues for Untrained Listeners in Judging Prominence

Stefan Baumann
IfL Phonetik, University of Cologne, Germany
steфан.baumann@uni-koeln.de

Abstract

The paper discusses the question of how untrained listeners perceive and interpret prosodic and non-prosodic features in a prominence rating task. We adopt the method of Rapid Prosody Transcription (Cole et al. 2010) to read German. Results suggest that tonal cues are most relevant for prominence perception, with rising accents reaching the highest prominence scores. Accent position, in particular the position of the nuclear accent, seems to be less important for untrained listeners. Furthermore, non-prosodic factors related to the accentability of a word based on language-specific knowledge and expectations (part-of-speech, word frequency, morpho-syntactic focus marking) also influence prominence ratings but to a lesser extent.

Keywords: prosody, prominence, untrained listeners, perception, nuclear accent, pitch accent type

1. Introduction

The factors determining the manual annotation of prosodic prominence are not yet fully understood. Usually, annotation is carried out by expert transcribers who may produce artefacts since they are biased by top-down knowledge derived from specific theoretical assumptions, e.g. in their treatment of secondary prosodic prominences in relation to a nuclear accent (defined as the last and both structurally strongests and semantic-pragmatically most important pitch accent in an intonation unit). However, the vast majority of prosody ‘users’ is untrained, and the question the present paper poses is how they perceive and interpret prosodic and non-prosodic features when asked to judge the prominence of words in connected speech.

In order to examine this question we broadly apply the method of Rapid Prosody Transcription (RPT), developed by Jennifer Cole and her colleagues (e.g. Cole et al. 2010) for collecting coarse-grained prosodic judgements from untrained listeners based on their integrated perception of form and function. That is, the resulting prominence scores for each test word are regarded as primary data that can be interpreted linguistically.

The present explorative study examines to what extent prominence perception is signal-driven (bottom-up) or expectation-driven (top-down) by investigating the influence on prominence judgements of a) phonetic-phonological (prosodic) factors and b) other factors related to a word’s ‘accentability’.

More specifically, the study asks the following questions:
1. Do untrained listeners perceive nuclear accents as prominent due to their – structural – importance for the message?
2. Do untrained listeners perceive secondary accents as prominent in spite of their weaker structural position (i.e. pre- and postnuclear)?
3. Are acoustic cues (F0 height, F0 movement and alignment, duration, intensity) dominant for prominence judgements?

2. Method

2.1. Material

The test material consists of 60 German sentences (containing between 5 and 18 words) read by 14 native speakers. They were selected from various databases and display different focus structures (including second occurrence focus and verum focus structures) as well as various information status categories of referring expressions (new, accessible and given) marked by varying degrees of prosodic prominence. All sentences are declarative and have low final boundary tones. The main selection criterion was the occurrence of various accent types (AT) in different accent positions (AP), according to a consensus annotation of three intonation experts (trained with GToBI; Grice et al. 2005). Fig.1 shows the categories tested, along a scale of assumed perceptual prominence.

![Figure 1: Accent types (AT) and accent positions (AP) tested in the experiment.](image)

The accent type (AT) classification takes the tonal movement in the vicinity of the accented syllable into account, in particular information about the onglide to the accented syllable. That is, a falling accent has higher pitch immediately before the accented syllable (comprising the ‘early peak’ accents H+L* and H+!H* in GToBI), whereas a rising pitch accent either has a low pitch target before the accented syllable (L+H* in GToBI) or rises from the accented syllable (L*+H). Low (L*) and high (H*) accents indicate tonal targets on stressed syllables without a considerable tonal movement in their immediate vicinity. The category ‘no pitch accent’ comprises complete lack of accent and phrase accents (see below).

Fig.2 indicates that the duration of a syllable differs as a function of the type of accent it carries: low (and falling) accents trigger the longest durations probably compensating for a lack of (rising) tonal movement as the main cue for prosodic prominence. These data confirm results by Röhr and Baumann (2010).

As to accent position (AP), we do not only differentiate between pre- and postnuclear accents but also between final and non-final nuclear accents. This latter distinction is made in order to capture the potentially secondary status of non-final nuclei, since they are often classified as nuclear only due to the occurrence of hesitation pauses or a slow speech rate.
Small breaks like these trigger the insertion of intermediate phrase boundaries in most ToBI systems (which renders the last pre-boundary accent nuclear by definition), although they often do not appear to be intended by the speaker. Prenuclear accents are fully-fledged pitch accents, while phrase accents are not. They are associated with a stressed syllable in postnuclear position and often lack a considerable tonal movement (see also Grice et al. 2000).

![Figure 2: Differences in mean duration of syllables as a function of accent type.](image2)

The duration differences given in Fig.3 show that syllables carrying nuclear accents are generally longer than syllables marked by pre- and postnuclear accents. As expected, unaccented syllables are the shortest.

![Figure 3: Differences in mean duration of syllables as a function of accent position.](image3)

The non-phonetic-phonological factors which have been claimed to affect the predictability and, in turn, accentbility of a word, and which are tested in the experiment are word frequency (based on the SubtLex corpus, see Brysbaert et al. 2011), part-of-speech and morpho-syntactic focus marking (i.e. presence or absence of a focus-sensitive particle like nur (‘only’) or sogar (‘ever’)).

2.2. Procedure

In a self-paced perception test with 28 prosodically untrained native speakers of German (aged between 18 and 58, with a mean of 24.8 years), the 60 utterances were presented over headphones in a silent room in pseudo-randomized order. According to the Rapid Prosody Transcription method, only a brief instruction without examples was given. After that, the subjects had to judge which words they perceive as prominent by underlining them on a sheet of paper displaying the written text. Capitalization and punctuation marks were removed in order to avoid structural cues that may influence the subjects’ judgements. The actual instruction was spelled out as follows:

[...]'Ihre Aufgabe besteht nun darin, sämtliche Wörter, die Sie in einer Äußerung als betont / hervorgehoben / wichtig wahrnehmen, auf dem Transkript zu unterstreichen.' 

([...] Your task is now to underline all the words on the transcript which you perceive as stressed/ highlighted/ important.)

The judgements result in a probabilistic prominence score (p-score) for each word expressing the proportion of subjects who labelled a word as prominent.

3. Results

The statistical analyses in this pilot study are restricted to descriptive measures (i.e. percentages/p-scores and correlations), since the utterances chosen and thus the data points analysed were not counterbalanced for all factors reported here. A follow-up study will investigate a strictly controlled data set that allows to explore interdependencies between the prosodic and non-prosodic variables, as well as transcriber-specific strategies in judging prominence.

3.1. Inter-transcriber agreement

The agreement among the 28 untrained annotators reached an intermediate value (Fleiss’ kappa $\kappa=0.53$; cf. Fleiss 1971, Geertszen 2012). This can be regarded as an acceptable basis for further analyses, since it shows that transcriber-specific variation was moderate.

3.2. Prosodic factors

3.2.1. Accent type and position

Fig.4 shows the interaction between accent type and position indicating that high and especially rising utterance-final nuclear accents received the highest p-scores. In comparison with (high and) rising accents, the scores for low accents are considerably lower in both nuclear and prenuclear position. Falling accents, however, were judged as more prominent than high accents in nuclear non-final and in prenuclear position.
This finding suggests a crucial role of tonal movement for the given task. Support for this view comes from the fact that subjects hardly recognised phrase accents as prominent, which generally lack a pronounced pitch movement. Interestingly, but as a plausible consequence, in utterances without perceptually salient accents many listeners did not mark any word as prominent. The strong influence of tonal movements also becomes obvious in the example illustrated in Fig.5. Here, both prenuclear rising accents on Bekannt (‘friend’) and gute (‘good’) trigger high p-scores, whereas the nuclear but low accent on Empfehlung (‘recommendation’) is not perceived as prominent.

3.2.2. Duration

Results for duration measures of vowels in all stressed syllables show the tendency for a linear correlation with prominence judgements: the longer the lexically stressed syllable of a given word, the more prominent the word is perceived. The Pearson's correlation coefficient was found to be slightly higher for phonologically long vowels ($r = 0.40**$), see scatter plot in Fig.6 than for phonologically short vowels ($r = 0.31**$).

3.2.3. Intensity

A similar tendency as for duration could be observed for intensity: the louder the peak intensity (RMS) of a lexically stressed syllable, the higher the p-score. Results reveal a moderate positive correlation between the two variables (Pearson’s $r = 0.42**$).

3.3. Non-prosodic factors

3.3.1. Part-of-speech

A word's part-of-speech (POS) was found to influence prominence perception as well. Highest p-scores were assigned to content words, in particular nouns and adjectives. Fig.7 shows the mean prominence values for POS in the experiment, following the categorisation of the Stuttgart Tübingen TagSet (STTS; Schiller et al. 1999).

However, the prominence of different POS was not judged independently of accentuation. Thus, if function words such as pronouns or particles (also modal verbs) carried a (nuclear) pitch accent in our data, this function word was perceived as particularly prominent (cf. Calhoun 2010a and discussion).

3.3.2. Word frequency

As to word frequency, results show the expected tendency for a negative correlation: the less frequent a word is (we used the SubtLex corpus as a reference) the more prominent it is perceived (Pearson’s $r = -0.33**$).

3.3.3. Morpho-syntactic focus marking

The test material contained words carrying phrase accents and nuclear accents which followed the focus markers nur (‘only’) and sogar (‘even’). An example of a second occurrence focus marked by a phrase accent on Bahber is the construction Auch eine Bachblütenkur kann nur Doktor Bahber machen ‘Also a cure with Bach flowers can only be done by Dr. Bahber’. Both phrase accented and nuclear accented words were judged as more prominent if they were preceded by a focus particle, as illustrated in Fig.8.
4. Discussion and conclusion

Our results suggest that untrained listeners base their judgements primarily on prosodic features, in particular F0 movement and alignment, encoded here in different accent types (rising accents being rated as most prominent). The position of an accent in the phrase also plays a role, as well as – to a small degree – duration and intensity, which are not fully independent of accent type and position, though (cf. Figs. 2 and 3). Secondary accents due to their prenuclear position are perceived as less prominent than nuclear accents, but their p-score crucially depends on the type of pitch movement – since both nuclear and prenuclear accents are fully-fledged pitch accents. In contrast, postnuclear secondary accents (phrase accents) are perceived as less prominent due to their lack of tonal movement. In general, bottom-up acoustic features seem to attract most attention in untrained listeners when judging prominence.

Nevertheless, prominence perception also is to some extent based on non-prosodic factors related to top-down knowledge, or rather intuition, about a word’s accentability. This knowledge triggers language-specific expectations which are structural or conceptual in nature, e.g. expectations on the appropriate metrical and information structure, on the alignment of the nucleus with the last argument near the right edge of an intonation phrase in German, or on the correspondence between semantic weight and perceived prominence, e.g. in terms of POS. Our data only show few hints at expectation-matching prominence ratings, such as the general prominence-ranking of POS and a slight effect of focus particles. A stronger influence on the p-scores could be observed in cases of mismatch with a listener’s expectations. Mismatches attract attention, demanding increased processing effort (Cole et al. 2010). They may either be triggered by the occurrence of unpredictable words (low frequency or non-derivable from the discourse) or by the difference between expected prominence (due to structural knowledge of an utterance’s metrical structure including the position of the nuclear accent) and evaluated prominence. This difference affects the interpretation of an utterance’s information structure (see Calhoun 2010a,b). A number of nuclear accents on function words in the test material which were perceived as highly prominent serve as evidence for this effect.

The results generally confirm the dual nature of prosody as reflecting the interdependency between acoustic patterns on the one hand and higher-level structures that account for these patterns on the other (Shattuck-Hufnagel and Turk 1996:196, Smith 2013:6). The relevance of both aspects for the perception of prominence and phrasing has also been confirmed to a large extent by the RPT data by Cole et al. (2010) on American English. However, it is important to point out that the wording of the instructions may strongly influence the subjects’ annotation of prosody. Smith (2013) showed in an RPT study on French that more prominences were marked under ‘meaning-based’ instructions than under ‘acoustically-based’ instructions. Our instructions were deliberately designed to leave both interpretations open, giving the options ‘important’ (meaning-based) next to ‘stressed’ and ‘highlighted’ (rather acoustically-based). Given the outcome of our study, the instructions may have been interpreted by most subjects as acoustically-based.

Finally, the data suggest that the main difference between trained and untrained listeners may be that the former attribute a special status to the nucleus (defined as the structurally most prominent and communicatively most important element in most common intonation theories), whereas the latter do not, at least not to the same extent. Instead, they determine the location of a prominent word according to their perception of prosodic, especially tonal, salience.

5. References


The effect of thalamic deep brain stimulation on speech production in subjects with essential tremor

Johannes Becker¹, Doris Mücke², Martine Grice², Jil Pochmann¹, Lars Timmermann¹, Michael T. Barbe¹,³

¹Department of Neurology, University Hospital Cologne, Germany
²IfL Phonetik, University of Cologne, Germany
³Institute for Neuroscience and Medicine, INM-3, Forschungszentrum Jülich, Germany

{doris.muecke, martine.grice}@uni-koeln.de
{johannes.becker, jil.pochmann, lars.timmermann, michael.barbe}@uk-koeln.de

Abstract
In subjects with essential tremor (ET) treated with chronic deep brain stimulation (DBS) of the nucleus ventralis intermedius (VIM) we investigated temporal- and intensity-related acoustic parameters to quantify DBS induced deterioration in motor speech performance. We found higher overall syllable durations (reflecting a decrease in articulation rates) as well as higher intensity ratios (reflecting an increase in intervocalic voicing and/or a reduction in the degree of constriction during stop consonant production) in on-DBS condition.

Keywords: deep brain stimulation, essential tremor, nucleus ventralis intermedius, dysarthria, acoustic analysis of speech

1. Introduction
In this study we investigate the effect of chronic deep brain stimulation (DBS) of the nucleus ventralis intermedius (VIM) on the production of speech in subjects with essential tremor (ET). VIM-DBS is performed to suppress medically resistant tremor, especially for essential tremor and tremor-dominant Parkinson’s disease (Benabid et al. 1996). VIM is regarded as a relay station in the tremor network connecting cerebellum and motor cortex (Schnitzler et al. 2009) and is therefore the classical neuroanatomical target for DBS in essential tremor. VIM-DBS is a highly effective treatment (Flora et al. 2010) and usually leads to a tremor reduction of 60-80% (Benabid et al. 1996). However, stimulation induced motor speech impairment is a common side effect of thalamic stimulation (Flora et al. 2010, Krack et al. 2002) and, as a result, the extent to which tremor can be suppressed in individuals with essential tremor is limited, since suboptimal parameter settings just below the threshold inducing dysarthria have to be selected by the treating clinician. Previous analyses in controlled essential tremor studies have indeed demonstrated that dysarthria is the most significant adverse event, affecting individuals with essential tremor (values reported in the literature range from almost 9% found in Flora et al. 2010, to 75% found in Pahwa et al. 2006) and inducing a severe impact on quality of life and social functioning. In a recent study (Mücke et al., accepted) we used temporal acoustic parameters to quantify stimulation induced speech deficits in subjects with ET performing oral diadochokinesis (rapid syllable repetition) tasks involving alternation of voiceless stop consonants and vowels. We found both an impairment of glottal control (decrease of voiceless intervals) and oral articulation (incomplete closure) under VIM-DBS.

Abstract
The aim of the present study is to confirm our earlier results with a different cohort of subjects with ET, and to extend our analysis to intensity-related acoustic parameters which have been used to quantify imprecise consonant articulation in DBS induced dysarthria in individuals with Parkinson’s disease (e.g. Dromey & Bjarnason 2011). Furthermore, we explore the feasibility of developing these measures for use as predictions of postoperative speech performance outcome.

2. Methods

2.1. Participants
We recorded sixteen native speaking German subjects with ET between 39 and 77 years (mean 65.13, SD 10.98; 11 male, 5 female). All of them had been bilaterally implanted with a DBS system in the VIM at least three months before the recordings took place. None of the individuals with essential tremor was diagnosed with pre-operative dysarthria.

2.2. Recordings
For each of the subjects, two separate recording sessions were carried out (on- and off-DBS condition). By using an AKG C420 condenser headset microphone, a mouth-to-microphone distance of approximately 5 cm could be kept constant across sessions and independently of the patient’s head movement. The acoustic signal was digitized by a Focusrite Scarlett 2i2 USB recording interface at 44.1 kHz/16bit and preprocessed using a high-pass filter in order to remove noise below 80 Hz. All recordings took place at the Department of Neurology of the University Hospital Cologne.

2.3. Speech materials
The speech material consisted of an oral diadochokinesis task in which speakers were instructed to produce the consonant-vowel sequence /kakaka/ as quickly and as often as possible on one single breath. The task was demonstrated by the examiner prior to the beginning of the first recording. We used velar oral diadochokinetic tasks /kakaka/, which have been shown to reliably reflect impairments of glottal and articulatory control in our earlier study (Mücke et al., accepted). Figure 1 shows the acoustic waveform, spectrogram and intensity curve of three subsequent syllable repetitions produced by a subject with ET in off-DBS and on-DBS conditions.
2.4. Labelling and measures

For analysis, we selected a sequence of ten consonant-vowel-syllables. To avoid durational effects associated with prosodic boundaries (such as utterance initial/final strengthening and lengthening, which have been shown to mark edges of prosodic domains, see e.g. Fougeron & Keating 1997), the first and last three instances within a sequence were excluded from the analysis. A total of 320 syllables (16 speakers x 10 syllable repetitions x 2 DBS states) went into the analysis. All data were labeled by hand in PRAAT (Boersma and Weenink 2010), using the speech waveform and a wide-band spectrogram. In order to capture the extent of motor speech impairment, we measured two acoustic parameters:

(a) Syllable durations (ms): Syllable duration is a measure related to the overall articulation rate (AR). Slowing down of AR has been described as an indicator for various forms of motor speech impairment (Ackermann et al. 1995).

(b) Intensity ratio (%): With this measure, the minimum intensity during the consonant is relativized to the maximum intensity of the following vowel. Intensity ratio has been shown to be a reliable acoustic measure for intervocalic voicing as well as for constriction degree during stop consonant production (Kaplan 2010, Parrell 2010). Periodic and aperiodic spectral components due to ongoing vocal fold vibrations and/or leaking oral closures lead to an increase in energy during the closure phase and concurrently increase the intensity ratio between the stop consonant and the vowel.

2.5. Statistics

Statistical analysis was performed using PASW Statistics 21. To detect significant differences between off-DBS and on-DBS conditions, we used paired-samples t-tests. Post-hoc nonparametric analysis (Wilcoxon signed rank tests) was performed when data did not match the assumptions for normality tested with the Kolmogorov-Smirnov test. Relationships among parameters themselves were analyzed by using Pearson’s/Spearman’s correlations. To consider multiple testing, p-values were corrected using the Bonferroni-Holm procedure. A value of \( p \leq .05 \) was defined as the level of significance.

2.6. Ethics

This study was approved by the Local Ethics Committee of the University of Cologne. Each subject with essential tremor gave written informed consent before study participation. Research was conducted in accordance with the Declaration of Helsinki.

3. Results

(a) When comparing syllable durations, we found a significant difference between off-DBS and on-DBS (paired-samples t-test; \( t(15) = -3.968, p < .01 \)). In the on-DBS condition, syllable durations increased by 10% (on average, 289ms in on-DBS compared to 262ms in off-DBS condition), reflecting a slower articulation rate under stimulation.

(b) For the intensity ratio measure, we also found a significant difference between off-DBS and on-DBS (Wilcoxon signed-rank test; \( Z = -2.844, p < .01 \)). Intensity ratio increased from 30% in off-DBS to 36% in on-DBS condition. This reflects an increase in acoustic energy during the intended silent gap of the voiceless stop consonant production, due to intervocalic voicing (periodic energy) and/or to an incomplete closure in the oral tract causing frication (aperiodic energy).
While there is a significant overall increase in syllable duration and intensity ratio, figures 2 and 3 suggest that not all individuals are affected by stimulation in the same manner: subjects with ET with a slower overall articulation rate and/or a higher intensity ratio already in the off-DBS condition tend to show a greater deterioration in the on-DBS condition. To quantify this relationship, we conducted a regression analysis between the severity of impairment in the off-DBS condition and the amount of deterioration caused by stimulation for both parameters separately (figure 4 and 5).

We found a significant positive correlation (Pearson; \( r = .516, p < .05 \)) between raw syllable durations in off-DBS condition and the amount of increase in syllable duration from off-DBS to on-DBS condition. Likewise, the positive correlation between intensity ratio in off-DBS condition and the amount of increase in intensity ratio from off-DBS to on-DBS condition turned out to be significant (Spearman; \( r = .691, p < .01 \)).

That is, those individuals with more severe impairments in speech performance in the off-DBS condition are more greatly affected by stimulation than those with no or less severe speech impairments.

4. Conclusion

In the present study, we used acoustic parameters reflecting a deterioration in speech performance of subjects with ET under chronic deep brain stimulation. In the temporal domain, we found that articulation rate slowed down under stimulation. This goes in the direction of Mücke et al. (accepted) who also found a tendency for syllables to be longer in duration under stimulation although their results did not reach significance.

In the spatial domain, we found an increase in intensity ratio between consonants and vowels under stimulation. This indicates imprecise articulation in terms of voicing and/or incomplete oral closures.

Furthermore, our results point to an important role of speaker-specific traits. Individuals with more severe impairments regarding speech performance in off-DBS condition are more greatly affected by stimulation than those with no or less severe speech impairments. Thus, these measures can be used to preoperatively identify a patient’s risk of future dysarthria as a severe adverse event under VIM-DBS.

5. Acknowledgements

We would like to thank the patients for participating in our study and for cooperation.

6. References


Disfluencies in Spontaneous Narratives and Conversations in Hungarian
András Beke, Mária Gósy, Viktória Horváth, Dorottya Gyarmathy, Tilda Neuberger

Department of Phonetics, Research Institute for Linguistics, HAS, Hungary
beke.andras@nytud.mta.hu, gosy.maria@nytud.mta.hu,
gyarmathy.dorottya@nytud.mta.hu, horvath.viktoria@nytud.mta.hu,
tilda.neuberger@nytud.mta.hu

Abstract

Disfluencies refer to monitoring and repairing errors in speech production processes that either occur on the surface (like false starts) or do not occur but are indicated by, for example, filled pauses. Although there are various factors that can trigger disfluencies, their effects might not be uniform in various speech styles. Speakers use diverse strategies to overcome their speech planning problems. They apply the strategies that have been found to be successful earlier on. Therefore these strategies may be connected with a specific speech style. Our hypothesis is that frequent disfluencies would show different patterns in the two presently analyzed speech styles. We analyzed eight types of disfluencies produced by 50 Hungarian-speaking young adults in narratives and conversations. Results show that the effect of speech style is more decisive in the occurrence of disfluencies than that of their types are. These findings can be applied as an additional factor in automatic speech style identification.

Keywords: disfluency types, speaker-specific patterns, speech style effects

1. Introduction

Various types of disfluencies – filled pauses, repetitions, fillers, false starts, etc. – received increasing attention in the past decades (e.g., Schriberg 1992, Nootbeoom, Clark and Wasow 1998). This direction of research is closely connected with the interest of representatives of speech-related fields in spontaneous speech. Speaking spontaneously means the quasi-parallel (or simultaneous) operations of speech planning and execution (Levelt 1989) that might result in disharmony at various levels of the mechanism. The consequences of such disharmony appear in various types of disfluencies that can be traced in spontaneous speech samples as early as from the age of three (Hudson and Carla 2008). Disfluencies refer to monitoring and repairing real or assumed errors in speech production processes that either occur on the surface (like false starts) or do not occur but are indicated by, for example, filled pauses. Disfluencies that are produced in narratives and conversations might be characteristic of the speech styles themselves (e.g., Shriberg 2001, Bortfeld et al. 2001, Mooshammer et al. 2008). Although there are various factors that can trigger disfluencies, their effects might not be uniform in various speech styles. Speakers use diverse strategies to overcome their speech planning problems. They apply the strategies that have been found to be successful earlier on. The different communication situations, the participants of the conversations and the topic to be discussed may lead to different frequency patterns of disfluencies. Therefore these strategies may be connected with a specific speech style. The aim of this study is to describe the disfluency patterns characteristic of narratives and conversations. The questions raised here are whether (i) it is the speech style or the disfluency type that influences the occurrences of disfluencies to a larger extent, and (ii) the disfluencies speakers produce can be used in automatic classification of the two speech styles. Our hypothesis is that frequent disfluencies would show different patterns in the two analyzed speech styles. Therefore we analyzed the occurrences of disfluencies that turned out to be frequent in our material – filled pauses, word repetitions, restarts, filler words, false starts, false words, and anticipations – and were produced by the same speakers in narratives and conversations.

2. Subjects, material, method

50 Hungarian-speaking young subjects (25 females and 25 males, aged between 20 and 32) were randomly selected from the BEA Spontaneous Speech Database of Hungarian (Gósy 2012). 50 narratives (10 hours) and 50 conversations (7 hours) elicited from the same speakers were used in this study. ‘Speaking time’ in conversations refers to speech samples that were produced by the same subjects whose narratives were used. The speech material was manually annotated at the segmental level by two of the authors while another two of them double-checked the annotations including marked disfluencies in the speech flow. The occurrence of 7 types of disfluency was measured: filled pauses, word repetitions, restarts, filler words, false starts, false words, and anticipations. For the processing of disfluency phenomena we used software specifically written for this task. 5,336 instances of disfluency were found in the conversations and 6,957 instances were found in the narratives. The occurrences and the types of disfluencies were analyzed across speech styles on the one hand, and the automatic classification of the two spontaneous speech styles were carried out using the Wilcoxon signed-rank test (with Monte Carlo simulation) for statistical examinations, as well as Fisher Linear Discriminant analysis were used for automatic classification of the speech styles.

3. Results

Seven types of disfluency were selected for this study that were the most frequent ones in our speech samples. Four of them belonged to the category that is usually labeled by the term ‘hesitations’; they were restarts, repetitions, fillers and filled pauses. The other three types of disfluencies were lexical and sublexical errors: anticipations, false starts and false word activations. Since there is no total agreement upon disfluency types and their names in the literature, we briefly define the types we used. Filled pauses are produced in the majority of cases by ñ- like (close to neutral vowel) and m-like sounds in Hungarian that can easily be differentiated from meaningful sound sequences (e.g., volt például ñ ñ egy tükör ‘there was, for example, ñ a mirror’). Filler words were identified when neither the grammatical structure nor the semantic context required the actual word uttered (e.g., meglátjuk hogy mi történik tehát hogy ‘we will see what happens well that’). Unintended repetitions were carefully identified considering also the
Narratives and conversations are characterized by different patterns of distribution of disfluency types. Looking at the details, we found that practically no difference was found in proportions of anticipations, and false starts within each speech style. Filled pauses and false words seem to be more characteristic of narratives while word repetitions, fillers and restarts are characteristic of conversations to a larger extent than they are of narratives. Fillers, filled pauses and word repetitions occupy the largest ratios among disfluency types. Filled pauses indicate a decisive ratio in the pattern of narratives while fillers and word repetitions are more characteristic of the pattern of conversations.

### 3.1 Occurrences

Although the most frequent disfluency types were selected for this study, their real occurrences show large differences among various types (see Shriberg 2001, Eklund 2004). Table 1 shows the frequency characteristics of the eight types of disfluencies produced by 50 speakers in their spontaneous speech materials irrespective of the speech style. As expected, fillers and filled pauses turned out to be the most frequent disfluencies. Fillers tehát (‘that is’), akkor (‘so’), and hát (‘well’) were the most frequent ones in all speakers’ speech materials while ó-like sounds with various durations were preferred in the case of filled pauses (see also Horváth 2010). The next most frequent hesitation phenomenon was word repetition used to gain extra time for speech planning processes. Activation of the mental lexicon resulted in two types of speech errors, false starts and false words; however, their occurrences were much rarer than those of the hesitation phenomena mentioned above.

#### Table 1: Occurrences of disfluencies in the spontaneous speech material.

<table>
<thead>
<tr>
<th>Types of disfluency</th>
<th>Frequency (occurrences/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
</tr>
<tr>
<td>anticipation</td>
<td>0.20</td>
</tr>
<tr>
<td>filled pause</td>
<td>3.66</td>
</tr>
<tr>
<td>word repetition</td>
<td>2.24</td>
</tr>
<tr>
<td>false start</td>
<td>0.65</td>
</tr>
<tr>
<td>false word</td>
<td>0.24</td>
</tr>
<tr>
<td>filler</td>
<td>4.41</td>
</tr>
<tr>
<td>restart</td>
<td>0.61</td>
</tr>
</tbody>
</table>

#### 3.2 Speech style

Disfluencies were analyzed depending on the two speech styles. Results show that there are four types of disfluencies (filler, word repetition, false start, restart and) occurring more frequently in conversations while the remaining three (filled pause, false word and anticipation) occurred more frequently in narratives. Fillers seem to be a better strategy for speakers to overcome their difficulties when they are supposed to speak for a relatively short time. On the contrary, filled pauses were preferred when speakers were forced to speak long in the narratives. Word repetition seems to be a good strategy in conversation to gain time for thinking and formulation of utterances on the one hand, and to signal to participants that the speaker wants to hold the floor, on the other hand. This latter function of word repetition is unnecessary in narratives. Restarts show larger differences between narratives and conversations than do false starts but both are more characteristic of conversation than of narrative. Although false words were not too frequent in our spontaneous speech material, they occurred more frequently in narratives. There
were no large differences in the occurrences of anticipations between narratives and conversations; however, anticipations were more frequent in narratives and less frequent in conversations (Fig. 2).

Figure 2: Occurrences of disfluency types depending on speech style (rep. = word repetition, ant. = anticipation, rest. = restart)

In narratives, 10.2 occurrences of all hesitation phenomena (std. dev.: 5.02) and 1.46 occurrences of lexical and sublexical errors per minute (std. dev.: 1.71) were found while 11.6 instances of hesitation phenomena (std. dev.: 5.32) and 1.47 instances of lexical and sublexical errors per minute (std. dev.: 2.01) occurred in conversations. Statistical results supported that speech style had an important effect on the occurrence of disfluencies (General Linear Model: $F(1, 799)=339.827$, $p=0.001$). However, no such effect was found for the factor of disfluency type, and there was no statistically significant result concerning their interaction, either. Speech style explains the occurrences of disfluencies in a relatively large proportion, in 63% (according to Partial Eta Squared).

Various disfluency types were analyzed in pairs depending on speech style. Results show that significant differences were found in five types (filled pause, word repetition, false word, filler and restart) in their occurrences between narratives and conversations (Table 2.). Fisher Linear Discriminant Analysis (FLDA) revealed that there were three types of disfluency that were statistically significant depending on speech style (Table 1). The data suggest that restarts, filler words and false words (respectively) had the greatest discriminant power effect to discriminate speech style.

Table 2: Statistical results of the FLDA

<table>
<thead>
<tr>
<th>Disfluency types</th>
<th>Tests of Equality of Group Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wilks’ Lambda</td>
</tr>
<tr>
<td>anticipation</td>
<td>0.986</td>
</tr>
<tr>
<td>word repetition</td>
<td>0.971</td>
</tr>
<tr>
<td>false word</td>
<td>0.953</td>
</tr>
<tr>
<td>restart</td>
<td>0.930</td>
</tr>
<tr>
<td>filled pause</td>
<td>0.971</td>
</tr>
<tr>
<td>false start</td>
<td>0.994</td>
</tr>
<tr>
<td>filler word</td>
<td>0.946</td>
</tr>
</tbody>
</table>

The 3-fold cross-validation method was used to train and test our classifier. Correct automatic classification of narratives and conversations was 60%. To improve the system accuracy we applied dimension reduction using PCA (Principal Component Analysis). The 7-dimension data (number of dimension is the number of disfluency) were decreased by PCA to 3-dimension data, and the FLDA was re-applied. The classification accuracy can be increased from 60% to 64% using PCA. This means that we can achieve 4% relative improvement. The real improvement concerns the narrative classification by an 8% increase of accuracy. The classification result showed further improvement (by 4% both in narratives and in conversations) when the data had been normalized to the means and ranges within each speaker before they were reduced to 3-dimension data and were classified using LDA. The best accuracy yielded by normalized 3-dimensional feature reduced by PCA was 68%.

3.3 Gender and individual differences

Disfluencies produced by females and males were analyzed irrespective of speech style. Males produced more disfluencies per minute than females did (1.67 instances vs. 1.25 instances within a minute). In addition, instances of five types of disfluencies were more frequent in males than in females. Two of them (anticipation and false start) showed no gender differences in their occurrences. However, there was only one type of disfluency, filled pause, where significant difference was found depending on gender ($F(1, 99) = 10.909$, $p = 0.013$), see Fig. 3.

Figure 3: Occurrences of filled pauses depending on gender

The occurrence of disfluencies is not uniform across speakers (see Fig. 4); however, groups of speakers sharing similar frequency of certain disfluency types can easily be found. However, no individual speaker-specific disfluency patterns could be found in our large spontaneous speech material of 50 speakers.
4. Conclusion

Eight of the most frequent disfluency types were analyzed in order to find out whether there is a speech style-specific pattern based on the frequency information of disfluencies. Two speech styles were considered to show their possible effects on the production of disfluencies. As expected, hesitation phenomena surpassed speech errors in both speech styles as had been shown and explained by the monitoring processes of spontaneous speech in a number of previous studies (Levelt 1989; Postma 2000). Fillers, filled pauses and word repetitions occupy the largest ratios among disfluency types. Narratives are characterized particularly by filled pauses while conversations are characterized by both fillers and word repetitions (see Lease et al. 2006). Speech styles significantly influence the occurrences of disfluencies in spontaneous speech which is the consequence of the different speaking tasks in narrative and in conversation (e.g., Shriberg 2005). Gender differences were found in the occurrences of disfluencies. Males used more disfluencies than females did, particularly those indicating uncertainty in their speech (filled pause, filler, word repetition and restart). The occurrence of filled pauses was significantly higher in males than in females. These findings suggest that males seem to be more careful about what to say in order to avoid any misunderstanding on the part of the listener(s).

5. Discussion

Speech style-specific information of certain types of disfluencies seems to be a decisive factor, this information characterizes the analyzed speech style, narratives and conversation. Results of the automatic classification of the speech styles based on disfluency occurrences confirm this statement. We can conclude that (i) speakers’ strategies to overcome their speech planning and execution difficulties seem to follow a universal scheme that is reflected by some types of disfluencies but (ii) disfluencies are not characteristic of the speakers themselves.

6. Acknowledgements

This work was supported by OTKA 108762.

7. References


Articulatory basis of the apical/laminal distinction: Tongue tip/body coordination in the Wubuy 4-way coronal stop contrast

Catherine T. Best1,2, Christian H. Kroos1,3, Rikke L. Bundgaard-Nielsen1, Brett Baker4, Mark Harvey5, Mark Tiede2, and Louis Goldstein2,6

1University of Western Sydney, Australia 2Haskins Laboratories, USA 3Curtin University, Australia 4University of Melbourne, Australia 5University of Newcastle, Australia 6University of Southern California, USA

c.best@uws.edu.au, chkroos@gmail.com, rikkelou@gmail.com, bjbaker@unimelb.edu.au, mark.harvey@newcastle.edu.au, tiede@haskins.yale.edu, louisgold@usc.edu

Abstract

We used electromagnetometry (EMA) to track tongue motion during productions of a rare 4-way coronal stop place series by three native speakers of the Australian language Wubuy, to assess differences in coordination of tongue tip (TT) and tongue body for Wubuy’s apical (alveolar [t], retroflex [?]I) versus laminal (dental [t], post-alveolar [c]) stops. Using two novel measures of TT-TB range and correlation of motion, we found support for our three initial hypotheses: 1) TB is relatively stable during TT gestures for apicals while TT-TB are coupled for production of laminals; 2) The anterior vs. posterior place contrast within each coronal class interacts with the apical-laminal TT/TB difference, which is larger for the posterior than the anterior stops; 3) The TT/TB difference between apicals and laminals is reduced in /i/ contexts (relative to /a/ and /u/) due to the high front TB position needed for /i/. Further work on other manners of coronals and other languages will help identify whether these are universal coronal characteristics (see Derrick, et al., 2014).

Keywords: coronal stop place contrasts, apical vs. laminal, tongue tip-body coordination, Wubuy, Australian languages

1. Introduction

Coronal consonants are those in which the primary articulatory constriction is achieved with the anterior third of the tongue, which includes the tip and blade, e.g., /l/, /j/, /w/, /n/, /l/. Coronal consonants are among the most numerous and varied consonants, across and within languages, due to the tongue tip’s (TT) range and speed of motion. This flexibility is exploited by a generally accepted orientation distinction that is unique to TT among the major articulators used to produce consonantal constrictions (lips, tongue tip, dorsum, and root). In apical coronals the constriction is made with the apex of the tongue tip (TT up), whereas in laminal constrictions it is made with the tongue blade (TT down) (Brownman and Goldstein, 1989; Butcher and Tabain, 2004; Flemming, 2003). In view of the fact that TT gestures are necessarily constrained by tongue body (TB) position, we posited that a TT-TB coordination distinction underlies the apical/laminal subclass distinction within coronals (Best, Bundgaard-Nielsen, Harvey, Baker, Goldstein, Kroos, Mooshammer, and Tiede 2009; Best, Bundgaard-Nielsen, Kroos, Harvey, Baker, Tiede and Goldstein, 2010); TT is relatively stabilized in apicals to support lever-like TT motion, whereas TB and TT move in tandem in laminals to yield thrust-like motion of the tongue (see Figure 1). If the hypothesis is correct, then these two patterns should be optimally clear in language that maintains a 4-way coronal series within a single manner class, such as stops. In 4-way coronal stop series, such as are found in a subset of Australian indigenous languages and Dravidian languages, each coronal subclass contrasts anterior-posterior constriction locations, yielding four distinct tongue tip postures (Butcher, 1993).

Such 4-way contrasts are rare in the world’s languages. Wubuy (Nunggubuyu: ISO 639-3/nuy), an endangered Australian language is one such language. Children no longer acquire it from birth and the current native speaker population is estimated at ~60 speakers, all over 55-60. Wubuy has an anterior vs. posterior contrast within both its apical (alveolar [t], retroflex [?I]), and laminal stops (dental [t], post-alveolar [c]) (Heath, 1984; Ladefoged & Maddieson, 1996). To examine our hypotheses about TT-TB coordination, we conducted an electromagnatic articulometry (EMA) study of Wubuy coronal stop productions. Preliminary examination of the four coronal stops in the /a/ vowel context suggested that TT-TB coordination differ as posited between apicals vs. laminals (Best et al., 2009, 2010). That report had provided descriptive observations of TT vs. TB motion and estimates of mid-sagittal tongue shape at gesture onset (GONS), Mid-constriction (Mid-C) and offset (GOFF) of the TT gestures.

Figure 1: Wubuy coronal stops schematized in a 2 x 2 matrix for apical vs. laminal closure x anterior vs. posterior release. Here the posterior stops (apical retroflex [?I]; laminal post-alveolar [c]) are represented with differing tongue position at onset of closure (solid black line) than at onset of release (dashed gray), and anterior stops (apical alveolar [t]; laminal dental [t]) with the same position at closure and release (solid black only). Figures adapted from a modeling study on the Wubuy EMA data for the aCa context (Proctor, Bundgaard-Nielsen, Best, Goldstein, Kroos & Harvey 2010).

2. The present EMA investigation

For the present report, we further examined the Wubuy EMA data to quantitatively assess the posited TT-TB coordination differences between apical and laminal coronal stops. We developed two novel indices of TT-TB coordination for this
purpose, which we describe in Measures of TT/TB coordination. One is a spatial index, capturing the range of motion transcribed by TB relative to that by TT during the TT constriction gesture for the coronal stops. The other addresses the dynamic temporal relationship between TT and TB movements, indexing the short-term average correlation between their velocity profiles during the TT gesture. In the present study, we also examined all three vowel contexts (/u/Ca/, /u/Cu/, /iCi/) that had been recorded. This allowed examining how the high front TB position for /i/, which more greatly constrains the position for TB leading into and out of the TT coronal gesture than does /a/ and especially /a/, affects the TT-TB coordination for apicals versus laminals.

2.1. Hypothesis 1: TB is relatively stable during TT gestures for apicals, while TT-TB are coupled for laminals, predicting a higher ratio of TT/TB range of motion for apicals than laminals, but a higher correlation between TT and TB velocity profiles for laminals than apicals, during the TT gesture.

2.2. Hypothesis 2: Although the anterior/posterior location distinction within apicals and laminals is not expected to show a main effect on either TT-TB measure, articulatory differences between anterior and posterior coronals will affect the size of the apical-laminal distinction in TT-TB coordination.

2.3. Hypothesis 3: The high front TB position for /i/, and possibly the high back TB position for /a/, will constrain TB motion in those vowel two contexts, relative to /a/. Thus, the apical-laminal difference in TT-TB coordination is likely to be smaller in /i/ and possibly /u/ contexts, relative to /a/ contexts.

3. Method

3.1. Participants
The participants were three female native speakers of Wubuy (ages at recording = 51-61 years), who had been born and raised in Numbulwar, Arnhem Land, Northern Territory, Australia. All were literate in Wubuy, had some basic linguistic training, and were involved in community language revitalization efforts. Two also spoke the Australian language Anindilyakwa with relatives other than their parents, and all spoke English. One (W1) had used English as a primary language since young adulthood, but also continued to use Wubuy on a daily basis with other speakers in the community. Wubuy community, and sometimes with her children and grandchildren.

3.2. Target stimuli
They produced 5 tokens each of /t/, /d/, /g/, and /c/ in Wubuy words containing /aCa/, /uCu/ and /iCi/ (as part of a larger study). Targets were produced within the Wubuy carrier phrase “nga-yamana _____ adaba” (“/na-ja mana _____ aฑaba”), which means “I say _____ now.” The participants discussed the selected words prior to the EMA task, affirming that all were words and known to them (see Bundgaard-Nielsen et al., 2012, for word lists and acoustic analyses of the target coronals).

3.3. Procedure
Articulatory data were recorded with a Carstens AG500 electromagnetic articulograph (EMA) at 200 Hz sampling frequency. Separate high-quality acoustic recordings were made concurrently using two Shoeps CMC6 microphones with a highly directional MK41 capsule. For tracking the tongue movements with EMA, three small wired magnetic-coil EMA sensors were attached mid-sagittally on the surface of the tongue with an approved dental adhesive, one ~1 cm posterior to the tongue tip (TT), one at the anterior tongue dorsum, and one at the most posterior position feasible without producing discomfort (e.g., gag reflex) which was designated as the tongue back (TB); a fourth sensor was attached para-sagittally on the tongue blade (see Figure 2). The sensors of interest for current analyses were TT and TB. Additional sensors were located at the vermillion border at the middle of the upper and lower lips to track lip motion, and at the gum line of the lower jaw for jaw motions. Reference sensors were attached at the upper incisor, the nasion and the left mastoid, for later head motion correction of the data.

Figure 2. Configuration of the EMA tongue sensor locations

The participant then sat in a non-metallic chair positioned within the open plexiglass cube of the Carstens AG500 device that contains the electromagnetic transmitters generating the magnetic field in which the sensor motions are tracked during speaking. During the session, the target words were presented in the carrier sentence on a computer monitor in Wubuy text, and the speaker produced five good tokens of each target (as judged by herself and the other speakers).

3.4. EMA data processing
The raw EMA data were converted to time-varying positional values using TAPAD-M, and were corrected for head motion using specially-developed software that utilized the returned orientation angles of the three reference sensors listed above. The FindGest function in MVIEW, a visualization application for viewing sensor motions in EMA data (Tiede, 2005, 2010), was then used to semi-automatically locate the key landmarks of the TT constriction gesture of the target coronals (see Figure 3): PVEL1 = peak velocity during closure gesture;
GONS = onset of closure gesture (20% threshold of PVEL1); NONS = nucleus onset (next 20% threshold of PVEL1: constriction onset); PVEL2 = peak velocity during release gesture; NOFF = nucleus offset (20% threshold of PVEL2); Mid-C = midpoint of NONS→NOFF closure period; GOFF = gesture offset (20% threshold of PVEL2 at offset of release).

4. Measures of TT/TB coordination

4.1. Quad-Ratio: TT/TB spatial displacement

If as posited TB is relatively stabilized during the TT gesture for apicals, but moves more in tandem with the TT for laminal gestures, then the TB should display less range of motion than TT in apicals, but TB and TT should show more equivalent ranges of motion in laminals. To index this coronal class spatial difference, we calculated the mid-sagittal quadrilateral areas transcribed by TT sensor, and by the TB sensor, at four key points in the TT gesture, for each target coronal: GONS, NONS, NOFF and GOFF (schematized in Figure 4). To normalize for speaker differences, the log ratio of the TT/TB quadrilateral areas, or Quad-Ratio, was calculated per token.

4.2. Velocity Correlation: TT/TB temporal dynamics

The proposed TT/TB coordination differences between apical and laminal gestures should also be reflected in the dynamic relationship between the velocity profiles of TT and TB over the course of the TT gesture. If as we propose the TB moves more in tandem with the TT during the TT gesture for laminal gestures, but is relatively stabilized during the TT gesture for apicals, then short-term velocity changes in TT and TB across the TT gesture should be more highly correlated in laminals than in apicals. To index this dynamic difference, for each target token a 15-sample (75 ms) sliding window was shifted sample-wise over the TT and TB velocity signals, from GONS through to GOFF of the TT gesture. To minimize undue influence from windows in which neither sensor was moving (velocities near 0), the short-term correlations were multiplied by the mean velocity of TT and TB within that window. This emphasizes faster and more synchronous TT-TB motions over synchrony of little-to-no motion of either sensor. The mean-velocity-weighted short-term correlations during the TT gesture were then averaged to yield one mean Vel-Corr value per token (schematicized in Figure 5).

5. Results

Generalized Linear Model (GLM) analyses were first run on the tokens (15) per coronal stop (4) per vowel context (3), for the factors coronal Class (apical, laminal) x Location of constriction (anterior, posterior) x Vowel context (a, u, i), separately for the Quad-ratio and Vel-Corr scores.

Hypothesis 1, our primary premise of TT/TB coordination differences between the apicals and laminals was supported by a Class effect for both measures. Quad-ratio showed a larger log ratio of TT/TB area of motion (higher median values) for apicals than laminals [F(1,14) = 28.19, p < .0001], as predicted. Also as predicted, Vel-corr TT/TB short-term velocity correlations were greater (higher medians) for laminals than for apicals.

![Figure 4: Schematic diagram of TT/TB areal range of motion index (Quad-Ratio) described above.](image)

![Figure 5: Schematic diagram of the short-term TT-TB velocity correlation index (Vel-Corr) described above.](image)

![Figure 6: Boxplots of Quad-ratio (left panel) and Vel-Corr (right panel) values by Coronal Class (apical/laminal) x Constriction Location within each Class (anterior on left, posterior on right). Box height indicates interquartile range; inner band shows the 2nd quartile (median); whiskers mark 1.5 * interquartile range. Small circles are outlier tokens.](image)

1. Problematic TT or TB data affected 7 tokens (of 180; <4%), never more than 1 per coronal x vowel cell for a given speaker. We replaced those cases by the mean good values for that cell by that speaker.

2. Studies on endangered languages often involve access to only a small number of speakers, and constraints on number of tokens per target that can be elicited. Three Wubay speakers agreed to this study (5% of the population), and we could elicit only 5 tokens of these targets within the larger EMA session. Thus, we used tokens as the random term for these main analyses. Given individual variation, this should provide a conservative test of effects. Speaker was used as a factor in further GLM tests; 2-way interactions with Class, with Location and with Vowel revealed speaker differences in magnitude but not pattern of Quad-ratio and Vel-corr scores for Class (smaller effect for W1), for Location (larger effect for W3), and for Vowel (W2 showed a different pattern for /a/ vs. /a/, but not /i/, contexts than W1 and W3).
Hypothesis 2 was that anterior/posterior Location would fail to show a TT/TB coordination difference, but might affect apical and/or laminal TT/TB coordination patterns. Indeed, neither measure showed a Location effect, but both Quad-ratio \( F(1,14) = 12.36, p < .003 \) and Vel-corr \( F(1,14) = 6.7, p < .021 \) displayed a Class \times Location interaction that suggests the TT/TB difference for apicals vs. laminals is accentuated in the posterior relative to the anterior (see Figure 6). Intriguingly, it is the posterior that show a tongue posture change between onset and release of closure (NONSTOP; see Figure 1).

Further work on other types of coronal consonants and other languages will help determine whether these are universal aspects of coronals, or are specific to languages with phonological systems using multiple coronal place distinctions within a given manner class, such as Wubuy. It will be important to confirm whether these patterns arise in the 4-way coronal stop series of other languages, e.g., certain Dravidian languages, as well as whether they are evident in the coronal place contrasts of other manner classes (e.g., nasals, liquids) in these languages. Regarding the universality/language-specificity issue more broadly, studies will be needed to determine whether apical and laminal coronals display these TT/TB differences and interactions even in languages lacking coronal place contrasts within any manner class, such as English. We report findings on English coronals that suggest both universal and language-specific effects (Derrick, Fiaisson & Best, 2014).

Acknowledgements

We are grateful to the Wubuy participants, Didamain Uibo, Leonie Murrungun and Jangu Nundhiribala, and Jijla Turner, for their invaluable contributions as experts on the language and as native speakers, and to their families, especially Mick Uibo. Thanks to Donald Derrick for preparing the figures. The Universities of Western Sydney (UWS), Newcastle, and New England provided funding for the project and analyses. The UWS Human Research Ethics Committee approved the study.

8. References


Enhanced area functions for noise source modeling in the vocal tract

Peter Birkholz

Department of Phoniatrics, Pedaudiology, and Communication Disorders
University Hospital Aachen and RWTH Aachen University
peterbirkholz@gmx.de

Abstract

The synthesis of natural-sounding fricatives based on a source-filter model is still a major challenge. While the filter is effectively modeled in terms of the vocal tract area function, noise sources are more difficult to model. Source properties critically depend on aerodynamic conditions and vocal tract geometry near the constriction in a way that is not fully understood. Therefore, noise source models usually assume different relations between noise source parameters and the aerodynamic state for different places of articulation. However, the place of articulation cannot be reliably determined from a dynamically changing area function, as in articulatory synthesis. Here we introduce the concept of an enhanced area function that adds the identity of the articulator that confines the vocal tract at the anterior-inferior side as a new layer of information to the classic area function. This allows to distinguish places of articulation and is therefore an effective representation of the vocal tract not only in terms of filter function but also for noise source modeling. A noise source model on this basis is presented.

Keywords: Area function, noise sources, fricatives, place of articulation

1. Introduction

Fricative production for articulatory speech synthesis is mostly modeled as a source-filter process (e.g., Shadle 1991; Narayanan and Alwan 2000). Hence, synthesis of fricatives requires the specification of source and filter parameters. The filter is usually specified in terms of the vocal tract area function \( A(x) \), which describes the cross-sectional area \( A \) of the vocal tract normal to the longitudinal dimension \( x \). The area function is an appropriate abstraction of the complex 3D shape of the vocal tract with regard to its acoustic filter effect for the relevant frequencies of up to about 4-5 kHz.

The more difficult problem is the prediction of the noise source parameters, namely the position, amplitude, and spectral shape of potential sources. They critically depend both on the aerodynamic conditions and vocal tract geometry in the vicinity of the turbulent jet, which may substantially differ from one fricative to the other (Shadle 1991; Ramsay and Shadle 2006). While it is not clear, which aspects of the complex shape of the vocal tract are relevant for this, it is sure that the abstract geometric information contained in the area function is not sufficient for predicting source parameters (Shadle et al. 2008). Therefore, the typical approach is to presume different noise source characteristics for different places of articulation. For example, Shadle (1991) identified significant differences between the noise source characteristics for /ʃ/ on the one hand, and for /s/, /s/ on the other hand. She termed the corresponding sources obstacle and wall sources, respectively, “to indicate a critical difference in the geometry presented to the turbulent jet downstream of the constriction”.

To account for such differences in modeling experiments, Badin, Mawass, and Castelli (1995) modeled the noise source amplitude \( L \) (and analogously the spectral tilt) of fricatives with the general equation

\[
L = k \cdot A^a \cdot \Delta p^b\]

where \( A \) is the cross-sectional area of the constriction, \( \Delta p \) is the pressure drop across the constriction, and \( k, a, b \) are parameters. The parameter values for the different places of articulation were experimentally determined.

This strategy – to use different values for noise source parameters depending on the place of articulation – currently seems to be the most promising way to synthesize high-quality fricatives. However, it is not straightforward to implement this idea for dynamic articulatory speech synthesis. If we assume the area function as a representation of the vocal tract shape, it is impossible to reliably discriminate places of articulation, and hence to predict position-dependent noise source parameters. For example, an apico-alveolar constriction cannot be distinguished from a labio-dental constriction based on the area function under all circumstances. Also when there are two or more constrictions, it is difficult to tell which of them produces a turbulent jet and hence a noise source. For example, /ʃ/ has two constrictions, one created with the anterior tongue, and one with the incisors (see the area function in Shadle 1991, p. 419), but only the lingual constriction gives rise to a significant turbulent jet. Furthermore, for /s/ and /ʃ/, the major source of noise is usually assumed at the incisors, whose position is not available from the pure area function.

In this paper, we propose the notion of an enhanced area function to represent the vocal tract shape. The purpose of the enhanced area function is to allow to determine the place of articulation of fricatives and so to support the optimal parametrization of noise sources. The basic idea is to add the identity of the articulator or structure that confines the vocal tract at the anterior-inferior side as a new layer of information to the area function. Therefore, each position along the longitudinal dimension of the vocal tract (or each tube section in the case of a discrete area function) is associated with a nominal value that specifies the articulator, i.e., the tongue, the lower incisors, or the lower lip. With the enhanced area function, the primary articulator forming a constriction can be identified and it becomes easily possible to distinguish places of articulation.

The following section describes the extraction of an enhanced area function from a 3D vocal tract model (Birkholz 2013), and Section 3 presents a simple noise source model on this basis.

2. Model-based extraction of the enhanced area function

Area functions are mostly obtained from either MRI or X-ray images of the vocal tract (e.g., Narayanan, Alwan, and...
Haker (1995), or from 2D or 3D models of the vocal tract (e.g., Birkholz 2013). Both the images and the models usually provide the information about the identity of the articulator or structure confining the vocal tract at the anterior-inferior side that is needed for enhanced area functions. Here we describe the calculation of the enhanced area function of the vocal tract model by Birkholz (2013) that is implemented in the articulatory synthesizer VocalTractLab 2.1 (www.vocaltractlab.de).

The first task is to find the center line of the vocal tract. Here, it is calculated in two steps (Birkholz 2005). In the first step, a grid system is superimposed on the vocal tract outline in the midsagittal plane, as shown in Figure 1a. This grid comprises a number of closely spaced horizontal grid lines in the pharyngeal region (sector I), vertical grid lines in the oral region (sector II) and radial grid lines in the velar region (sector II). The most inferior horizontal grid line represents the position of the glottis, and the most anterior vertical grid line represents the vocal tract termination at the mouth. The latter is positioned about half-way between the corners of the mouth and the most anterior points of the lips. In this way, the acoustic effect of the notch-shaped vocal tract termination at the lips is roughly accounted for according to the data by Lindblom et al. (2007). The boundaries between the three sectors intersect in the center point of a circle that represents the (movable) tongue body in the vocal tract model (cf. Figure 1a).

Each grid line intersects the posterior-superior outline and the anterior-inferior outline of the vocal tract as shown in Figure 1a. The first estimate of the center line is the sequence of straight-line segments joining the midpoints of the grid lines delimited by the outlines. At positions where the cross-sectional area of the vocal tract changes abruptly, this center line exhibits sharp bends, which are unlikely to represent the true path of acoustic wave propagation. Therefore, this initial center line is smoothed in the second step with a 2 cm long moving average filter to obtain the final center line that is shown in Figure 1b. At each of 129 equally-spaced points along the final center line, the 3D wire-frame meshes that constitute the vocal tract model are intersected with a plane perpendicular to the center line in the respective point. For each cut, the cross-sectional area is obtained as one sample of the area function. Figure 1c shows the piece-wise linear area function corresponding to the vocal tract shape and center line in Figure 1b.

The idea of the enhanced area function was to associate each position \( x \) along the tube axis not only with the cross-sectional area \( A(x) \), but also with the articulator \( \alpha(x) \) that confines the vocal tract at the anterior-inferior side. In our model, this information is directly available from the identity of the wire-frame mesh that is intersected by each cutting plane at the anterior-inferior end. In the current implementation, the associated articulator \( \alpha \) can assume one of four nominal values, namely “tongue”, “lower incisors”, “lower lip”, and “other”. In Figures 1b and c, the segments for tongue and lower lip are red, the segment for the lower incisors is white, and the remaining segments for the laryngeal region and for the sublingual cavity are gray.

For computational reasons, the vocal tract is mostly represented in terms of a discrete area function that corresponds to a sequence of incremental cylindrical tube sections. In our case, this requires to map the piece-wise linear area function \( A(x) \) and the associated articulators \( \alpha(x) \) to \( N \) cylindrical tube sections, each having a cross-sectional area \( A_i \), a length \( l_i \), and an associated articulator \( \alpha_i \). Hence, if \( x = \sum_{i=0}^{l} l_i \) (with \( x_0 = 0 \) and \( 1 \leq i < N \)) denotes the position of the \( i \)-th section, \( A(x) \) and \( \alpha(x) \) in the intervals \( [x_i, x_i + l_i] \) have to be mapped to single values for \( A_i \) and \( \alpha_i \), respectively. With regard to the area, the obvious approach would be to assign \( A_i \) the average of \( A(x) \) in the respective interval. However, in this way, very short closures in the vocal tract (shorter than a tube section length) may be “released” and become a narrow constriction instead, because the areas greater than zero directly next to the closure contribute to the average value. To prevent this problem, we propose to use the minimum of \( A(x) \) in the interval, i.e., \( A_i = \min_{x=x_i} \{ A(x) \} \). With regard to the associated articulator, \( \alpha(x) \) may have different nominal values in the interval \( [x_i, x_i + l_i] \). Here, \( \alpha_i \) should take the value of \( \alpha(x) \) at the position of the minimal area in the interval, i.e., \( \alpha_i = \alpha(x_0) \) with \( x_0 = \arg \min_{x \in [x_i, x_i + l_i]} \{ A(x) \} \), because this is the relevant articulator in the case that the tube section forms a constriction for noise generation.

As a compromise between low computational cost, which requires as few tube sections as possible, and high spatial detail, which requires as many sections as possible, we represent the area function with longer tube sections in the posterior part, and shorter tube sections in the anterior part, where most fricatives are produced and spatial accuracy is preferable. In fact, we...
use 16 tube sections of equal length between the glottis and the velo-pharyngeal port position, and an additional 24 tube sections with decreasing length between the velo-pharyngeal port and the mouth opening (40 sections in total). As an example for the discretization, Figure 2a shows the discrete enhanced area function corresponding to the piece-wise linear area function in Figure 1c.

![Figure 2: Discrete enhanced area functions of four fricatives. Noise source locations are marked by gray circles, constricted regions are marked by gray horizontal bars, and the positions of the upper incisors are marked by yellow triangles.](image)

### 3. Noise source modeling

Based on the discrete enhanced area functions, we propose the following method for predicting noise sources. First of all, potential critical constrictions formed with the tongue or the lower lip are identified. Therefore, in each of the two regions where \( \alpha_s = \text{"tongue"} \) and \( \alpha_s = \text{"lower lip"} \), the tube section with the smallest area \( A_{\text{min}} \) is identified. If \( A_{\text{min}} < 1 \text{ cm}^2 \), all contiguous tube sections left and right of this section for which \( A_t < A_{\text{min}} + 0.2 \text{ cm}^2 \) are considered as one continuous constriction. In each example in Figure 2, there is one such constriction in the lingual region (marked by the gray bars below the area functions). In the example for /f/ in Figure 2d, there is also such a constriction formed with the posterior section of the lower lip region (the gray bar is hidden under the yellow triangle). In the lingual region, there may generally be a second constriction if it is not connected to the first one and satisfies \( A_{\text{min}} < 1 \text{ cm}^2 \). This is important to prevent acoustic artifacts when a constriction in the lingual region suddenly "jumps" from one position to another in running speech, for example from the tongue tip to the tongue back in /su/ when the tongue tip constriction is released so far that it becomes greater than the tongue back constriction of /u/. As Figure 2 shows, the area function has been extended with two short tube sections to the left that represent the glottis. These glottis sections are considered as a permanent additional constriction for aspiration noise with a variable gain.

Each of the identified constrictions gives rise to one noise source. All noise sources are modeled as localized turbulent sound pressure sources and characterized by their magnitude, spectral shape and location. According to Stevens (1998), p. 103, the source magnitude is calculated as \( p_s = G \cdot |r_s| \cdot \sqrt{A_{\text{min}}} \), where \( G \) is a position-dependent gain and \( r_s \) is the low-frequency part of the air particle velocity in the constriction. The latter was obtained by low-pass filtering the actual air particle velocity \( r \) with a first-order low-pass filter with a cutoff frequency of 500 Hz. The values of \( G \) for different constriction locations were determined in re-synthesis experiments as described further below.

Each noise source generates Gaussian white noise that is shaped with a first-order low-pass filter with a specific cutoff frequency \( f_c \). The filter produces a high-frequency tilt of the source spectrum that is evident from a variety of experimental studies (Narayanan and Alwan 2000). For lingual constrictions, \( f_c \) is set to 0.15 \cdot v_s / d (following Stevens 1998, p. 104), where \( d = \sqrt{4 \cdot A_{\text{min}} / \pi} \) is the diameter of the constriction. For labiodental and glottal constrictions, \( f_c \) is set to 6000 Hz to generate an essentially flat source spectrum (Stevens 1998).

The position of a noise source in our model depends on the source type. When \( x_{\text{sep}} \) denotes the assumed point of flow separation at the anterior end of the respective constriction, and \( D \) is the distance between \( x_{\text{sep}} \) and the tip of the incisors, the noise source is assumed 0.15 cm downstream of \( x_{\text{sep}} \) for a labiodental source, 0.25 cm downstream of \( x_{\text{sep}} \) if \( D > 2 \text{ cm} \) ("wall source" type), directly at the tip of the incisors when \( D \leq 2 \text{ cm} \), and 1.5 cm above the glottis for the glottal constriction. The information about the position of the upper incisors is currently not contained in the enhanced area function but obtained separately from the vocal tract model.

In the transmission-line model of the vocal tract underlying the acoustic simulation, actual noise sources can only be inserted at boundaries between tube sections (Birkholz, Jackel, and Kröger 2007). Therefore, a predicted noise source located at some point within a tube section is always realized as two actual pressure sources in the transmission-line network – one at each end of the section, with the amplitudes linearly scaled according to the distance between the predicted and actual sources. The two actual sources generated for each constriction are shown as gray circles in Figure 2.

All source properties discussed above are summarized in Table 1. The relative gain \( G \) of the different types of sources was determined with analysis-by-synthesis. Therefore, a natural production of the utterance /afasaʃaʃaʃaʃaʃ/ spoken with a flat intonation, was re-synthesized with VocalTractLab 2.1, where the proposed noise source model was implemented. The values of \( G \) for the different noise source types were incrementally adjusted such that the noise levels of the synthetic fricatives approached the levels of the natural fricatives. Figure 3 shows the natural and (final) synthetic productions of /ʃ,ʃ,ʃ/ are very similar to the original levels, it is somewhat too high for the synthetic /ʃ/ (which was realized with an "obstacle source" like /ʃ,ʃ/).
Table 1: Properties of modeled noise sources. $v_c$ is air particle velocity in the constriction and $d$ is constriction diameter.

<table>
<thead>
<tr>
<th>Articulator forming the constriction</th>
<th>Glottis</th>
<th>Tongue</th>
<th>Obstacle source at the incisors</th>
<th>Lower lip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source type</td>
<td>Aspiration source</td>
<td>Wall source</td>
<td>Obstacle source at the incisors</td>
<td>Labiodental source</td>
</tr>
<tr>
<td>Typical phonemes</td>
<td>many</td>
<td>/k, w/</td>
<td>/z, j, ʃ, ʒ, ʃ, ʒ/</td>
<td>/l, v/</td>
</tr>
<tr>
<td>Condition for source</td>
<td>none</td>
<td>Distance from flow separation point to tip of upper incisors greater than 2 cm</td>
<td>Distance from flow separation point to tip of upper incisors smaller than 2 cm</td>
<td>Constriction area must be smaller than that of a potential tongue tip constriction</td>
</tr>
<tr>
<td>Source position</td>
<td>1.5 cm downstream from the flow separation point (glottal exit)</td>
<td>0.25 cm downstream from the flow separation point</td>
<td>Tip of the upper incisors</td>
<td>0.15 cm downstream from the flow separation point</td>
</tr>
<tr>
<td>Cutoff freq. $f_c$ in Hz</td>
<td>6000.0</td>
<td>0.15 · $v_c/d$</td>
<td>0.15 · $v_c/d$</td>
<td>6000.0</td>
</tr>
<tr>
<td>Relative gain $G$</td>
<td>0.005...0.5</td>
<td>5.0</td>
<td>10.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Figure 3: Oscillograms and spectrograms of the utterance /afasafacaga/ - natural production (bottom) and re-synthesis (top).

4. Discussion and conclusion

The main goal of this paper was to introduce the idea of the enhanced area function. It was devised as a coherent extension of the classic area function to support modeling of noise sources, without giving up the simplicity of the area function as an abstract representation of the complex 3D shape of the vocal tract for acoustic simulations. Based on the enhanced area function, a noise source model was presented, which was recently implemented in the articulatory speech synthesizer VocalTractLab 2.1 (www.vocaltractlab.de). According to informal listening tests, the model could generate all German fricatives in high quality in connected synthetic utterances.

The noise source model can still be improved in many ways. For example, not only dipole sources contribute to the spectrum of fricatives, but also monopole sources should be modeled. Furthermore, in voiced fricatives, the friction source is known to be modulated by voicing, where the phase of modulation depends on the distance between constriction and obstacle (Jackson and Shadle 2000). This delay in the modulation is perceptually relevant and should be considered. Also the fixed threshold of 2 cm between the exit of a lingual constriction and the incisors to discriminate between wall and obstacle sources should be replaced by a smooth “blending” of sources. Finally, a formal perceptual evaluation of the synthesized fricatives is needed.

5. References


Voice onset time in language acquisition: Data from Hungarian

Judit Bóna, Anita Auszmann

Eötvös Loránd University, Hungary

bona.judit@btk.elte.hu, auszmannanita@gmail.com

Abstract

Voice onset time (VOT) provides important information about the language acquisition especially how the motor speech skill is being refined. Our hypotheses were the following: 1) If the age increases, the mean and standard deviation of VOTs decrease. 2) The VOTs of Hungarian-speaking children become similar to adults' VOTs at the age of 11. 3) The differences between genders depend on the age and the types of plosives. In this study 30 Hungarian-speaking children were participated in three age groups (9-, 11- and 13-year-olds). The VOTs of voiceless plosives ([p], [t], [k]) were analyzed in spontaneous speech. Our hypotheses were only partially proved by the results. There were significant differences between VOTs of the different plosives in all age groups. There were age differences in VOTs of all three plosives: with increasing age, the means of the VOTs of [p], [k] decreased, and the mean of the VOTs of [t] increased. Gender differences in VOTs depended on the age of the speakers.

Keywords: voice onset time, voiceless plosives, Hungarian speaking children

1. Introduction

Plosives occur in practically all languages of the world but they can be rather varied as far as their articulatory and acoustic properties are concerned (Maddieson 1984; Laver 1994). The silent closure phase of voiceless plosives is followed by a burst phase and then transition to the next sound; whereas in voiced plosives, voicing may be present in the closure phase and during the burst, too (Ladefoged & Maddieson 1996; Stevens 1998).

Voice onset time (VOT) is the length of time that elapses between the burst and the onset of voicing for the next voiced segment (Lisker & Abramson 1964; Zlatin 1974; Lieberman & Blumstein 1988). VOT values may differ from one language to the other, and they are influenced by a number of factors other than the voicing of the plosive including place of articulation (Volaitis & Miller 1992), the quality of the following vowel (Pind 1999), speech rate (Baum & Ryan 1993; Pind 1995), and speakers' sex and age (Whiteside–Marshall 2001; Whiteside–Dobbin–Henry 2003), too.

The analysis of VOT in children’s speech is very important, because it provides us with information on how this aspect of motor speech skill is being refined (Whiteside–Dobbin–Henry 2003). VOT indicates atypical language acquisition, thus, the measurements of typical children’s speech could serve as control for analysing the speech of atypically developing children. In addition, acquisition of VOTs is a challenge to many bilingual speakers, too (Fabiano-Smith–Bunta 2012).

In childhood, VOTs are affected by language acquisition, refinement of speech perception processes, and the fact that the motor control and articulation gestures are becoming more accurate (Whiteside–Marshall 2001; Whiteside–Dobbin–Henry 2003). In English-speaking children’s speech production the variability of VOT gradually decreases, and VOTs become similar to adults’ VOTs by the age of 11 (Whiteside–Marshall 2001).

There is also difference between males and females in VOT; however gender differences have little attention in the literature. Whiteside and Marshal (2001) analysed 7, 9, and 11 year-old children’s VOTs, and they found gender differences at age of 11. With increasing age (between the analysed 7 and 11 years of age), VOTs of [p] and [b] decrease in males, while VOT of [t] increase in females. In another study, Whiteside, Dobbin and Henry (2003) found that differences between males’ and females’ VOTs were higher around 13 years of age, and gender differences occurred to different degree in various phonetic contexts. In analysis of Swedish children’s speech at the ages 3 and 9, Karlsson et al. (2004) found significant gender effects in the aspirated plosives in the 3-year-old children’s group that were not present in the plosives produced by adults. The authors hypothesised that the effect of gender at this early childhood might have been due to the differences in trans-glottal airflow properties (Karlsson–Zetterholm–Sullivan 2004).

We have only sporadic data about Hungarian children’s VOT (for example Gósy 1984). Analysing Hungarian VOTs could lead us to different results from the above mentioned languages (English and Swedish), because they have plosives with aspirated VOTs, while Hungarian plosives are not aspirated (Gósy 2004). Hungarian VOT has been investigated both in word lists and in adults’ spontaneous speech (Gósy 2000; Gósy & Ringen 2009; Grácz et al. 2009; Neuberger–Grácz 2013; Bóna 2011). The results show that VOT in Hungarian depends on the place of articulation, the following vowel, and the type of speech.

Bóna (2012) analysed voice onset time in voiceless plosives produced by members of three age groups in spontaneous speech. The three groups included 9–10-year-old children, 22–31-year-old young adults and 70–90-year-old speakers. The results showed that all three plosives under discussion ([p, t, k]) exhibited significant differences in VOT between age groups. The widest scattering of VOT values was found in the children’s productions for all three consonants. Gender differences occurred only with [k] in the children’s group, with [t] and [k] in the young adults’ group, while in the group of old subjects, VOT values for all three consonants significantly differed between male and female subjects.

The aim of this paper was to investigate the VOT of voiceless plosives ([p], [t], [k]) in Hungarian-speaking children’s speech. We had three preliminary hypotheses: 1) If the age increases, the mean and standard deviation of VOTs decrease. 2) The VOTs of Hungarian-speaking children are stabilized in spontaneous speech at the age of 11, and become similar to adults’ VOTs. 3) The differences between genders in the velar plosives is detectable at the age of 9, but in the alveolar plosives can only be detected at the age of 13.
2. Subjects, material, and methods

2.1. Subjects

30 children’s recordings were selected from speech samples recorded for earlier studies (Bóna 2012; Auszmann 2013; Neuberger 2013) for this analysis. The subjects were divided in three age groups: 9, 11, and 13 years old. All of them were Hungarian-speaking children, and none of them reported any hearing or speech disorders. In all groups there were 5 males and 5 females.

2.2. Material

Spontaneous speech was recorded from all children in a quiet room. Subjects were tested individually. They were asked to talk about their family, school and free time activities. We recorded spontaneous speech instead of word or sentence recall or read speech, because 1) in these ages articulation rate differs in these speech styles from spontaneous speech (it is slower in recall and reading); 2) in recalls the production of the item to repeat can influence the children’s speech production.

In Hungarian VOT is not influenced by the phonetic position of the plosives. In a previous study, we found that phonetic position did not significantly affect the VOT values across the positions #CV, V.CV, and CCV (Bóna 2011). A similar result was obtained by Gósy (2010) from her analysis of the speech of young men and women with respect to [t]. Therefore, in the present study, we did not separate the VOT data by position.

We analyzed about 100–150 data from each speaker to compensate the effects of other factors based on the characteristics of spontaneous speech. Altogether 4146 VOTs were analyzed.

2.3. Methods

The VOTs of voiceless plosives ([p], [t], [k]) were analyzed using Praat 5.0 (Boersma–Weenink 2008). We annotated all of the files by hand.

VOT measurements raise a number of methodological issues (Francis et al. 2002; Gráczi & Kohári 2012). On the one hand, the burst cannot always be seen in the spectrogram, and the consonant may switch into fricative articulation after the closure period, or multiple bursts can be attested (Gráczi & Kohári 2012). On the other hand, researchers may disagree about the end point of voice onset time. Francis et al. (2003) compared four different methods of measurement with respect to VOT: the end point can coincide with the onset of regular voicing, that of F1, or F2, or F3. They performed both acoustic and electro-glotographic measurements and found that the most “accurate” method of measuring VOT (i.e., one that yielded the least amount of variability) was based on the onset of regular voicing.

Thus, in this study voice onset time was defined as the time span between the beginning of the burst and the absolute onset of voicing as observed on the oscillogram and on the spectrogram in parallel (Beckman et al. 2011). First we compared all VOTs of the same plosives across the age groups. Then we calculated the means and standard deviations for all speakers for the comparision. VOTs were also compared between males and females. Statistical analyses (one-way ANOVA for normal distribution, and Mann-Whitney U-test, the non-parametric alternative to the independent samples t-test) were carried out by SPSS 13.0 at the 95% confidence level.

3. Results

There were significant differences between VOTs of the different plosives in all age groups. There was only one exception: there was no significant difference between the VOTs of the 9-year-old children’s bilateral and alveolar plosives (at this age between [p] and [k] $Z = -9.262; p \leq 0.001$; between [t] and [k] $Z = -9.262; p \leq 0.001$). (In 11-year-old children between [p] and [t] $Z = -4.939; p \leq 0.001$; between [p] and [k] $Z = -8.991; p \leq 0.001$; and between [t] and [k] $Z = -11.675; p \leq 0.001$.

In 13-year-old children between [p] and [t] $Z = -4.125; p \leq 0.001$; between [p] and [k] $Z = -9.908; p \leq 0.001$; and between [t] and [k] $Z = -15.135; p \leq 0.001$.)

With increasing age, the means of the VOTs of [p] and [k] decreased, while the mean of the VOTs of [t] increased (Table 1). The mean of VOTs of [p] in native Hungarian speaking young adults was 18 ms (St. dev. 5.3); the mean of VOTs of [t] was 18 ms (St. dev. 5.8); and the mean of VOTs of [k] was 36 ms (St. dev. 10.7) (Bóna 2014).

Standard deviations in the cases of [p] and [k] were lower in 13-year-old children than in 9-year-old children, while in the case of [t] standard deviation was higher in the older groups.

Table 1: Means and standard deviations of VOTs.

<table>
<thead>
<tr>
<th>Age</th>
<th>VOT (ms)</th>
<th>St. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[p]</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>25</td>
<td>9.9</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>11.5</td>
</tr>
<tr>
<td>13</td>
<td>21</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>[t]</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>7.9</td>
</tr>
<tr>
<td>11</td>
<td>25</td>
<td>12.2</td>
</tr>
<tr>
<td>13</td>
<td>25</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>[k]</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>36</td>
<td>12.4</td>
</tr>
<tr>
<td>11</td>
<td>33</td>
<td>14.5</td>
</tr>
<tr>
<td>13</td>
<td>34</td>
<td>11.6</td>
</tr>
</tbody>
</table>

There were age differences in VOTs of all three plosives according to the statistical analyses (Figure 1). In case of [p], there were significant differences between 9- and 11-year-old children ($Z = -4.482; p \leq 0.001$), between 9- and 13-year-old children ($Z = -2.430; p = 0.015$), and between 11- and 13-year-old children ($Z = -2.374; p = 0.018$), too. In case of [t], there were significant differences between 9- and 13-year-old children ($Z = -5.322; p \leq 0.001$) and between 11- and 13-year-old children ($Z = -3.916; p \leq 0.001$), but there was no significant difference between 9- and 11-year-old children. In case of [k], there were significant differences between 9- and 11-year-old children ($Z = -3.80; p = 0.001$), between 9- and 13-year-old children ($Z = -2.96; p = 0.003$), and between 11- and 13-year-old children ($Z = -3.279; p = 0.001$), too.
We calculated the means and standard deviations for all participants, and compared them between the age groups. The statistical analyses showed that there were no significant differences in the individual values between the age groups.

We compared VOTs of males and females in all plosives in each age group. In case of [p], there was no significant difference between boys and girls in any age group (Figure 2).

In case of [t], there was no difference between the 9-year-old boys and girls, while the difference was significant between males and females in the 11-year-old group ($Z = -2.362; p = 0.018$) and in the 13-year-old group ($Z = -2.467; p = 0.014$) (Figure 3).

The results were unexpected in case of [k]: there was significant difference between the 9-year-old boys and girls ($Z = -3.966; p \leq 0.001$), but in the other two age groups the difference was no significant between males and females (Figure 4).

In this paper, we analysed voice onset times in voiceless plosives produced by 9-, 11- and 13-year-old Hungarian children in spontaneous speech. We had three initial hypotheses, which were only partially confirmed.

Our first hypothesis was that with increasing age, the mean and standard deviation of VOTs decrease. This were confirmed only for [p] and [k], but in the case of [t] we observed opposite trend: the mean of the VOTs of [t] increased in the two older group compared to the 9-year-old children’s VOTs.

The second hypothesis was that VOTs of Hungarian-speaking children become similar to adults’ VOTs at the age of 11. This hypothesis was partly confirmed, too. The VOTs of [k] were similar to the values measured in adults’ speech at the age of 9, while in the case of [p] at the age of 11. The VOTs of [t] were longer also at the age of 13 than adults’ VOTs (Bóna 2014).

The confirmation of our third hypothesis also depended on the type of plosives. The differences between genders in the velar plosives was detectable at the age of 9, but unexpectedly there was no difference between males and females in the other two age groups. In the alveolar plosives gender differences could be detected at the age of 11 and 13.

The results show that the place of articulation determines the age-related changes of VOT. In the analyzed age groups...
there are great individual differences between the children (as the statistics also showed that). It seems that the stabilization of the articulation of plosives is not completed at the age of 13. Our results show that the acquisition of voiceless plosives and VOT is language-specific.

5. Acknowledgements

The authors would like to thank Tilda Neuberger for the speech samples provided available for this research.

6. References


Bóna, Judit (2014). “Voice onset time and speakers’ age: Data from Hungarian”. Clinical Linguistics and Phonetics. Accepted paper readable in Early Online.


Plosive and Fricative Geminates in Tarifit
An Articulatory and Acoustic Study

Fayssal Bouarourou1, Béatrice Vaxelaire1, Yves Laprie 2, Rachid Ridouane3, Marion Bechetî & Rudolph Sock1,4

1-U.R. 1339 LILPA/ Equipe de Recherche Parole et Cognition & Institut de Phonétique de Strasbourg (IPS)
Université de Strasbourg (Uds)
22, rue Descartes 67084 Strasbourg, France
2- LORIA/CNRS, Nancy, France
3- Laboratoire de Phonétique et Phonologie (CNRS-Sorbonne nouvelle), Paris, France
4- Pavla Jozefa Safarika University, Faculty of Arts, Košice - Slovakia

Fayssalbouarourou@gmail.com

Abstract

This investigation, based on acoustic data for six native speakers, and on X-ray data for two native speakers, reports on gemination in Tarifit Berber (spoken in Northern Morocco). It presents results of articulatory and acoustic investigations of singleton and geminate voiced and voiceless consonants, produced in word initial, word medial, and word final positions, at a normal and at a fast speaking rate. Speech rate is varied in order to evaluate the robustness of the phonological contrast. Special attention is paid to the timing of tongue gestures in producing this phonological contrast.

1. The problem

Several studies have sought to determine acoustic cues for gemination in many languages: see e.g. [1] to [19]. One consistent acoustic characteristic shared by geminates is that they are significantly longer than their singleton counterparts (see [18] for a review of 24 languages opposing singletons to geminates). Lahiri & Hankamer [9], for example, investigated the timing properties of singleton/geminate voiceless stops in Turkish and Bengali and showed that closure duration is the most important correlate of the geminatesingleton voiceless stop in both languages. In addition, VOT is longer for geminates in Turkish while vowel duration is unaffected. In Bengali, vowel duration is shorter before geminates, but VOT is unaffected. In his acoustic study on gemination in four unrelated languages – Levantine Arabic, Standard Hungarian, Indonesian Madurese, and Swiss-German Bernese – Ham [5] found that the only acoustic correlate that significantly distinguishes geminates from singletons is closure duration. Positive VOT or burst duration does not contribute to the contrast between these consonants in any of these languages. Other findings suggest, however, that rather than being restricted to durational differences alone, the implementation of gemination may have implications for most if not all of a form’s phonetic shape involving vowel and consonant qualities and resonances. In Malayalam, for example, forms containing geminates differ systematically from those without geminates in terms of phonation, tense vs. lax articulations, consonant and vocalic resonances as well as patterns of articulatory variability in adjacent consonants [11]. The acoustic and articulatory characteristics of geminates have been extensively examined in Tashlihyt, a related Berber language (see e.g. [12], [15], [17], and [18]). These studies have reported that geminates are significantly longer than their singleton counterparts. Based on the temporal information about the articulatory closure using EPG, Ridouane [17] showed that these durational differences were robustly maintained for voiceless stops even in utterance-initial position, though these durational differences are not perceptible. In addition to closure duration differences, gemination was also found to significantly shorten the preceding vowel. The majority of the production studies on geminates has been limited to intervocalic geminates, a fact which is unsurprising given that these segments are most widely found in this environment. The acoustic characteristics of initial and final geminates have not been subject to as much investigation.

Tarifit Berber contrasts singletons and geminates in intervocalic as well as in initial and final positions. In this study we present results of an investigation of the acoustic and articulatory correlates that distinguish singleton stops and fricatives from their geminate counterparts. The novelty in the present study, compared with previous work [22] and [23], is that investigations consider all obstruents in three positions: word-initial, word-medial, and word-final positions. The aim is to determine whether position in the word shapes variability in singleton/geminate contrast. In addition, we varied speech rate in order to determine whether geminate contrast is resistant to increased speaking rate. It is hypothesised on the acoustic level that, as reported in the literature, geminates would have longer closure durations than singletons (hypothesis 1). The duration of flanking vowels may be affected by that of geminate consonants (hypothesis 2): they would be shorter in this environment [10], in case of syllable isochrony. VOT could be longer for geminates, as their occlusion phase is usually remarkably long, thus retarding onset of voicing, due to high intra-oral pressure (hypothesis 3). On the articulatory level, contact-extent for plosives and length of maximum constriction for fricatives, partly underlying consonantal closure, respectively for these two categories, would be correlatively longer for geminates (hypothesis 4). If geminates do shorten adjacent vowels, vowel constriction opening may vary as a function of this coarticulatory influence; the size of the constriction would be reduced as vowel duration reduces (hypothesis 5) [21].
2. Method

The entire corpus (plosives and fricatives) consisted of 54 sentences of 4 to 6 syllables, comprising 27 minimal pairs that were inserted in these meaningful carrier sentences. The speech material analysed here consists of all the 27 minimal pairs, contrasting singleton stops and fricatives with their geminate counterparts, in three positions: word initial, word medial, and word final. The plosives examined were: /t/, /d/, /k/, /g/, /v/ vs. /t/, /d/, /k/, /g/, /q/ . The fricatives were: /s/, /z/, /j/, /v/ vs. /s/, /z/, /j/, /v/. All target sequences were inserted in the same carrier sentence: /sinɪvɪtʃɪŋ/ umar /. Meaning “Say _ once”. For the acoustic investigation, the six subjects (two women and four men, from 24 to 30 years old) were seated comfortably at a distance of 20 cm from the microphone, in an anechoic room. All tokens were repeated twelve times by the six speakers, in the two rate conditions. All pairs of sentences had the same number of syllables.

In the X-ray experiment (25 frames/sec), these minimal pairs were produced once at a normal (self-selected) speaking rate, by two speakers (two men).

The X-articulator software, developed at LORIA in Nancy within the DOCVACIM project [20], includes various tools devoted to processing cine-radiographic data. These tools comprise semi-automatic algorithms to monitor speech articulators, a graphic interface which allows editing these contours, and also tools devoted to data analyses and elaboration of articulatory models. These X-ray data processing tools have allowed creating entire contours corresponding to the position and movements of the speech articulators. Hence, for this specific investigation, measurement parameters for vocal tract configurations were determined related to constriction opening within the vocal tract. They provide for plosives: tongue tip to alveolar ridge, tongue body-to-soft palate, and tongue body-to-uvula contact-extents (mm). For fricatives and vowels adjacent to or flanking target consonants, constriction opening (mm) was monitored throughout the entire vocal tract, from the alveolar region to the larynx.

Temporal events were detected on the audio signal, and specific intersegmental and intrasegmental timing relations between these events allowed determining acoustic durations (ms) which correspond respectively to articulatory opening and closing gestures, and also to timing between supraglottal and glottal gestures. Thus, for intersegmental timing relations, vowel durations were specified as intervals between onset and offset of a clear formant structure, for V1 and V2. Corollary, closure duration was measured, between vowels, from offset to onset of clear vocalic formant structures. As concerns intrasegmental timing relations, VTT (Voice Termination Time, measured from vowel offset to the last voicing pulse within the voiceless plosive), plosive occlusion (i.e. closure duration excluding VOT for voice plosives), the acoustic silent phase (for voiceless plosives) and VOT (the interval between the burst-release of the plosive and onset of a clear formant structure of the subsequent vowel) were also acquired.

General remark on acoustic measures: It was expected following results usually reported in the literature on quantity contrasts that, in spite of any eventual compression that measured parameters might undergo, due to increased speaking rate, differences in consonantal closure (the privileged parameter of the phonological contrast) would nonetheless be maintained. Taking into account the elasticity of speech signals [4], which vary as a function of speakers, speaking rates, diverse contexts, ... differences in absolute values between geminates and singletons were normalised. Thus, the proportion of consonantal closure within the CV2 syllable was calculated. It has indeed been shown [10] that it is within this CV domain that temporal contrasts for consonantal quantity are maximised. In fine, therefore, fine grained analyses of the data, together with our conclusions, will be drawn from these relative values.

3. Results

All statistical analyses of the acoustic data were made using Prism® software. Two-way analyses of variance (ANOVA) were thus carried out for all variables (V1, VTT, consonantal closure, occlusion / acoustic silent phase, VOT, V2) in order to determine the statistical significance of main effects (gemination, voicing, place of articulation and speech rate) followed by a Bonferroni post-hoc pair-wise test, so as to compare the reproducibility of mean values within speakers. Only results significant with a probability of less than five per cent (p<0.05) were retained. Since the measured parameters differed between plosives and fricatives, separate statistical analyses were carried out for the two categories. Two main effects proved to be statistically significant for both the intersegmental consonantal closure and intrasegmental occlusion/silent phase variables: gemination (p<0.0001) and speech rate (0.0001). Hence, post-hoc pair wise comparisons (Bonferroni) were carried out on mean values of absolute and relative values only for these variables.

The acoustic data reveal that consonantal closure and the occlusion/silent phase of geminates, in absolute values, are significantly longer than corresponding singletons, for all consonants (alveolars, velars and uvulars), and in both the voiced and voiceless contexts, for all six subjects, in the two speech rate conditions. This result is in line with hypothesis 1. This hypothesis is further consolidated as durational differences between geminates and singletons are maintained in fast speech, although consonantal closures undergo compression; this compression is more pronounced for geminates. It was noticed that consonantal gemination did not affect the duration of adjacent vowels V1 and V2, as expected, given statistical results reported supra. Hypothesis 2 is consequently not verified. Likewise for intrasegmental VTT and VOT values which are also similar for both categories (hypothesis 3). The acoustic data further show that consonantal closure and the occlusion/silent phase of geminates, in relative values take up a higher proportion of the CV domain (p<0.0001), compared with their singleton counterparts, thus highlighting the robustness of the phonological distinction, regardless of compression induced by increased speaking rate. Proportions remain relatively stable in fast speech, as they are comparable in this speaking condition for geminates and for singletons.

Articulatory results given here are based on raw data, and rarely on statistics, due to experimental conditions (exposure to X-rays). Some of them should therefore be considered as tendencies. Measurements obtained from mid sagittal profiles show, for plosives, that contact-extents (maximum value for contact) are longer for geminate consonants than for their singleton counterparts. Figure 1a shows contact-extent (frames 187 and 188) for the apical singleton, accompanied by a concomitant enlargement of the pharynx. It can be seen that, conversely, while constriction opening is large in the alveolar region for the flanking /a/ vowels (frames 183 to 186, then frames 190 to 191), pharyngeal constriction is maintained for these vowels. The scenario is structurally...
similar for the geminate plosive (Figure 1b), but contact-extent is remarkably longer (frames 633 to 636). Hence tongue-palate contact perturbs the size of pharyngeal constriction only during occlusion, and not when vocal tract configurations, associated to the production of the vowels, emerge. Constriction opening in the critical pharyngeal region for vowel /a/ does not vary as a function of any coarticulatory influence from the consonant and this is seemingly why no shortening of the adjacent vowels by geminates was observed on the acoustic level (hypothesis 5 not confirmed). These observations are valid, in an intra-speaker pair-wise comparison, for all the linguistic categories examined i.e. alveolars, velars and uvulars, for voiced and voiceless consonants, for fricatives (based on noticeable differences of constriction length, see Figures 2a and 2b) and for the two speakers (thus corroborating hypothesis 4). The constriction duration differences are maintained for all positions, including word-initial and –final positions (see Figure 3a and 3b for final t/tt contrast).

4. Conclusions

A close look at both the articulatory and the acoustic data suggests that all speakers adopt comparable strategies in contrasting singletons and geminates. Specifically, it is shown that the most systematic acoustic and articulatory correlate distinguishing Tarifit Berber singletons from geminates is consonant duration. This difference holds for all types of obstruents in all word positions, including word-initial and word-final voiceless stops, and operates at a normal as well as at a fast speaking rate. The phonetic characteristics of Tarifit geminates may be captured from a phonological standpoint by a structural representation of these segments as two timing units associated with one segmental slot, the relevant timing measure being the duration of the consonant (and closure duration for stops).

5. References


**Figure 1 a and b:** Frame-by-frame analyses of tongue gestures during the production of the apical singleton /t/ (Figure 1a) vs. geminate /tt/ (Figure 1b), in intervocalic position. Speaker F. See text for explanation.

**Figure 2 a and b:** Frame-by-frame analyses of tongue gestures during the production of the apical singleton /s/ (Figure 2a) vs. geminate /ss/ (Figure 2b), in intervocalic position. Speaker F.

**Figure 3 a and b:** Frame-by-frame analyses of tongue gestures during the production of the apical singleton /t/ (Figure 3a) vs. geminate /tt/ (Figure 3b), in word final position. Speaker F.
The Influence of Lexical-Perceptual Integration on Auditory-Motor Adaptation in Speech Production

Nicolas J. Bourguignon¹,²,⁴, Shari R. Baum³,⁴ & Douglas M. Shiller¹,²,⁴

¹Centre de Recherche du CHU Sainte-Justine, Montreal, QC, Canada
²Ecole d’orthophonie et d’audiologie, Université de Montréal, Montreal, QC, Canada
³School of Communication Sciences and Disorders, McGill University, Montreal, QC, Canada
⁴Centre for Research on Brain, Language and Music, McGill University, Montreal, QC, Canada

Abstract

Lexical bias and altered auditory feedback (AAF) manipulations were combined within a single paradigm to investigate the shared influence of lexical and sensory information on speech motor control: Subjects produced real words or pseudo-words containing the vowel [ε] (as in “head”) under conditions in which F1 frequency was decreased, making it more similar to [ı] (as in “hid”). This changed the lexical status of spoken items from real words to pseudo-words (i.e., less – liss), or from pseudo-words to real words (i.e., kess – kiss). These subjects’ adaptive articulatory patterns were compared with those of a different group of subjects for whom the same AAF did not change the lexical status of real words to pseudo-words (mess – miss) or pseudo-words to real words (ness – niss). Subjects exhibited different patterns of adaptation under these two lexical-change conditions indicating that AAF-based speech motor learning is sensitive to the lexical status of the words produced.

Keywords: speech production, sensorimotor integration, lexical effect

1. Introduction

Because speech processing typically takes place in ambiguous and noisy circumstances, the human nervous system must be able to extract and integrate as many types of sensory, motor and lexical information as possible from the incoming speech stream so as to distinguish linguistically relevant cues from irrelevant sounds (Zion Golumbic et al., 2012; Hickok et al., 2011). Evidence from speech production and perception studies indicates that this challenge holds for both input and output ends of the language system. Beginning with speech perception, it is known that linguistic context determines how speech sounds or words are interpreted. An example of this is Ganong’s (1980) lexical effect on phoneme identification, in which the perception of a phonetically ambiguous sound (e.g., between a [d] and a [t], according to voice onset time differences) will be perceived as a [d] in a given linguistic context (e.g., -ash) so as to favor an interpretation consistent with an existing lexical entry (i.e., dash) over a non-lexical one (i.e., tash; see Figure 1). This suggests that the identification of a target phoneme is biased towards real-words, leading to a shift of perceptual boundaries that matches participants’ lexical knowledge.

The production of speech sounds also rests upon the integration of articulatory gestures with sensory and linguistic information, though the respective influence of sensory cues and linguistic knowledge on the motor processes of speech production has so far been considered within largely non-overlapping approaches (Hickok et al., 2011; Hickok, 2012). On the side of sensorimotor integration, many studies have stressed the importance of acoustic-phonetic targets in guiding articulatory control, as indicated by studies introducing perturbations in subjects’ auditory feedback while producing words. For instance, producing the target word “head” under conditions in which perceived vowel F1 is decreased yields a compensatory F1 increase in participants’ speech output. This phenomenon illustrates participants’ reliance on auditory feedback in setting and adapting future productions (Houde & Jordan, 1998). Separately, within the psycholinguistic study of spoken language, results suggest an influence of abstract lexical knowledge on speakers’ articulatory productions, notably a propensity to substitute phonemes more often in word strings that yield real words (e.g., barn door – darn bore) as opposed to pseudo-words (e.g., dart board – *bart – doart, cf. Baars et al., 1975). This effect echoes, on the production side, Ganong’s evidence for a bias of the language processing system towards real words.
While these separate demonstrations illustrate that lexical status influences the perception and production of speech sounds, and that both sensory and lexical information drives the neural control of speech production, it is still unknown to what extent sensory and lexical information contributes to speech motor control in an interactive fashion. We set out to examine this question by combining Ganong’s (1980) lexical manipulation with the paradigm of sensorimotor adaptation to altered auditory feedback (AAF) while subjects produced monosyllabic words containing the vowel [e], but varying in lexical status. Specifically, two groups of subjects underwent a lexical change (LC) procedure, producing [e] words that were selected in such a way that the feedback alteration resulted in the perception of real words as pseudo-words (i.e., less – liss; Group 1), or in the perception of pseudo-words as real words (e.g., kess – kiss; Group 2; see Table 1a). The LC groups were compared with two other groups of speakers in a no-lexical-change (NLC) condition, whereby the same F1 decrease was applied during the production of [e] words that did not involve any shift in lexical status (see Table 1). In line with previous demonstrations of a lexical bias in production and perception (Baars, 1975; Ganong, 1980), we predicted that in participants’ phonetic perception of their self-generated auditory feedback, the phonetic boundary would be biased in accordance with lexical status – with the boundary shifting in the direction of the non-lexical item, thus enlarging the area along the continuum containing the real word (see Figure 2).

Table 1 – Stimulus words grouped as a function of condition and groups

<table>
<thead>
<tr>
<th>Lexical change (LC)</th>
<th>No Lexical change (NLC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC1 – G1</td>
<td>LC2 – G2</td>
</tr>
<tr>
<td>e (I)</td>
<td>e (I)</td>
</tr>
<tr>
<td>death (dith)</td>
<td>weth (with)</td>
</tr>
<tr>
<td>depth (dith)</td>
<td>bet (bit)</td>
</tr>
<tr>
<td>nest (nist)</td>
<td>fet (fit)</td>
</tr>
<tr>
<td>less (liss)</td>
<td>ket (kit)</td>
</tr>
<tr>
<td>chess (chiss)</td>
<td>kess (kiss)</td>
</tr>
<tr>
<td>chest (chist)</td>
<td>steff (stiff)</td>
</tr>
<tr>
<td>test (tist)</td>
<td>ked (kid)</td>
</tr>
<tr>
<td>vest (vist)</td>
<td>detch (ditch)</td>
</tr>
<tr>
<td>best (bist)</td>
<td>steck (stick)</td>
</tr>
<tr>
<td>keg (kig)</td>
<td>stall (still)</td>
</tr>
</tbody>
</table>

2. Materials and Methods

2.1. Subjects

Forty native speakers of English (18-30 years) without history of speech, language or hearing disorders took part in the study. All subjects passed a pure-tone hearing screening (threshold < 20 dB HL at octave frequencies between 250 and 4000 Hz) and provided written informed consent prior to testing. Procedures were approved by the Institutional Review Board, Faculty of Medicine, McGill University.

2.2. Stimuli and group assignment

Participants were randomly assigned to one of four groups (10 subjects per group, 5 males and 5 females) and underwent an identical series of mono-syllabic word production tasks under normal-feedback conditions (baseline) followed by the production of words under conditions of AAF (speech adaptation). All words produced by these groups during the speech adaptation task contained the vowel [e] (e.g., “bed” or “geck”). A real-time acoustic manipulation carried out while subjects produced these words yielded a perception of the vowel [e] as being closer to [I] (see section 2.3.2 for details). Depending on the conditions and groups to which participants were assigned, this phonetic manipulation introduced a shift from real words to pseudo-words (Group LC-1), a shift from pseudo-words to real words (Group LC-2), a shift from real words to real words (Group NLC-1) or a shift from pseudo-words to pseudo-words (Group NLC-2).

2.3. Procedure

Speech was recorded in a quiet testing room using a headset-430 mounted microphone (CS20, AKG, Germany) and digitized at 431 16-bit/44.1 kHz on a PC using custom software written in Matlab 432 (Mathworks, Boston, MA, USA). Auditory speech signals were presented to subjects using circumaural headphones (880 pro, 434 Beyerdynamic, Germany).

All subjects underwent the following sequence of tasks:

2.3.1. Baseline speech production

Subjects in each group produced a set of [e] and [I] words (see Table 1) under normal auditory feedback. Each word was presented visually on a computer screen for 1.5 s (ISI: 2 s). Subjects were instructed to produce each item as soon as it appeared on screen.

Figure 2 – Schematic illustration of the effects predicted in the present study (from Bourguignon et al., in press). Vowels are altered by a change in F1 frequency, resulting in [e] productions being perceived more like [I] (black arrow at top). The resulting compensatory change in speech output (red and blue arrows) has the effect of restoring the subject’s perception of their vowel to the original vowel category [e]. Top panel: F1 vowel alterations that have the effect of changing the lexical status of a real-word (e.g., less) to a pseudo-word (e.g., liss) are predicted to yield reduced compensatory changes in participants’ productions (red arrow) as a result of the perceptual lexical effect, whereby the boundary for the vowel contrast is shifted in the direction of [I] (red dashed line). Bottom panel: When the same F1 vowel alteration has the effect of changing the lexical status of a pseudo-word to a real-word, a larger compensatory change is expected (blue arrow) due to the perceptual boundary shifting in the direction of the vowel [e] (blue dashed line).
2.3.2. Speech motor adaptation

After the baseline production task, subjects in each group produced 160 randomized stimulus words (or pseudo-words), all of which contained the vowel [ε] from the group’s target stimulus list. Subjects underwent a series of auditory feedback conditions involving an initial period of normal feedback (30 trials, null phase), followed by a period of practice under condition of altered auditory feedback (100 trials, hold phase), then a final period under normal feedback to test for learning after-effects (30 trials, after-effect phase). The auditory feedback manipulation corresponded to a 30% decrease in the frequency of the first spectral peak (F1, average shift: 216.3 Hz), inducing the perception of the vowel closer to [I]. The system used to carry out the feedback manipulation combines a digital signal processor (DSP) designed for manipulating vocal signals in near-real-time (VoiceOne, TC Helicon) with low-/high-pass filters to restrict the formant shift to F1. A detailed description of the system has been published previously (Rochet-Capellan & Ostry, 2011; Mollaei et al., 2013).

2.4. Acoustic analyses

For each word produced in the baseline and speech adaptation tasks, a 30 ms segment centered around the midpoint of the vowel was selected. Mean F1 and F2 frequency for each segment was then estimated by LPC analysis in Matlab. LPC parameters were chosen on a per-subject basis to minimize the occurrence of spurious formant values. Values of F1 and F2 frequency were used to directly compare vowel acoustic properties during baseline productions of the different stimulus words among the different groups. Analysis of vowel acoustics during the speech adaptation task was restricted to F1 frequency, as the compensatory response was primarily observed in this acoustic parameter. Vowel acoustic changes during the speech adaptation task were computed as the proportion change in F1 frequency relative to the mean values during the null phase (averaged over trials 1-30). Differences in speech adaptation between the different groups were evaluated at three time points: (1) the beginning of the hold phase (trials 31-60) under conditions of altered auditory feedback, (2) the end of the hold phase (trials 101-130) under conditions of altered auditory feedback and (3) during the after-effect phase following removal of the feedback manipulation (trials 131-160).

3. Results

3.1. Baseline

Mean formant values for the baseline productions of [ε] words in the four different groups were compared using one-way ANOVAs. No reliable differences were obtained between the four groups for either F1 (F3,36 = 1.12, p = 0.35) or F2 (F3,36 = 0.30, p = 0.822).

3.2. Speech Adaptation

The results of the speech adaptation task for the two lexical change groups (L1C-1 and L1C-2) are shown in Figure 3A, with mean changes in formant values relative to baseline at the three time-points in the testing sequence shown in Figure 3B. A compensatory increase in F1 frequency occurred in both groups in response to the auditory feedback manipulation. However, the magnitude of this increase can be seen to differ markedly between the two groups (with L1C-1 showing a greater degree of compensation). Figure 4A illustrates the results of the speech adaptation task for the two non-lexical-change groups (NLC-1 and NLC-2). As in the LC groups, altered auditory feedback provoked a compensatory increase in F1 frequency. In contrast with the LC groups, however, little if any difference was observed in the magnitude of compensation between NLC-1 and NLC-2.

An omnibus 3-way ANOVA was carried out with WORD (word vs. non-word) as one between-subjects factor, LEXICAL (lexical change vs. non-lexical-change) as a second between-subjects factor, and PHASE (Early-hold, Late-hold, After-effect) as a within-subjects factor. The two between-subjects main effects (WORD and LEXICAL) were found to be non-significant (WORD: F[1,34]=0.87, p=0.36; PHASE: F[1,34]=0.22, p=0.64), however a reliable 2-way interaction between WORD and LEXICAL was observed (F[1,34]=4.36, p = 0.04). A highly significant main effect of the within-subject variable PHASE was also observed (F[2,68]=22.49, p < 0.001), but there was no reliable 2-way interaction between PHASE and either of the two group variables (PHASE x WORD: F[2,68]=2.52, p = 0.09 PHASE x LEXICAL: F[2,68]=1.20, p = 0.31). The 3-way interaction was also found to be non-significant (F[2,68]=2.04, p = 0.14).

The 2-way interaction between WORD and LEXICAL conditions is of particular interest, since our main prediction involves a difference in the degree of adaptation between the word and non-word production under the lexical-change condition, with no such difference predicted between word groups under the non-lexical-change condition. Post-hoc pair-wise comparisons were carried out using Holm-Bonferroni-corrected t-tests to examine the reliability of these simple group effects. The tests were carried out on adaptation performance at the end of the training period (i.e., the Late-hold phase), as this represented the moment at which speech adaptation was maximal for all groups. A reliable difference in the magnitude of speech adaptation was observed between the
groups producing real-words and pseudo-words in the lexical-change condition ($t[16]=2.91$, $p = 0.04$), while no significant difference was observed between the two word conditions in the non-lexical-change condition ($t[18]=0.84$, $p = 0.41$). A reliable difference was also observed between the two groups producing real words, one in the lexical-change condition and one in the non-lexical-change condition ($t[17]=2.86$, $p = 0.03$). No such difference was observed between the two groups producing pseudo-words ($t[17]=1.55$, $p = 0.28$).

Figure 4 – Results of the Speech adaptation task for NLC-1 (producing pseudo-words) and NLC-2 (producing real-words, from Bourguignon et al., in press). (A) Mean F1 frequency (normalized units) averaged over successive blocks of 10 trials each. (B) Mean normalized F1 at three key time-points during the adaptation task.

4. Discussion and conclusion

The present study investigated the possible interaction between lexical and perceptual information in guiding articulatory movements during spoken speech. Our hypothesis was that compensatory changes in speakers’ articulatory patterns following an auditory-feedback perturbation affecting vowel quality would reflect sensitivity to the lexical status of the words produced. In particular, we predicted that the degree of compensation would vary when the auditory feedback manipulation had an effect on the lexical status of the word being produced (changing a real word into a pseudo-word, or vice versa). Consistent with this prediction, participants’ articulatory compensation to a decrease in F1 inducing a shift from real words to pseudo-words (LC-1) was significantly less than when the same F1 perturbation produced a shift from pseudo-words to real words (LC-2). The speech compensatory response for the LC-1 group was also reliably less than that observed for the group producing real words under non-lexical-change conditions (NLC-1). This pattern of difference was absent between groups producing real words and pseudo-words that did not change in lexical status under AAF.

The difference in adaptation magnitude between LC-1 and LC-2 extends Ganong’s (1980) original finding of a lexical effect on the perception of acoustic-phonetic aspects of speech to language production, showing that speakers tend to keep their acoustic speech outcomes within the sensory space corresponding to the task-related side of the word/pseudo-word boundary (see Figure 3). The conceptual and methodological payoffs of combining altered auditory feedback and lexical bias into a single experimental paradigm are twofold. First, evidence for lexically driven motor adaptation to auditory perturbations demonstrates for the first time a concurrent influence of phonetic and lexical information on the control of spoken speech, indicating that articulatory plasticity is in part constrained by the structure of abstract lexical knowledge. Second, observing a lexical effect through participants’ speech productions bypasses the methodological shortcomings associated with explicit perceptual decision-making tasks, and thus strongly supports the view that the lexical influence on perception involves a change in phonetic processing (e.g., Guether et al., 2006), and is not simply the result of a bias in lexically-driven decision-making.

The present findings indicate that the perceptual and motor sub-systems of the speech apparatus interact to a certain extent with higher-order lexical information. Future work examining this question may help to enrich current neurocognitive perspectives of speech production (e.g., Guether et al., 2006) in a way that would bring together psycholinguistic and motor control theories of language (Hickok, 2012).

5. Acknowledgements

This work was supported by the NSERC-Canada and the National Institute of Health (NIDCD-R01DC012502).

6. References


**The Intonation of Imperatives in Mexican Spanish**

Annika Brehm\(^1\), Alina Lausecker\(^1\), Ingo Feldhausen\(^1,2\)

\(^1\) Goethe-Universität Frankfurt am Main, Germany
\(^2\) UMR 7018-LPP, Université Paris-Sorbonne Nouvelle, France

a.brehm@stud.uni-frankfurt.de, alinalausecker@hotmail.de, ingo.feldhausen@gmx.de

---

**Abstract**

This paper discusses new insights into the intonation of Mexican Spanish imperatives and addresses the question of whether differences exist between lab speech and (semi-) spontaneous speech. Results from a production experiment based on semi-spontaneous speech indicate that the imperative sentences exhibit two different nuclear configurations, depending on whether the verb (\(V_I\)) stands in the sentence-initial or sentence-final position of the utterance: (a) \(L^+H^* L\%\) with \(V_I\) in sentence-final position, and (b) \(L^* L\%\) with \(V_I\) in non-final position. Furthermore, the pitch accent on \(V_I\) in sentence-initial position varies between a delayed peak (\(L^+\rightarrow H^*\)), an early peak (\(L^+\rightarrow H^*\)) and a peak with no distinct rise (\(H^*\)). While these results contradict claims made by De-la-Mota et al. (2010) concerning the nuclear configuration of imperatives, they (partly) confirm the findings in Willis (2002) for pitch accent variation on \(V_I\). The results further suggest that (semi-)spontaneous speech does not necessarily display more tonal variation than lab speech.

**Keywords:** Intonation, Imperatives, Mexican Spanish, Pitch Accent, Nuclear Configuration, Lab and Spontaneous Speech

---

**1. Introduction**

Imperatives are characterized by the illocutionary force of an order, i.e. the speaker wants the interlocutor to do something. While imperatives in Northern Peninsular Spanish have been subject to different intonational studies for decades (e.g. Navarro Tomás 1944, Estebas-Vilaplana & Prieto 2010, Robles-Puente 2011), Mexican Spanish has been examined to a lesser extent (e.g. Willis 2002, De-la-Mota et al. 2010). In addition, the studies on the Central American variety have pursued different goals. Willis (2002) concentrates on the pitch accent realization of the imperative verb (\(V_I\)) in sentence-initial position as Lee in ¿Lee la novela! ‘Read the novel!’; and neglects the nuclear configuration (i.e. the nuclear accent and the following boundary tone; here the sentence final DP la novela ‘the novel’). De-La-Mota et al. (2010), on the other hand, concentrate on the nuclear configuration and fail to address the pitch contour on \(V_I\). Consequently, these scholars present different findings. As for the pitch accent on \(V_I\), Willis (2002) reports a three-way variation (a predominant late \(H\) peak, followed by an early \(H\) peak and a peak with no distinct rise; p. 355f), whereas De-La-Mota et al. (2010) show only a single instance in the corresponding figure ([1] presents the tonal configuration of De-la-Mota et al.’s figure (p.340); the metrical strong syllables are underscored and the syllable with nuclear stress is written in capitals). The pitch accent on \(V_I\) (here \(ven\ ‘come’) is transcribed by \(L^+H^*\) (i.e. early \(H\) peak in Willis’ terms). As for the nuclear configuration, De-la-Mota et al. (2010) state that imperatives have a rising-falling nuclear configuration \(L^+(\text{¡})H^* L\%\), as shown in (1). However, due to two questionable points, we wonder whether such an example in fact demonstrates the actual imperative contour.

---

**First,** since \(ven\) is a mono-syllabic word, the rising pitch accent might be due to the following high edge tone \(H\). Second, the description of the nuclear configuration seems to be based on imperatives including sentence-final facultative elements (such as \textit{ahorita mismo} in ([1])), which are separated from the preceding elements by a prosodic break. As such, we wonder whether the nuclear configuration describes the intonation of some added, facultative material. Furthermore, neither Willis (2002) nor De-la-Mota et al. (2010) address this open issue in a study based on lab speech (here: scripted speech elicited in an experimental, laboratory setting). They show that \(V_I\) in sentence-final position (in both one-word utterances and conditional clauses with material preceding \(V_I\)) is predominantly realized by \(H^*\) and by \(L^+H^*\). The nuclear configuration of sentences with \(V_I\) in initial position, in turn, is typically realized by a low contour (\(L^* L\%\)), and only sometimes by \(L^+H^* L\%\). Thus, the results of Lausecker et al. (2014) largely support the assumptions made for Spanish and Catalan that short declaratives (consisting of one prosodic word) have a rising nuclear accent, whereas long declaratives (with a length of two or more prosodic words) have a low or falling nuclear accent (e.g. Prieto 2002).

It is unclear, however, whether this pattern also holds in (semi-) spontaneous speech.

This paper combines the different approaches of Willis (2002) and De-la-Mota (2010) and investigates both (i) the pitch accent on \(V_I\) in sentence-initial and sentence-final position and (ii) the nuclear configuration of these structures in semi-spontaneous speech. By using verb-argument structures, we attempt to keep the prompts from evoking additional, extrasentential material, and thus try to circumvent the controversial aspects of (1). In order to facilitate a direct comparison with the results from De-la-Mota et al. (2010), we apply their methodology here. The present study thus differs from Lausecker et al. (2014) by considering semi-spontaneous speech and not scripted speech – even though the hypotheses are similar. Two hypotheses (H1 and H2) guide our study:

- **Hypothesis 1:** short imperatives, i.e. declaratives consisting of only one imperative verb (such as \(Bé bebe!\ ‘Drink that!’), are realized with a rising nuclear accent (in accordance with Prieto 2002) and as such do not show any variation in pitch accent (as opposed to Willis 2002).

- **Hypothesis 2:** long imperatives (such as \(Bé bebe la limonada!\ ‘Drink the lemonade!’) should be realized with a low or falling nuclear configuration and a rising pitch accent on \(V_I\) (in accordance with Prieto 2002, but contradicting De-la-Mota et al. 2010); consequently, we do not expect any pitch accent variation either (as opposed to Willis 2002).
2. Methodology

Our production experiment was conducted based on semi-
spontaneous speech in which three native speakers of Mexican
Spanish (one female from Torreón aged 26 [FT], and two
males, one from Mexicali [MM] and one from Monterey
[MY], aged 23 and 32) uttered a total of 120 imperative sen-
tences (3 subjects x 20 situations x 2 repetitions). The data
were gathered by means of a Discourse Completion Task
(Blum-Kulka et al. 1989), in which the subjects had to reply
spontaneously to a given situation. The questionnaire for
the present study consisted of two conditions (C1, C2). For each
condition, ten different situations were created along the lines
presented in (2) and (3). The situations were created analogous
to those used in De-la-Mota et al. (2010) and Prieto &

- **Condition 1**: situations that evoke short imperatives
  consisting of one prosodic word, as shown in (2);
- **Condition 2**: situations that evoke long imperatives,
  with \( V_1 \) in sentence-initial position, as shown in (3).

\[ \text{Tú eres el alumno y quieres escuchar al profesor, pero un compañero está hablando demasiado alto. Pídele que se calle.} \]

Expected response: ¡Callate!

\[ \text{Un chico está comiendo algo durante la clase, pero no debe. Pídele que deje de comer.} \]

Expected response: ¡Deja de comer!

Bold letters indicate the imperative verb \( V_1 \) in (2) and (3). In
situation (2a), the intransitive verb callarse ‘to keep quiet’ is
introduced. Due to the typical subject elimination in impera-
tive contexts, we expect the short imperative response ¡Calla-
tarse!. (2b). In this condition, the pitch accent of \( V_1 \) is at
the same time the nuclear accent. In (3a), the transitive verb dejar
‘to stop doing something’ is introduced. Such verbs require an
internal argument (here: comer ‘to eat’). Consequently, we
expect a long imperative response in which \( V_1 \) is followed by
the argument (e.g. ¡Deja de comer! as in (3b)). Accordingly,
the pitch accent of \( V_1 \) in C2 is a prenuclear accent and does not
form part of the nuclear configuration.

The subjects were recorded in three sessions in a quiet room at
the Goethe-Universität Frankfurt (Germany) with an Edirol
recorder by Roland (44.1 kHz, 16 bit). In order to allow for the
most natural speech possible, the speakers only listened to the
stimuli, which were read out to them, and did not read the
situations themselves. The subjects were then asked to reply
spontaneously to the presented situations. The entire set of
situations was repeated once (2 x 20 situations). Each speaker
went through a small practice session before the recording.
We used praat to analyze the F0 tracks (Boersma & Weenink
2013) and annotated the data with the help of the Sp_ToBI
annotation system (Aguilar et al. 2009, Prieto and Roseano
2009-2013). In the end, 11 sentences were deemed inadequate
and thus had to be excluded: the subjects MM and MY uttered
a question instead of an imperative. Consequently, 109 sen-
tences were used for analysis.

3. Results

Theoretically, we expected 60 sentences per condition (3
speakers x 10 situations x 2 repetitions). However, due to the
semi-spontaneous nature of the study, the speakers did not
always realize the expected responses in each situation. As for
C1, a total of 13 sentences were uttered with a short impera-
tive, while 44 replies were uttered by means of a long impera-
tive. For C2, the subjects realized 52 long imperative respons-
es. Thus, a total of 13 short imperatives and 96 long impera-
tives were uttered (i.e. 109 in total), which each represent
100% in the following calculations.

3.1. Results between the speakers

Our results show that the pitch accent on \( V_1 \) is either high or
rising. The realizations differ between the conditions in that the
pitch accent on \( V_1 \) is prevalently realized by an early peak
(L+H*) in C1 (nuclear region) and by a delayed peak (L->H*) in
C2 (prenuclear region). As for the nuclear configurations, the
most prevalent configurations are L+H* L% in C1 (short
imperatives) and L*L% in C2 (long imperatives).

### Table 1: Nuclear configurations for each condition for all speakers (in %).

<table>
<thead>
<tr>
<th>Condition</th>
<th>L* L%</th>
<th>L+H* L%</th>
<th>L+&gt;H* L%</th>
<th>L+H* H%</th>
<th>H* H%</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>74%</td>
<td>11%</td>
<td>8%</td>
<td>3%</td>
<td>10%</td>
</tr>
<tr>
<td>C2</td>
<td>91%</td>
<td>5%</td>
<td>4%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 1 shows the results for the attested nuclear configuration
for each condition between the speakers. In Condition 1 (C1,
short imperatives), the tonic syllable is predominantly realized
by a rising nuclear peak followed by a low boundary
tone, L+H* L% (85%, N=11), see Figure 1 for a correspond-
ing pitch contour. In two cases, the nuclear accent was realized
by a high or rising accent (H* L% and L->H* L%; each 7.5%,
N=1). The IP edge tone is always low (L%; 100%, N=13).

Since the total number of short imperatives is very small, we
refrained from applying a significance test. As for Condition
2 (C2, long imperatives with sentence-initial \( V_1 \)), the most
commonly uttered nuclear configuration was a low nuclear
accent followed by a low edge tone (L* L%; 74%, N=71), see
Figure 2. The second most popular choice was a falling nuclear
accent with a low boundary (H*L* L%; 12%, N=11). A chi-
square test showed that the difference in frequency be-
tween L* L% and H*L* L% on the one hand and the other
attested nuclear configurations (to be described below) on the
other is highly significant \( \chi^2 = 48.17, df = 1, p = 0.00 \).
Preacher 2001 was used for the calculations. In addition to the
low or falling configurations, the speakers also uttered a rising
accent (L+H* L%; 5%, N=5), and a high accent (H* L%; 2%,
N=2), both followed by a low edge tone, the predominant
choice in C2. With a total of 89 instances, L% was realized
significantly more often than the high edge tone H%, which
occurs in only seven instances \( \chi^2 = 70.04, df = 1, p = 0.00 \).
In these seven utterances, the nuclear accent was found to be
either rising (with or without a delayed peak; L+->H* H% and
L+H* H%; each 3%, N=3) or high (H* H%; 2%, N=1). As is
evident in Table 1, the total number of attested nuclear configurations was broader in C2 (N=7) than in C1 (N=3). In terms of the pitch accent realization on VI, the speakers consistently realized either a rising or high nuclear accent in C1 (L+H*, L+>H*, H*). However, with 11 out of 13 instances, L+H* was realized considerably more often than L+>H* and H* (an example of L+H* is given in Figure 1). As for C2, in which VI is located in the prenuclear area, the speakers also realized only rising or high pitch accents on VI. Here, however, the most prominent choice was the delayed peak L+>H* (52%, N=50, Figure 2), followed by the rising accent L+H* (31%, N=30) and the high accent H* (17%, N=16).

![Waveform, spectrogram, and F0 trace for the short imperative (C1)](image1)

Figure 1: Waveform, spectrogram, and F0 trace for the short imperative (C1) ¡Siéntanse! 'Sit down!' of speaker MM (sentence 2_1), produced with an early peak (L+H*) and a low edge tone (L%).

![Waveform, spectrogram, and F0 trace for the long imperative (C2)](image2)

Figure 2: Waveform, spectrogram, and F0 trace for the long imperative (C2) ¡Comparen sus tareas! 'Compare your homework!' of speaker MM (sentence 2_10), produced with a delayed peak (L+>H*) and a low nuclear configuration (L* L%).

3.2. Individual results

With respect to the nuclear configurations attested in C1, it is important to point out that speaker MY did not realize any short imperative at all, but rather only long imperatives. For this reason, no column exists for MY in C1 in Table 2. In contrast, speakers FT and MM did realize short imperatives, with the number of realizations being nearly identical between them (N=6 and N=7 respectively). Note that in utterances in C1, the nuclear accent is simultaneously the pitch accent on VI, meaning that the nuclear configurations include the pitch accent on VI. While MM only realized the prevalent configuration L+H* L% (100%, N=7), speaker FT shows some variation. In addition to a prevalent realization of the typical configuration L+H* L% (67%, N=4), she also realized L+>H* L% and H* L% (each 16.5%, N=1), see the two leftmost columns in Table 2.

As for C2, due to the test design, the number of utterances of long imperatives differed between the speakers. FT and MY each realized 34 long imperatives, while MM realized 28. Each speaker showed a strong preference for the prevalent pattern L* L%, though to different extents (FT: 71%, N=24; MM: 93%, N=26; MY: 61%, N=21), see the three rightmost columns in Table 2. The three speakers also showed variation with respect to the total number of realized nuclear configurations. MM only realized two: the prevalent L* L% pattern, and L+H* H% (7%, N=2). FT realized three configurations: H*L% (26%, N=9) and H* L% (3%, N=1), in addition to the prevalent L* L%. The third speaker, MY, realized seven different configurations: L+H* L% (15%, N=5), L+>H* H% (9%, N=3), H+L* L% (6%, N=2), L+H* H%, H* L%, and H* H% (each 3%, N=1), in addition to the prevalent L* L%.

In terms of the realization of the prenuclear pitch accent located on VI in C2, speakers FT and MY strongly preferred the delayed peak L+>H* (FT: 59%, N=20; MY: 64.7%, N=22), followed by L+H* (FT: 30%, N=10; MY: 29.4%, N=10) and H* (FT: 11%, N=4; MY: 5.9%, N=2). Speaker MM realized L+H* and H* (each 35%, N=10) slightly more often than L+>H* (30%, N=8).

![Nuclear configurations for each condition and each individual speaker (in %)](image3)

Table 2: Nuclear configurations for each condition and each individual speaker (in %).

4. Discussion and conclusion

Hypothesis 1 (short imperatives are realized with a rising nuclear accent and do not show any variation in pitch accent) can be considered as fulfilled. Even though three pitch accents were attested (L+H*, L+>H*, and H*), the early peak (L+H*) was realized most frequently, with only individual instances of L+>H* and H*. This finding supports the claims made by Prieto (2002) and contradicts those of Willis (2002). Willis (2002), however, did not consider short imperatives in his study. Furthermore, while the relatively small number of short imperatives in this study (N=13) may cast doubt on the robustness of the results, a comparison with Lausecker et al. (2014) reveals similar findings. Their results, which are based on a broader corpus (of scripted speech), also verify the hypothesis for short imperatives. In their study, the nuclear accent was realized either as H* or L+H* - exactly as predicted by Prieto (2002).

Hypothesis 2 (long imperatives have a low or falling nuclear configuration and a rising accent on VI, which is located in the prenuclear area) could be supported only in part: the assumption concerning the nuclear configuration could be verified, whereas the pitch accent on VI shows considerable variation. As for nuclear configuration, the significantly frequent occurrence of a low or falling pattern clearly supports the claims...
made by Prieto (2002) and contradicts those of De-la-Mota et al. (2010), who state that the nuclear configuration has a rising-falling pattern. Although we also found some circumflex contours in C2, their total number was rather small and not significant. As such, the broader corpus of this study (compared to that in De-la-Mota et al. 2010) suggests that the typical nuclear configuration is rather low or falling as opposed to rising-falling – a particularly significant finding, considering that we applied the same method as De-la-Mota et al. (2010).

As for the pitch accents on VI in C2, although we found three different pitch accents (52% L+>H*, 31% L+H*, and 17% H*) which were either rising or high pitch accents (supporting claims by Prieto 2002), their distribution was found to be more or less balanced. Consequently, considerable variation in pitch accent exists. In addition, it is a three-way variation, exactly as in Willis (2002). As such, the assumption of a non-existing pitch accent variation in H2 is falsified.

Interestingly, our results concerning pitch accent variation challenge the results for scripted speech presented in Lau secker et al. (2014), who failed to find considerable pitch accent variation in long imperatives: the realization of L+>H* accounts for 92% of the pitch accents in their study. As a result of this difference, one could jump to the conclusion that more variation exists in (semi-) spontaneous than in scripted speech (see, for example, Face 2003, who argues in favour of this view as well as Feldhausen et al. 2011 and Pešková et al. 2012 for empirical data arguing against it). This conclusion would not be entirely correct, however. Exactly the opposite can be seen in the example of the nuclear pitch accent on VI in C1. While the prevalent pitch accent in this study was found to be L+H* (85%), a generally balanced frequency between H* (55%) and L+H* (45%) was attested in Lausecker et al. (2014). Thus, a considerably greater degree of variation appears to exist in scripted speech than in semi-spontaneous speech (based on the frequency-dependent realization of the attested nuclear accents). This finding is somehow surprising, as Face (2003:125) states that the “clear-cut definitions” of when a tone is used in lab speech do not hold in spontaneous speech, as “there is more to the story than what has been discovered in studies of lab speech”. Thus, additional research is needed in order to answer the complex question of whether a greater degree of tonal variation exists in spontaneous or lab speech, specifically through the inclusion of a greater number of speakers. Furthermore, the type of spontaneous speech has to be controlled. While Face’s (2003) study is based on corpus data of spontaneous speech, the present study is based on semi-spontaneous speech as described in the second paragraph of this work. The differences in speech type may lead to international differences.

In summary, the goal of the present study was to combine different perspectives on imperatives in Mexican Spanish as established in Willis (2002) and De-la-Mota et al. (2010) and to discuss whether differences exist between spontaneous speech and lab speech. While the three-way pitch accent variation described in Willis (2002) in long imperatives could be found in our semi-spontaneous data, the typical nuclear configuration was L* 1% and not L+H* 5%, as stated in De-la-Mota et al. (2010). The L+H* 5% configuration was uttered frequently, though only in short imperatives. Our comparison between lab and semi-spontaneous speech has shown that each type of speech shows tonal variation which is not attested to the same extent in the other type of speech. This finding leads us to suggest that differences between lab speech and spontaneous speech cannot be reduced to the claim that spontaneous speech shows more variation than scripted lab speech.

5. Acknowledgements

Our gratitude goes to Pilar Prieto and the two anonymous reviewers for helpful comments. Special thanks go to Ellenit Hernández Mendoza, Miguel Linan and our subjects for their help during the recording session. Moreover, we would like to thank Audrey MacDougall for her assistance with editing.

6. References


Anticipatory nasalization in four languages: American English, French, Bosnian and Norwegian

Altijana Brkan, Angélique Amelot, Jacqueline Vaissière

Phonetics and Phonology Laboratory, LPP-CNRS, UMR7018, University Paris3 Sorbonne Nouvelle
altijana.brkan@gmail.com, angelique.amelot@univ-paris3.fr, jacqueline.vaissiere@univ-paris3.fr

Abstract

The aim of this paper is the investigation of anticipatory nasalization in CVNC sequences in four different languages: American English, French, Bosnian and Norwegian for three speakers per language. Phonological nasal vowels occur only in French. Data were recorded using a piezoelectric accelerometer to capture surface vibrations from the lateral bone of the nose. Cross-language differences of anticipatory nasal vibrations were observed: the largest anticipatory nasal vibrations in American English and the least in Norwegian. Furthermore, the results reveal that American English and Norwegian can be linked to the model of Beddor (2007). These two languages have very similar total duration of nasal vibrations while American English has a long vowel followed by a short nasal consonant and Norwegian has a short vowel followed by a long nasal consonant. Bosnian and French can be linked to the 'time-blocked model' (Bell-Berti et al., 1981) which postulates that the anticipatory nasalization start at a fixed time before the onset of the nasal consonant. American English without nasal vowels also shows data to confirm the Henke model (1966): the nasalization begins as soon as possible if compared with French where vowel nasalization is distinctive.

Keywords: anticipatory nasalization, piezoelectric accelerometer, nasal vibrations, American English, French, Bosnian, Norwegian

1. Introduction

There are at least three models for nasal anticipation. The first one, Henke's 'look-ahead' model (1966) postulates the velum to lower “as soon as it can”: an oral vowel before a nasal consonant should be completely nasalized in a language where there are no nasal vowels. The second one, the 'time-blocked model' postulates the velum to lower at a temporal fixed distance ahead from the nasal consonant (Bell-Berti et al., 1981), independently of vowel duration. For what concerns the third model, Beddor (2007) postulated that the temporal and spatial extent of the nasal gesture is relatively constant across a range of contextual conditions, although the temporal alignment of nasal and oral gestures can differ across contexts in predictable ways: the shortest the nasal consonant, the longest the duration of coarticulatory nasalization during the vowel.

It was generally predicted that the span of coarticulatory nasalization is more limited in languages with phonological distinctiveness of vowel nasalization, such as French (as in Cohn, 1990). However, some studies (Clumeck, 1976 and Solé, 1995), indicate that distinctiveness of vowel nasalization in some languages does not have reliable influence on the amount of anticipation of nasality and that the span of nasal anticipation has a language-specific nature.

The languages under investigation in the present study are French, where vowel nasalization is distinctive, American English, Bosnian and Norwegian, three languages without phonological distinctiveness of vowel nasalization. The questions under study are the following: (1) does the span of anticipatory nasal vibrations (ANV) vary across the four selected languages? (2) what is the influence of the duration of the nasal consonant or of the preceding vowel on the amount of ANV? (3) Is it possible to state that the coarticulation in some languages is either time-blocked, look-ahead, or temporal fixed?

2. Material and method

The same apparatus (a piezoelectric accelerometer fixed on the nose) and a similar corpus (nonsense words) were used for the four languages. The piezoelectric accelerometer is a noninvasive method to study nasality. Several investigators have used this instrument: Stevens et al (1975), Horii (1980, 1981), Lippman (1981).

A double piezoelectric accelerometer (KdK Sound model) was fixed on the speaker’s lateral bone of the nose (like in Lippman, 1981) to provide the “nasal signal”). The “oral signal” was recorded simultaneously using a headset microphone (AKG C420L). Each signal was recorded from an acquisition card at 44100 Hz and 16 Bits. The recordings were made in a quiet room. Prior to the recording of the corpus, the signals from the accelerometer and microphone were calibrated: the level of the recording for the piezoelectric accelerator during the production of a nasal sequence (sustained [m]) was set to be the same as the level of the recording for the oral microphone during an oral sequence (sustained [a]).

2.1. Subjects and speech corpus

The subjects were three native speakers of each of the four languages (from 22 to 35 year old). The corpus is composed of CVNC nonsense words, inspired by Beddor (2007). C={/p,t,k,b,d,g,f,s,ʃ/}, V={/a/,/e/,/i/,/o/,/u/,/ɪ/,/r/,/a/} in Norwegian, N={/m/,/n/} and /j/ are not part of the phonological system of Norwegian and we decided to exclude them so that we can have as similar data as possible for all the four languages. CVNC sequences don’t exist in French but were used to allow cross-language comparison. Each nonsense word was embedded in a carrier phrase: “Repeat (nonsense word) Kate” in English, “Dites (nonsense word) deux fois” in French, “Si (nonsense word) Katharina” in Norwegian; “Kazes (nonsense word) dva puta” in Bosnian. Subjects repeated each sequence two times.

2.2. Measurements

We measured the root mean square amplitude (RMS) of both signals. The piezoelectric accelerometer records all the
vibrations on the nose surface due to nasality but also (unfortunately) also vibrations due to voicing in oral vowels (Audibert et al., 2011). Furthermore, the amount of vibrations for the oral vowel depends on the vocal tract configuration and on the vowel. A vowel dependent “threshold” was determined to decide the frontier between vibrations due to voicing, and vibrations due to nasality for each oral vowel, measured in a non-nasal context (CoralVC) for each language. For example, in the sequence “dites pimp deux fois” (‘say pimp two times”), we use the “threshold” found for /i/ in non-nasal context.

The difference in average 91% nasalized in American English, 82% in Bosnian, 78% in French and 62% in Norwegian; \( F(3,957)= 73,537 \).

For each nonsense word, the duration of the nasal consonant (D on figure 1), and of the preceding vowel (C) were calculated from a spectrographic representation, and the duration of the anticipatory nasalization (A) and the duration of the sequence VN (B) were estimated from the nasal signal (A). We also derived the amount of anticipatory nasalization in % depending on the duration of the preceding vowel (A/C).

3. Results

Anticipatory nasal vibrations during V in sequences CVNC occur in all productions and in all languages \( (n=957) \).

3.1. Vowel acoustic duration (C)

Figure 2 illustrates V length in CVNC sequences. Mean values are: 143 ms for American English \( (n=248) \), 110 ms for Bosnian, 92 ms \( (n=240) \) for French 91 ms \( (n=235) \) and for Norwegian 91 \( (n=238) \). C is the longest in American English and the shortest in Norwegian \( (F(3,957)=84.093, p<0.0001) \). The difference is not statistically significant between French and Norwegian \( (p=0.9077) \).

The ratio between ANV and the threshold was determined \( p \) by applying a variation threshold found for /i/ in non-nasal context.

The nonsense word “pump” pronounced by a Bosnian speaker. From top to bottom: oral signal, nasal signal, RMS value derived from the nasal signal. Dotted line corresponds to the “threshold” found for the vowel /i/ in non nasal context. See text.

3.2. Nasal consonant duration (D)

Figure 3 illustrates N length in CVNC sequences. Mean values are: 157 ms for Norwegian, 107 ms for American English, 110 ms for French and 86 ms for Bosnian. The nasal consonant /m/ is the longest in Norwegian and the shortest in Bosnian \( (F(3,957)=174.529, p<0.0001) \). The difference is however not statistically significant between American English and French \( (p=0.2635) \).

3.3. Anticipatory nasalization (A & A/C)

Figure 4 illustrates ANV length in CVNC sequences. Mean values are: 137 ms for American English, 95 ms for Bosnian, 73 ms for French and 67 ms for Norwegian. Anticipatory nasal vibrations are the longest in American English: the anticipatory coarticulation starts earlier during the vowel than for the three other languages. We observed the least anticipatory nasal vibrations in Norwegian. The difference of anticipatory nasalization is significant between the four languages \( (F(3,957)=83.422, p<0.0001) \). The difference is however not significant between French and Norwegian \( (p=0.2196) \). The difference is significant between other languages \( (F\text{-test}, p\text{-value inferior to 0.0001}) \).

The ratio between ANV and the duration of the preceding vowel was calculated. Vowels are in average 91% nasalized in American English, 82% in Bosnian, 78% in French and 62% in Norwegian; \( F(3,957)= 73,537 \).
The p-value for the difference between French and Bosnian is 0.0271 while for the difference between the other languages p-value is <0.00001. The difference between 82% of nasalization for French and 79% of nasalization for Bosnian is very small and it leads us to conclude that the amount of anticipatory nasal vibrations in % depending on the duration of the preceding vowel is very similar between these two languages.

3.4. Total duration of nasal vibration (B)

The sum of the nasalization (ANV) plus the nasal consonant duration differs among languages. Mean values are: 266ms for American English, 265 ms for Norwegian, 214 ms for French and 209 ms for Bosnian; F(3,957)= 51.433. The total duration of the nasal vibrations is the longest in American English and Norwegian and the shortest in Bosnian and French. The difference between American English and Norwegian is not significant with a p-value of 0,9081. The p-value for the difference between French and Bosnian is 0.0361. However, the difference between the mean values for these two languages is small.

4. Discussion and conclusion

Our results confirm those of Clumeck (1976) and Solé (1995). Anticipatory nasalization does not depend on the phonology of the language (note however that CVNC syllable does not exist in French, a fact which may have an influence). ANV is the longest in American English and the shortest in Norwegian. The total duration of nasal vibrations are almost the same for American English and Norwegian, while American English has a long vowel followed by a short nasal consonant and Norwegian has a short vowel followed by a long nasal consonant. The constancy of the duration nasal vibrations is consistent with the hypothesis of Beddor (2007). However, according to our data from nasal vibrations, Beddor’s hypothesis is not valid for all languages. French and Bosnian data are best modeled using Bell-Berti et al.’s model (1981). French is a language with phonological distinctiveness of vowel nasalization and Bosnian isn’t. However, these two languages have similar duration of anticipatory nasal vibrations and of total nasal vibrations. These results are consistent with those of Brkan (2009).

Vaissière et al. (2010) have shown that there is no exact correspondence between different types of data (articulatory vs. aerodynamic vs. acoustic vs. perception), leading to different conclusions by the authors. Comparing languages using exactly the same instrumentation (and the same corpus) as done here is very useful. This limited study show 1) that the modeling of the spreading of nasality is very much language-dependent: the fact that the language has distinctive vowel nasalization does not seem a leading factor; 2) several models are adequate, and one model may fit one language better than the others. Beddor’s model fits the American English and the Norwegian data, with a very similar total duration of nasal vibrations while American English has a long vowel followed by a short nasal consonant and Norwegian has a short vowel followed by a long nasal consonant. The ‘time-blocked model’ (Bell-Berti et al., 1981) which postulates that the anticipatory nasalization starts at a fixed time before the onset of the nasal consonant fits well to the French and Bosnian data. Henke’s model (1966), which postulates that the nasalization begins as soon as possible, fits also the American English data.

5. Acknowledgements

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the “Investissements d’Avenir” program (reference: ANR-10-LABX-0083).

6. References


Phonetic effects in the timing of onset clusters

Jana Brunner1,2, Christian Geng1 and Adamantios Gafos1

1Universität Potsdam, 2Humboldt-Universität zu Berlin

jana.brunner@uni-potsdam.de, geng@uni-potsdam.de, gafos@uni-potsdam.de

Abstract

The temporal organization of syllable onsets has been described by the term C-center stability. The aim of the study is to investigate the influence of phonetic mechanisms on the C-center effect. EMA data of German and English singleton and cluster onsets suggest that the C-center effect depends on the coarticulatory surroundings of the onset. Furthermore, a C-center effect can almost always be found when a consonant that is aspirated as a singleton looses its aspiration in a cluster (e.g. /k/ in /sk/ in German and English).

Keywords: C-center, coarticulation, articulation

1. Introduction

The temporal organization of syllable onsets has been described by the term C-center stability. Browman and Goldstein 1988 show that, comparing a CVC sequence such as pats and a CCVC sequence such as spat, the interval from the temporal midpoint (the C-center) of the singleton onset C in pats to the onset of the coda C is comparable in duration to the interval between two successive temporal midpoint of the complex onset CC in spats to the onset of the coda C. In other words, the /p/ in spats is shifted towards the vowel; it is assumed to overlap more with the vowel than the /p/ in pats. As a result of that, the vowel, as it can be measured in speech data, is shorter in spats than in pats. According to the Articulatory Phonology literature the C-center effect is the result of competing coupling relations between, on the one hand, each consonant with the vowel and, on the other hand, the consonants with each other (Browman and Goldstein 1988).

Previous studies investigating the C-center effect for example in English (Marin and Pouplier 2010), Italian (Hermes et al. 2008), German (Pouplier 2012) and other languages allowing for complex onsets gave inconsistent results. In some cases the effect was found, in others it was not. The aim of the present study is to investigate in how far this inconsistent evidence is a result of two phonetic mechanisms (i.e. coarticulation and oral-laryngeal coordination) influencing articulatory patterns.

Coarticulation. Bladon and Al-Bamerni 1976 show that sounds differ with respect to their coarticulatory resistance. Some sounds such as /s/ are not easily influenced by their surroundings, but the articulation of others, such as laterals, depends very much on the context. If a consonant with little resistance occurs in a cluster it is usually different from the same phoneme occurring as a singleton. As a result of that, for sounds with little resistance, the path the tongue has to travel to reach the vowel differs between the singleton and cluster item. /l/ for example, is less resistant than /g/. An /l/ following /g/ therefore has a higher tongue position than a singleton /l/. The distance the tongue needs to travel to reach a vowel following /l/ might therefore be larger coming from the cluster /gl/ than coming from singleton /l/. Taking into account the widely attested relationship between movement duration and movement amplitude (e.g. Nelson 1983), a larger movement needs more time, and this might prevent a C-center effect. The first aim of the study is to investigate the influence of a difference in movement amplitude between a CV and a CCV item on the C-center-effect. If the articulator (tongue or lip) has to travel more in the CV item than in the CCV item we expect a C-center effect whereas if it travels less in the CV item than in the CCV item we expect no C-center-effect.

Oral-laryngeal coordination. Oral-laryngeal timing differs in clusters as compared to singletons. In the German cluster /sk/ for example, there is just one glottal opening and closing gesture, resulting in a lack of aspiration for /k/. Singleton /k/, on the other hand is aspirated. Earlier studies consistently report a C-center effect for German /sk/ (e.g. Pouplier 2012) and for English /sk/ and /sp/ (Marin and Pouplier 2010, Goldstein et al. 2009), another language where aspiration disappears if the sound occurs as the second component of a cluster. These consistent findings could be explained in a very simple manner: due to the aspiration the singleton stop is longer, so it starts articulatorily and acoustically earlier relative to the end of the vowel as compared to the shorter cluster stop. The second aim of the study is to investigate this difference between clusters where aspiration disappears and other clusters.

2. Methods

Articulatory movements of six German and eight American English native speakers were recorded via Electromagnetic Articulography (NDI Wave system). The German corpus consisted of items with the clusters /gl/, /kv/, /pl/ and /sk/ and the corresponding singletons. Examples of the target words are Plätze; Plätze; Plätze. The English corpus contained items with the clusters /kl/ and /sk/ and the corresponding singletons. Examples are - Lätze; queren - wehren. As a result of that the /kl/ in the German cluster /kl/ for example, is less resistant than /g/. An /l/ following /g/ therefore has a higher tongue position than a singleton /l/. The distance the tongue needs to travel to reach a vowel following /l/ might therefore be larger coming from the cluster /gl/ than coming from singleton /l/. Taking into account the widely attested relationship between movement duration and movement amplitude (e.g. Nelson 1983), a larger movement needs more time, and this might prevent a C-center effect. The first aim of the study is to investigate the influence of a difference in movement amplitude between a CV and a CCV item on the C-center-effect. If the articulator (tongue or lip) has to travel more in the CV item than in the CCV item we expect a C-center effect whereas if it travels less in the CV item than in the CCV item we expect no C-center-effect.

Coarticulation. Bladon and Al-Bamerni 1976 show that sounds differ with respect to their coarticulatory resistance. Some sounds such as /s/ are not easily influenced by their surroundings, but the articulation of others, such as laterals, depends very much on the context. If a consonant with little resistance occurs in a cluster it is usually different from the same phoneme occurring as a singleton. As a result of that, for sounds with little resistance, the path the tongue has to travel to reach the vowel differs between the singleton and cluster item. /l/ for example, is less resistant than /g/. An /l/ following /g/ therefore has a higher tongue position than a singleton /l/. The distance the tongue needs to travel to reach a vowel following /l/ might therefore be larger coming from the cluster /gl/ than coming from singleton /l/. Taking into account the widely attested relationship between movement duration and movement amplitude (e.g. Nelson 1983), a larger movement needs more time, and this might prevent a C-center effect. The first aim of the study is to investigate the influence of a difference in movement amplitude between a CV and a CCV item on the C-center-effect. If the articulator (tongue or lip) has to travel more in the CV item than in the CCV item we expect a C-center effect whereas if it travels less in the CV item than in the CCV item we expect no C-center-effect.

Oral-laryngeal coordination. Oral-laryngeal timing differs in clusters as compared to singletons. In the German cluster /sk/, for example, there is just one glottal opening and closing gesture, resulting in a lack of aspiration for /k/. Singleton /k/, on the other hand is aspirated. Earlier studies consistently report a C-center effect for German /sk/ (e.g. Pouplier 2012) and for English /sk/ and /sp/ (Marin and Pouplier 2010, Goldstein et al. 2009), another language where aspiration disappears if the sound occurs as the second component of a cluster. These consistent findings could be explained in a very simple manner: due to the aspiration the singleton stop is longer, so it starts articulatorily and acoustically earlier relative to the end of the vowel as compared to the shorter cluster stop. The second aim of the study is to investigate this difference between clusters where aspiration disappears and other clusters.

Three sensors were placed midsagittaly on the tongue in equal distances, one sensor was placed at each lip, one below the lower incisors. In order to correct for head movements further sensors were attached to immobile structures (behind each ear, on the bridge of the nose and above the upper incisors). After the recording the data were head corrected and rotated to the occclusal plane. Velocities and lip aperture were calculated.

Consonantal plateaus were measured as 80% velocity thresholds of the vertical tongue movement towards and away from the consonantal target position (cf. for example Marin and Pouplier 2010). For bilabial sounds the consonantal plateau was
calculated on the lip aperture signal.

Vowel compression. A C-center analysis was carried out by measuring the difference in vowel length between the CCV and the CV sequence (e.g. Murin and Pouplier 2010). Vowel length was calculated as the interval from the offset of the consonantal plateau of the prevocalic consonant to the onset of the coda consonant. Afterwards, vowel compression was calculated as the difference between the vowel length of the CV item (e.g. Lütze) and the CCV item (e.g. Plätze). Positive vowel compression values mean that the vowel is shorter in the cluster than in the singleton item (as would be expected for a C-center effect), negative values mean that the vowel is longer in the cluster than in the singleton item (no C-center effect).

Movement amplitude difference. In order to investigate the influence of the distances to travel in the CV and the CCV item on the C-center effect, the amplitude of the movement from the prevocalic consonant to the vowel was calculated as the sum of Euclidean distances between the sample points for the CCV and the CV items. The difference between these movement amplitudes was calculated. If this value is positive there was more movement in the CV than in the CCV item (which would facilitate a C-center-effect because the shorter distance in the CCV item can easily be travelled in less time), if it is negative there was more movement in the CCV item. This would make it difficult to produce a C-center effect since a larger distance in the CCV item must be travelled in less time.

3. Results

3.1. Vowel compression

Figures 1 and 2 give the results for vowel compression for all the clusters in German and English respectively. Positive values mean that the vowel in the cluster was shorter (C-center-effect). For German the C-center effect is consistently found for /kv/ and /sk/, but not for /gl/ and /pl/. A linear mixed model with vowel compression as dependent variable and speaker and vowel as random effect showed that the influence of the cluster was significant. Vowel compression was estimated to be 29 ms larger for /sk/ as compared to /gl/; /kv/ was 14 ms larger than /gl/ and /pl/ was 2 ms larger than /gl/.

In the English data the C-center effect is more consistent, almost all the values are positive. A linear mixed model was calculated as described above. The influence of the cluster on vowel compression was significant. Vowel compression was estimated to be 27 ms longer in /sk/ than in /kl/.

Figure 1: Vowel compression (German) for the four clusters investigated. Left: onsets preceding tense vowels, middle: onsets preceding lax vowels, right: onsets preceding diphthongs. Positive values: CCV-vowel is shorter than CV-vowel (C-center effect). No /gl/ clusters preceding tense vowels were available.

Figure 2: As figure 1 but for English. No /sk/ clusters preceding tense vowels were available.

3.2. Movement amplitude vs. vowel compression: German

Figure 3 shows the relation between movement amplitude difference (abscissa) and vowel compression (ordinate) for the contrast /kv/-/kv/. Values >0 on the abscissa mean that the tongue travels more in the singleton (favouring a C-center effect), values <0 mean that the tongue travels more in the cluster item. Values >0 on the ordinate mean that the singleton vowel is longer (C-center effect), values <0 mean that the cluster vowel is longer. There is a correlation between the two parameters: if the articulator travels more in the singleton than in the cluster item the vowel compression values are higher. In other words, a C-center effect is more likely in this case. In most cases the movement amplitude values are above 0, so the tongue travels more in the singleton item. Nearly all the exceptions are productions of queilen-weilen (white squares) where the onset is followed by a low vowel (first component of the diphthong). This is easily explained: the tongue comes from a higher position in the cluster than in the singleton, so it has to travel further to reach the vowel in the cluster than in the singleton. This results in negative movement amplitude values. Unless the speaker interferes by varying the velocity a larger distance takes more time to travel, this results in negative vowel compression values.

The results for the German clusters /gl/ (Figure 4) and /pl/

Figure 3: kv-clusters (German)
(not shown here, \(\rho = 0.529, r = 0.000\)) differ from /kv/ in that there are many more items with a negative movement amplitude and, consequently, negative vowel compression values. The difference between /kv/ on the one hand and /gl/ and /pl/ on the other is therefore not to be found in the phonological structure of the clusters or the language German but simply in articulatory configurations that either favour or disfavour a C-center effect.

In both languages a very clear C-center effect was found for

---

3.3. Movement amplitude vs. vowel compression: English

For English /kl/ and /sk/ the same effect can be found: the greater the difference in movement amplitude the more likely is a C-center effect. Compared to the German data, however, the correlation is less strong. Looking at /kl/ for example (Figure 5) there is more dispersion. For clean-lean (triangles) a C-center effect was often not found whereas for climb-lime (white squares) it was almost always observed. What is remarkable as well is that although the correlation between vowel compression and movement amplitude difference exists for English as well it cannot be linked to the height of the vowel. For German we have shown that, for our material with a carrier phrase ending in a low vowel, a C-center-effect is more likely if the onset is followed by a high vowel (compare queren-wehren to queilen-weilen). For the English /kl/ clusters the opposite can be observed: for clean-lean there was often not C-center effect whereas for climb-lime there usually was. More in depth analysis of the data has shown that in English the vowel is more strongly coarticulated than in German. Whereas in queren-wehren the tongue position during /ei/ is approximately the same, the tongue is much higher in the first component of the diphthong /ai/ in climb than in lime. As a result of that the tongue has to travel more in lime although it is lower.

---

3.4. Influence of velocity

One other possibility to produce a C-center effect when the tongue has to travel further in the CCV sequence than in the CV sequence is to increase the velocity. In order to investigate whether speakers increase the velocity in the CCV sequence, the average velocity during the movement from the prevocalic consonant to the anchor was calculated. Afterwards a linear mixed model was fitted with velocity as dependent variable and complexity (singleton or complex onset) and vowel duration as fixed factor. Cluster and speaker were entered as random factors. For German the model suggests that the mean velocity is 2.9 mm/s lower in CV sequences than in CCV sequences (\(N=1436\)). The effect is significant. For English the model suggests that the velocity is 1.6 mm/s lower in the CV item. This effect is not significant (\(N=585\)).

The significant result for German suggests that, if the vowel duration is the same in the CV and CCV sequence, speakers speed up in the CCV sequences, possibly in order to produce a C-center effect. Looking at the estimates, however, the differences in velocity cannot account for the vowel compression values shown in figure 1. A velocity difference of 2.9 mm/s means that in an interval of 30 ms (which corresponds to the vowel compression in /kv/) the tongue moves 0.09 mm more in the CV than in the CCV item. This is below the accuracy of EMA (1 mm). And compared to the movement amplitude differences discussed here (up to 20 mm) this amount of movement is meaningless.

For the present purpose the question is not so much whether the velocity is higher in a CCV than in a CV sequence when the vowel duration is the same but whether the velocity in the CCV sequence increases when the vowel is longer in the CCV sequence than in the CV sequence. The negative slopes for vowel duration (-37.39 for German, -76.15 for English), however, suggest that within a speaker and item pair velocity and vowel duration correlate negatively. If the vowel duration increases the velocity decreases so that the difference in velocity between a CCV sequence with a longer vowel and a CV sequence with shorter vowel is even smaller than when the vowel durations are the same.

---

3.5. Oral-laryngeal coordination

Both the German and the English cluster /sk/ are different from the other clusters (Figures 6 to 7). Looking at the German data, on first sight the patterns looks similar to /kv/, but there are even less points with a vowel compression and a movement amplitude below 0. In the English data for /sk/ the majority of data points has a movement amplitude above 0 too.

In both languages a very clear C-center effect was found for
/sk/-/fk/, much clearer than for the other clusters. This effect is independent of the environment. The difference in vowel length between the CV and the CCV item (Figures 1 and 2) is between 20 and 60 ms, which is a typical range for aspiration in the two languages. It is therefore likely that, although the movement amplitude certainly plays a role for the timing of the articulatory movements, the clear C-center effect in this pair is influenced by the lack of aspiration in the cluster.

4. Discussion

The present study shows that the C-center effect depends on phonetic mechanisms. The C-center effect is found in cases where it is easy to produce, i.e. when the movement from the prevocalic consonant to the vowel is shorter in the cluster than in the singleton or when aspiration in the cluster gets lost. The analysis of velocity differences gave unclear results. Whereas the velocity is significantly lower in German CV than in CCV items, which would be in line with the assumption that speakers speed up in the CCV items in order to produce a C-center effect, the estimated difference cannot account for the entire C-center effect.

Our results suggest that coarticulatory behaviour influences the timing in syllable onsets. For the same cluster a C-center effect might be observed with one vowel but not with another one. This effect can be observed in earlier data as well. Hermes et al. 2008 for example, found a C-center effect in Italian rimaprima (with a high vowel), but in rema-prema (with a lower vowel) the effect was less clear.

Our results furthermore suggest that clusters involving sounds that lose their aspiration when they become part of a cluster should reliably show a C-center effect simply because the loss of aspiration causes a right shift of the prevocalic consonant. Earlier studies have frequently found a C-center effect for exactly those clusters. Brownman and Goldstein 1988, Honorof and Browman 1995 and Goldstein et al. 2009 found a C-center effect for /sp/ in English, and Marin and Pouplier 2010 found one for /sk/ and /sp/ in English. Importantly, no C-center effect was observed for French /sp/ (Tilsen et al. 2012) and Italian /sp/ (Hermes et al. 2008), and we suggest that this is because French and Italian voiceless stops are frequently unaspirated even as singletons.

5. Acknowledgements

This work was supported by an Advanced ERC grant 249449 to A. Gafos.

6. References


Oralofacial gestures and emotion: Articulatory feedback or sound symbolism?

Francesco Cangemi¹, Martine Grice¹, Ralf Rummer², Judith Schwegge²

¹Phonetik, University of Cologne
²Psychology, University of Erfurt
{fcangemi,martine.grice}@uni-koeln.de, {ralf.rummer,judith.schwegge}@uni-erfurt.de

Abstract

Research on embodied emotion suggests a link between muscle activation and emotional state. For example, activation of the muscles involved in smiling leads subjects to find the same objects funnier than when these muscles are not activated. Strack et al. (1988) provided evidence with a non-linguistic task, asking subjects to hold a pen with their teeth or their lips while rating cartoons. Rummer et al. (in press) expanded the paradigm to speech production, asking subjects to articulate [i] or [o] while providing ratings. However, the effect of vocal on mood could also be mediated by phonosymbolic aspects. In order to control for this potential confound, in this study we further expand the paradigm to a listening experiment, in which subjects provide ratings while listening to repeated articulation of [i] or [o]. Results show an effect of vowel type in the production experiment only. This is incompatible with an interpretation based on sound symbolism, but does not rule out the possibility of a link between vowel perception and emotional state, especially in frameworks suggesting a recruitment of the motor system during speech perception.

Keywords: embodied cognition, sound symbolism, frequency code, articulatory feedback, motor theory

1. Introduction

1.1. Embodied emotions

Research on embodied cognition sees cognitive processes as deeply rooted in the body’s interactions with the world (Wilson 2002). Work in this framework appears to elaborate on Gratiot’s (1865: 65-67) early claim of a bidirectional link between thoughts and actions, quickly incorporated by Darwin (1872) in the contention that emotions are intensified by their overt expression, and later recast by Zajonc and Markus (1984) as a critique of the computational theory of mind. One segment of research in this framework has concentrated on emotion and peripheral physiological (re)actions (Niedenthal 2007) in particular on the link between emotion, orofacial muscular activity and, more recently, language. Pioneering experimental work by Strack et al. (1988) suggests that activation of the Zygomaticus Major Muscle (ZMM), which is involved in smiling and laughing, has an effect on emotional state (facial feedback hypothesis). They had subjects rate the funniness of cartoons in two conditions: (A) while holding a pen with their teeth (thus activating ZMM) and (B) while holding a pen with their lips (thus activating the Orbicularis Oris Muscle, OOM, an antagonist of ZMM). Funniness ratings for condition A were significantly higher than for condition B.

This paradigm was extended to the study of language by Havas et al. (2007), who showed that the reenactment of congruent emotions through muscle activation facilitates sentence comprehension. They had subjects judge the pleasantness of events described in sentences, again while (A) holding a pen with their teeth, thus activating ZMM, or (B) holding a pen with their lips, thus activating OOM. Reading times where faster when sentence content and facial posture were matching, e.g. for subjects reading about pleasant events while holding a pen with their teeth.

A more recent extension of the Strack et al. (1988) paradigm to speech production is provided by Rummer et al. (in press). ZMM and OOM are also used in speech production, notably in the articulation of vowels. Rummer et al. (in press) asked subjects to rate the funniness of cartoons in the two original pen-holding conditions, and in two novel conditions: (A) while repeatedly articulating [i] (activating ZMM) and (B) while repeatedly articulating [o] (activating OOM). Results for the pen-holding conditions, though without reaching significance, are compatible with those from Strack et al. (1988). Moreover, as for the vowel articulation conditions, funniness ratings for condition A’ were significantly higher than for condition B’. The authors conclude that vowel articulation affects emotional state through muscle activation (articulatory feedback hypothesis).

1.2. Sound symbolism

An alternative explanation for the findings by Rummer et al. (in press) is that the effect of vowel type on emotional state is not mediated by articulatory feedback, but rather by sound symbolism – that is, by non-arbitrary relationships between signifier and meaning. The arbitrary nature of linguistic signs has been one of the pillars of structuralism in particular and modern linguistics in general (e.g. de Saussure 1916), but it has been questioned at multiple points in time and from different perspectives (Sapir 1929, Bolinger 1949, Ertel 1969, Hinton et al. 1994, Ramachandran and Hubbard 2001).

Within this framework, work on size symbolism is particularly relevant to the evaluation of the articulatory feedback hypothesis. Ohala (1980, 1994) suggests a link between resonant frequencies, pitch and vocal tract size (frequency code). From an ethological perspective, vocalizations with [i]-like and high pitched sounds (which are characteristic of small vocal tracts) might have been associated with small size and appeasing or submissive stance. Facial expressions corresponding to the articulation of [i] are used by primates to a similar effect. On the other hand, “o-faces” and vocalizations with [o]-like and low pitched sounds are used to appear larger and more aggressive or threatening. It is difficult to say whether frequency code and articulatory feedback hypothesis should be seen as complementary or as mutually exclusive. In the first interpretation, size symbolism would provide phylogenetic grounding for the mechanics of smiling, which in turn provide ontogenetic grounding for the articulatory feedback hypothesis. However, from an embodied cognition perspective, it is not clear how submissiveness could be equated to pleasantness, since reenacting a submissive stance should also entail reenacting a threatening context. In the second perspective, the two hypotheses would provide two
different explanations for the findings in Rummer et al. (in press). The effect on emotional state reflected by pleasantness ratings is driven either by muscle activation or by a fossilized but ethologically motivated association between resonances, pitch, vocal size and stances.

1.3. Rationale

In order to test these two claims, we expand the Strack et al. (1988) paradigm to a new condition, in which muscle activation is not required, but phonosymbolic aspects are maximized. We thus compare results from a production experiment, in which subjects rate cartoons while repeatedly articulating either [i] or [o] (conditions A’ and B’), as in Rummer et al. in press), with results from a perception experiment, in which subjects rate cartoons while listening to a repeatedly articulated high-pitched [i] or low-pitched [o] (conditions A’’ and B’’).

Table 1: Synopsis of studies, conditions, hypotheses.

<table>
<thead>
<tr>
<th>Strack</th>
<th>Havas</th>
<th>Runmer</th>
<th>This study</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: teeth</td>
<td>A: teeth</td>
<td>A’: i</td>
<td>A’: i</td>
</tr>
<tr>
<td>B: lips</td>
<td>B: lips</td>
<td>B’: o</td>
<td>B’: o</td>
</tr>
</tbody>
</table>

If driven by articulatory feedback (H1), the effect of vowel type on emotional state is only expected in the production experiment. Funniness ratings (R) are expected to be higher for subjects articulating [i] (A’) than articulating [o] (B’). If driven by sound symbolism (H2), the effect of vowel type on emotional state in the perception experiment is expected to be at least equal in size to the effect in the production experiment. Funniness ratings (R) are also expected to be higher for subjects listening to [i] (A’’) than listening to [o] (B’’).

Table 2: Predictions.

<table>
<thead>
<tr>
<th>Feedback Hypothesis</th>
<th>Embodied cognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fun</td>
<td>Articulatory Feedback Hyp.</td>
</tr>
</tbody>
</table>

2. Method

2.1. Participants and procedure

448 students at the University of Erfurt were asked to rate the funniness of cartoons in two experiments. The experiments took place in sound proof cubicles; each participant was tested in separately. In both experiments, participants provided ratings for cartoons on a 9-point Likert Scale, ranging from 1 (not funny at all) to 9 (very funny). The cartoons were the same as in Rummer et al. (in press), experiment 2. In a pretest, 24 cartoons out of 50 were selected around the midpoint of the funniness scale (M=5.10, SD=.33). All participants were presented with the same cartoons. After two practice trials, the cartoons were separately presented on a computer monitor, each with a Likert Scale in the form of 9 radio buttons below.

As soon as the participant clicked on one of these buttons, the next cartoon (and scale) appeared on the screen, and so on.

2.2. Experimental conditions

In the first experiment (production task), while reading and rating the cartoons, 112 participants articulated the vowel [i] approximately once per second, and 112 participants articulated the vowel [o] approximately once per second. This design replicates two of the four conditions in Rummer et al. (in press), experiment 2.

In the second experiment (perception task), while reading and rating the cartoons, 112 participants listened to repeated articulation of the vowel [i], and 112 participants listened to repeated articulation of the vowel [o]. For each vowel type, a single sound file was recorded and presented to subjects once per second. Both vowels were chosen among several tokens uttered by the same female native speaker of German. The chosen pair maximized phonosymbolic differences with respect to both pitch (average f0 of 252Hz for [i] and 207Hz for [o]) and resonant frequencies (e.g. average F2 of 2290Hz for [i] and 802Hz for [o]).

3. Results

Figure 1 shows counts (x-axis) for funniness ratings (y-axis) pooled across subjects and cartoons, and split by task (perception: left panel; production: right panel) and vowel ([i]: black bars, [o]: grey bars). The results indicate an effect of vowel type on funniness ratings in the production task (right panel): compared to counts for [o] (grey bars), counts for [i] (black bars) are higher for positive funniness values (i.e. from 5 to 9) and lower for negative funniness values (i.e. from 1 to 4). This is not the case in the perception task (left panel).

Figure 1: Funniness ratings pooled across subjects and cartoons. Counts split by vowel and task.

The effect of vowel reaches significance in the production task but not in the perception task. As for production, a Likelihood Ratio Test reveals a significant difference between mixed effects models with random slopes and intercepts for cartoons and subjects, and predicting funniness ratings using (a) vowel or (b) no fixed term (p = 0.05). Funniness ratings for [o] are significantly lower than for [i] (respective mean values are 5.03 and 5.53 on the 9-point scale).
As for perception, a Likelihood Ratio Test reveals no significant difference between mixed effects models with random slopes and intercepts for cartoons and subjects, and predicting funniness ratings using (a) vowel or (b) no fixed term (p = 0.65). Funniness ratings for [o] are however slightly lower than for [i] (respective mean values are 5.2 and 5.33 on the 9-point scale). Effect size analysis using Cohen’s d supports this finding, showing a small effect in production (d = 0.39) but a negligible effect in perception (d = 0.09).

However, a joint analysis shows no significant interaction between task (production vs perception) and vowel ([i] vs [o]). A Likelihood Ratio Test between mixed effects models with random slopes and intercepts for cartoons and subjects, and predicting funniness ratings using (a) task and vowel and their interaction or (b) task and vowel only did not reach significance (p = 0.35).

4. Discussion

4.1. Articulatory feedback and frequency code

Our results are interesting in three respects. First, model comparisons for the production task replicate findings from Rummer et al. (in press) on the effect of articulated vowel type on emotional state (i). Moreover, model comparisons for the perception task and effect size analysis for both tasks fail at detecting an effect of perceived vowel type on emotional state (ii). Taken together, these findings seem to support the hypothesis that the effect of vowel type on emotional state is indeed mediated by articulatory feedback, rather than by phonosymbolic associations (§1.3, H1).

However, in the absence of a significant interaction between task and vowel in the joint analysis, and given the small effect sizes, it would not be cautious to dismiss the hypothesis of an effect of perceived vowel type on emotional state altogether. That is, while emphasizing the role of articulatory feedback over that of sound symbolism in the link between vowel type and emotional state (ii), our results also point to the necessity of further investigating the perceptual aspects of this link (iii).

4.2. Articulatory feedback and motor theory

Indeed, an effect of perceived vowel type on emotional state does not need to be interpreted as evidence for sound symbolism. It could also be accounted for by theories linking speech production and perception, and more specifically by theories positing recruitment of the motor system during speech perception. Recruitment of the motor system during perception has been suggested in the framework of the motor theory of speech perception (Lieberman and Mattingly 1985, *inter alia*) and has received experimental support by a number of studies featuring transcranial magnetic stimulation (Roy et al. 2008, Meister et al. 2012). Fadiga et al. (2002), for example, show that subjects’ motor-evoked potentials recorded from tongue muscles are higher when listening to words containing apical trills (whose production require strong tongue tip activation) than when listening to words containing labiodental fricatives.

In this scenario, perception of a certain vowel type might entail (simulation or facilitation of) activation of the muscles involved in its production, which in turn could trigger articulatory feedback. The effect can be expected to be smaller than the one appreciable in production (Table 3, H3). It is actually difficult to say whether vowel perception would grant a covert muscle activation which is strong enough to yield an appreciable effect on emotional state (which, in our paradigm, should in turn be reflected by significant differences in funniness ratings).

However, recasting a possible perception effect of vowel type on emotional state in terms of motor theory of speech perception rather than in terms of sound symbolism has another advantage. Early accounts of the motor theory suggested that recruitment of the motor system is specific to speech perception. This is in contrast with other theories on the link between perception and action (e.g. Fowler, 1996), and has been extensively questioned in the light of further developments of the motor theory itself (Galantucci et al. 2006). However, this perspective has the advantage of providing an additional testing ground for our hypothesis.

In our study, the stimuli employed in the perception task were long sustained vowels, which do not qualify as prototypically speech-like as words (or, better yet, nonce words, which have the advantage of minimizing semantic confounds). If recruitment of the motor system involved in speech articulation is specific to perception of speech, one might expect articulatory feedback to be stronger with nonce words than with isolated vowels (H4). However, the articulation of non-words necessarily entails activation of a high number of orofacial muscles, thus reducing the proportional activation of the ones specifically relevant to the articulatory feedback hypothesis. If the effect of perceived vowel type on emotional state reflects the degree of activation of the relevant muscles, one might expect articulatory feedback to be stronger with isolated vowels than with nonce words (H5).

Table 3: New predictions.

<table>
<thead>
<tr>
<th></th>
<th>H3</th>
<th>H4</th>
<th>H5</th>
</tr>
</thead>
</table>
|   | R_v > R_b | R_v ≥ R_b | R_v ≥ R_b

By allowing the formulation of clear predictions, the exploration of differences between nonce words and isolated vowels in perception sounds thus particularly promising for future research on the articulatory feedback hypothesis.

5. Acknowledgements

We would like to thank the two anonymous reviewers for valuable comments and Bodo Winter for discussing the quantitative analyses. For collecting the data we would like to thank René Schlegelmilch.

6. References


Gratiolet, P. (1865). *De la physionomie et des mouvements*


Prosodic marking, pitch and intensity in spontaneous phonological self-repair in Dutch

Paul Carter1, Leendert Plug2

1University of Sheffield, United Kingdom
2University of Leeds, United Kingdom

paul.carter@sheffield.ac.uk, leendert.plug@leeds.ac.uk

Abstract

We report on a phonetic analysis of instances of spontaneous phonological error repair sampled from Dutch spontaneous speech. We investigate whether phonological error repairs can be ‘prosodically marked’ and what factors constrain repair prosody. Previous studies of ‘prosodic marking’ in self-repair have suggested that while lexical error repairs are regularly realized with marked prosody, phonological error repairs are generally unmarked. Moreover, it has been asserted that the temporal organization of a repair has no bearing on its pitch and intensity characteristics. Our findings suggest that in fact, phonological error repairs are realized with prosodic marking as frequently, and through the same acoustic correlates as lexical repairs. We also show that repair timing is a significant predictor of prosodic marking judgments and F0 and intensity measurements.

Keywords: self-repair, prosody, spontaneous speech

1. Introduction

This paper reports on a phonetic analysis of instances of spontaneous phonological error repair such as sa ... fat soap, in which a mispronunciation is corrected. We address two questions: first, whether phonological error repairs can be ‘prosodically marked’; second, what factors constrain repair prosody. The concept of prosodic marking in repair was first introduced by Cutler (1983), who describes a collection of spontaneous speech error repairs. She describes an ‘unmarked’ repair as one in which the pitch, intensity and speaking rate of the repair component — in the case of phonological error repair, the correct pronunciation of the target word — are not noticeably different from those of the reparandum — in the case of phonological error repair, the erroneous target word attempt. A ‘marked’ repair, on the other hand, ‘is distinguished by a quite different prosodic shape from that of the original utterance’ (Cutler 1983: 81). By leaving a repair unmarked, the speaker ‘minimises the disruptive effect of the error on the utterance as a whole’, while marking assigns ‘salience’ to the correction (Cutler 1983: 80).

Cutler (1983: 83) states that unlike lexical repairs, in which one lexical item is rejected in favor of another, phonological error repairs are, as a rule, unmarked: her data set contains no prosodically marked phonological error repairs. Levelt (1989: 495) reiterates the claim, and emphasizes that repair prosody is ‘semantically motivated’. According to Levelt, a speaker’s decision to mark or not to mark a repair is made at the message level. The greater the perceived semantic contrast between two lexical items, the greater is the likelihood of a decision to assign ‘contrastive prominence’ to one of the items. Phonological error repairs do not involve semantic contrast at the message level, as they involve two attempts at a single target word — hence their consistently unmarked form.

While Levelt’s reasoning is appealing, there is reason to doubt that Cutler’s observations generalize beyond her data set. Shattuck-Hufnagel & Cutler (1999:1485) report that in their data set, prosodic marking is ‘less likely’ in phonological error repair than in lexical repair, but the difference does not reach statistical significance. This must mean their data set does contain prosodically marked phonological error repairs. Furthermore, Nooteboom (2010) reports a pattern of prosodic differentiation of phonological error repairs that suggests that at the very least, labeling them all ‘prosodically unmarked’ is a simplification.

With reference to the second question, Cutler (1983), Levelt & Cutler (1983) and Levelt (1989) have described semantic factors that constrain speakers’ choices for or against prosodic marking in lexical repair. They explicitly describe the choice as ‘orthogonal to the time course of error detection and correction’ (Cutler 1983: 81), or the ‘interruption-and-restart structure of the repair’ (Levelt & Cutler 1983: 211). Still, in Nooteboom’s (2010) pattern of prosodic differentiation among phonological error repairs, it is exactly repair timing that is the most significant factor: Nooteboom observes that instances in which the repair comes in very early, as in sa ... fat soap, tend to have a repair component with a high pitch and intensity prominence on the first vowel compared with the reparandum. Instances in which the mispronounced word is completed before the onset of repair, as in sat soap ... fat soap, tend to have a repair component with a low pitch and intensity prominence on the first vowel compared with the reparandum. Furthermore, Kapatsinski (2010) has shown that highly frequent words resist interruption of the reparandum item in lexical repair: in other words, a repair’s timing is correlated to some extent with its lexical frequency contour.

These considerations warrant detailed analysis of the prosody of phonological error repair in spontaneous speech. In this paper we model the pitch and intensity characteristics of a collection of phonological error repairs sampled from the Spoken Dutch Corpus, complementing a previous study of lexical repairs drawn from the same corpus (Plug & Carter 2013). We derive the characteristics from auditory judgments of prosodic marking, following Cutler (1983) and Levelt & Cutler (1983), as well as acoustic measurements, following Nakatani & Hirschberg (1994), Nooteboom (2010) and others. We explore the relationship between the auditory judgments and measurements, and evaluate the role of temporal and frequency-related factors in accounting for both.

2. Data and method

2.1. Data

The data for this paper comprise 325 instances of phonological error repair extracted from four sub-corpora of the Spoken Dutch Corpus containing spontaneous speech. We extracted instances of speech which were coded as mispronounced or
interrupted and did a number of additional, unsystematic data trawls. We only included repaired mispronunciations containing at least one consonant and one vowel with primary or secondary lexical stress. We left aside repairs occurring in utterance-initial and utterance-final positions. We included instances ambiguous between phonological and lexical repair if the immediate context contained a plausible trigger for phonological error. Representative examples include [b]aartbij – [w]aartbij ‘with which’, van[a][l] de – van[a][l] de ‘from the’ and mer[e]rol- mer[ero]logisch ‘meteorological’.

2.2. Acoustic analysis

We segmented all instances, placing boundaries at the start and end of the erroneous target word attempt and the start and end of the repair stretch. We included any lexical items following the target word attempt in the reparandum, for the purpose of calculating target-to-offset and target-to-repair durations. We delimited all vowel portions within the erroneous target word attempt and correct realization. We labeled the first vowel with primary or secondary stress separately.

We measured f0 (in Hertz) and intensity (in decibels) at every millisecond across the segmented vowel portions, and log-transformed f0 values. We then calculated mean, median and maximum values. We did this for the first stressed vowels in the erroneous target word attempt and its correction, and across all of the vowels in these repair components. In each case we calculated a delta value by subtracting the value derived from the erroneous target word attempt from the value derived from the correct production. This yields a measure of the prosodic difference between the crucial components of the repair, as well as introducing some speaker normalization.

2.3. Auditory analysis

Following Levelt & Cutler (1983), we classified all instances as prosodically marked or unmarked based on auditory analysis. The question in each case was whether the correct target word realization sounds particularly salient because of its pitch or loudness, or both, relative to the erroneous attempt. We allowed for the intermediate classification of ‘possibly marked’ (see Plug & Carter 2013).

The classification was done by two raters: the second author and a Dutch discourse analyst with no particular knowledge of the phonetics of self-repair. The two raters classified all instances independently. They reached the same judgment in 250 cases (77%). Of the 75 instances for which the raters proposed a different classification, 24 involved one rater proposing ‘possibly marked’ and the other ‘marked’. In order not to overestimate the proportion of prosodically marked repairs, we coded these instances as ‘possibly marked’. The remaining 51 instances either involved ‘possibly marked’ vs ‘unmarked’ or ‘marked’ vs ‘unmarked’. All of these instances were reconsidered independently by both raters. In nine cases, this resulted in straightforward agreement, while in 42, the raters confirmed their initial judgments. Remaining cases of ‘possibly marked’ vs ‘unmarked’ were coded as ‘unmarked’: remaining cases of ‘marked’ vs ‘unmarked’ as ‘possibly marked’. In what follows, we will refer to the marking classification by its variable name, Prosodic marking.

2.4. Temporal analysis

In order to assess whether repairs with an erroneous target word attempt that is interrupted have different prosodic characteristics from repairs with a completed attempt, we classified each reparandum item as interrupted or completed prior to repair (Completeness). In addition, we explored the relevance of continuous measures, on the assumption that these might capture more fine-grained differences between ‘early’ and ‘late’ repairs. First, we measured the duration from the start of the erroneous target word attempt to the abandonment of speech prior to repair (Target-to-offset duration) and the duration from the start of the erroneous target word attempt to the onset of repair (Target-to-repair duration). Repairs with a low target-to-offset duration tend to have a low offset-to-repair duration too (Nooteboom 2010), so that Target-to-repair duration might show greater differentiation between ‘early’ and ‘late’ repairs. All other things being equal, the higher these measures, the later the repair. We also included a binary classification of whether the repair is preceded by an editing term — uh or of ‘or’ in our data set — or not (Editing term).

Second, we took a proportional measure of target word completeness (Proportional completeness). This is appropriate since our target words are not independently controlled for word length or speaking rate; as a result, raw duration measures can only partially capture repair timing. We divided the number of segments in the erroneous target word attempt by the number of segments in the correct realization. We ignored segment deletions: the crucial question was which target word segment was reached in the first attempt. The measure is bounded by 1, which corresponds to the level ‘complete’ of our binary variable Completeness. All other things being equal, the higher the value, the later the repair.

2.5. Lexical frequency and control variables

We took two measures of the frequency of the target word in our quantitative analysis: its word form frequency (Word frequency) and its lemma frequency (Lemma frequency) as represented in the CELEX lexical database.

In modeling our prosodic parameters, we considered several other potentially relevant variables. First, we included a classification of each repair as involving consonantal error, vowel error or both (Error type), on the expectation that if prosodic marking is attested in phonological error repair, the correction of a vowel error is more likely to be marked than that of a consonantal error. Second, we included several measures of similarity between the vowels of the erroneous target word attempt and correction, to control for any effects of the relative number or nature of the vowels compared through our prosodic delta values. None of these yielded significant effects, so we leave them aside here. Similarly, we included several speaker-related factors, which yielded no significant effects.

3. Results

3.1. Occurrence of prosodic marking

In our final coding for Prosodic marking, 63 instances (19%) are classified as marked, 49 (15%) as possibly marked and 213 (66%) as unmarked. These proportions are very similar to those we have reported for lexical repairs (20%, 11% and 69% respectively; Plug & Carter 2013). Therefore, our findings provide no support for Cutler’s assertion that while lexical repairs are regularly prosodically marked, phonological error repairs are generally unmarked.

3.2. Acoustic measures and marking judgments

The pitch and intensity characteristics of our phonological repairs are very similar to those reported in Plug & Carter...
We focus here on delta values derived from the first stressed vowels in the erroneous target word attempt and its correction, as these allow for the most direct comparison with Nooteboom’s (2010) data. Figure 1 shows corresponding f0 maximum and intensity maximum delta measures plotted against each other, with marked, possibly marked and unmarked instances labeled ‘m’, ‘p’ and ‘u’ respectively. (Equivalent plots for median and mean delta values show similar patterns.) If f0 and intensity are manipulated independently in the prosody of self-repair, we would expect data points to fall into discrete clouds. Moreover, if our acoustic parameters are among those on which the perception of prosodic marking is based, we would expect data points representing marked, possibly marked and unmarked instances to cover distinct subareas of the plot — following Cutler (1983), marked instances should cluster around the periphery of the plot, where data points represent instances with a large delta value for one or both parameters.

Figure 1 shows that most instances have delta values around 0 for both f0 and intensity. Moreover, the scatter shows what looks like a single cloud of data points with a positive correlation between the two dimensions, and only small numbers of instances around the peripheries of the plots (p=0.3817, p<0.0001). Most data points corresponding to prosodically marked instances occupy the top right quarter of the plot; these instances have a rise in f0 and intensity between the reparandum and repair. Instances with a fall in f0 and intensity do occur in our dataset; however, few of them are perceived as marked. On the whole, the distributions suggest that the higher the increase in f0 and intensity maximum between a reparandum and repair item, the greater the likelihood that the repair is perceived as prosodically marked. This is in line with Cutler’s (1983: 80–81) observation that ‘typically’, a marked repair ‘is uttered on a higher pitch and with greater intensity than the erroneous material’.

These observations are confirmed by further analysis. Modeling the marking judgments on the basis of our f0 and intensity maximum delta measures using conditional inference regression trees (Tagliamonte & Baayen 2012) reveals three homogeneous subsets of data: 185 instances with intensity delta values up to 2.76 and f0 delta values up to 0.05, of which less than 20% are perceived as possibly marked or marked; 73 instances with intensity delta values up to 2.76 and f0 delta values above 0.05, of which about 50% are perceived as such; and 67 instances with an intensity delta above 2.761, of which about 50% are perceived as marked.

3.3. Modeling the prosodic parameters

In order to establish the predictive value of the factors described above, we modeled each of our acoustic parameters and our marking judgments, again using conditional inference regression trees. The modeling revealed consistent effects across dependent variables; we focus here on f0 maximum delta measured across the first stressed vowels in the erroneous target word attempt and its correction. In modeling this parameter, we included the corresponding intensity measure as a control variable, so that candidate predictor effects that are observed only in subareas of Figure 1 may emerge in the form of interactions between those predictors and the control prosodic variable.

Figure 3 shows the resulting regression tree. We see that the data is first split on Stressed vowel intensity maximum delta, such that instances with a high intensity maximum delta also have a high f0 maximum delta. This is in line with the positive correlation visible in Figure 1. More noteworthy is that in the subset of instances with a relatively low intensity maximum delta — and therefore a relatively low f0 maximum delta, at or below 0 on average — two further splits are possible, on Completeness and Target-to-repair duration. First, the 101 instances within this subset with an incomplete erroneous target word attempt have a higher f0 maximum delta (0 on average) than the remaining 37 instances with a completed target word attempt (below 0 on average). These 37 instances can be further split, such that instances with a higher target-to-repair duration have a lower (on average negative) f0 maximum delta.

Figure 3 illustrates that there is a systematic relationship between repair timing and repair prosody in our data. The relationship is not strong: instances with f0 deltas above 0 are mostly unconstrained by repair timing. Still, the direction of the relationship is consistent with the pattern of prosodic differentiation of early and late repairs described by Nooteboom (2010): early repairs are associated with higher delta values than late ones. None of our frequency-related...
variables or control variables feature in the analysis, including whether the mispronunciation concerns a vowel or consonant. However, it is worth noting that direct modeling of our timing variables (not shown in detail here) does reveal that higher frequency target words are more likely to be completed prior to repair than lower frequency target words. This provides some support for the notion that higher-frequency lexical items form more cohesive units in speech production (Kapatsinski 2010), although this interaction between timing and frequency variables does not appear to have a significant effect on repair prosody.

Figure 3: Conditional inference regression tree modeling the difference in f0 maximum between the first stressed vowels in the erroneous target word attempt and its correct production.

4. Discussion and conclusion

As indicated above, our findings provide no support for Cutler’s (1983) assertion that while lexical repairs are regularly prosodically marked, phonological error repairs are, as a rule, unmarked. Comparing the findings with those of our previous study of lexical repairs (Plug & Carter 2013), we can only conclude that the prosody of phonological repairs is not as different from that of lexical repairs as Cutler suggests. This arguably has implications for our understanding of the function of prosodic marking in self-repair. In Levelt’s (1989) model of speech production, a speaker’s decision to mark or not to mark a repair prosodically is a semantic one, made at the message level. While phonological repairs do not, strictly speaking, involve semantic contrast at the message level, it seems that contrastive prominence can be assigned to the repair. Repair prosody is perhaps not exclusively ‘semantically motivated’—or perhaps an incorrect and correct attempt at the same target word should be considered semantically distinct, at least in the context of self-repair.

It is worth highlighting that we found no prosodic differences between repairs of vowel and consonant mispronunciations. If we assume that segments can function as prosodic domains, attracting narrow focus (Van Heuven 1994), it would seem plausible that in phonological repair, contrastive prominence is assigned at the segment level, while in lexical repair, it is assigned to the entire repair item. If so, corrections of vowel errors should be more likely than corrections of consonant errors to be associated with pitch marking, simply because many consonants cannot be marked through pitch. The absence of this pattern in our data suggests that there is little qualitative difference between prosodic marking in lexical repair and prosodic marking in phonological repair: both are equally frequent, and both appear to be implemented through similar speech production processes, across similar domains.

Our findings are consistent with Cutler’s (1983: 80–81) assertion that ‘typically’, a marked repair ‘is uttered on a higher pitch and with greater intensity than the erroneous material’, and suggest that the independence of pitch and intensity parameters implied by Levelt & Cutler’s (1983: 206) definition of prosodic marking — ‘a noticeable increase or decrease in pitch, in amplitude, or in relative duration’ — should not be overestimated. Nakatani & Hirschberg (1994) have reported similar results.

Our findings are also consistent with those of Nooteeboom (2010). We predicted that early repairs are associated with higher delta values for f0 and intensity than late ones: early repair should be associated with a mean rise in pitch and intensity between reparandum and repair, while late ones should be associated with a mean fall. We found some evidence that this is the case in our data. When modeling the three f0 delta measures between the first stressed vowels in the erroneous target word attempt and its correct production, several of our measures of repair timing yielded significant effects. The effects are all in the predicted direction: for example, incomplete erroneous target word attempt have a higher f0 maximum delta than complete ones in one subset of the data, and instances with a relatively high target-to-repair duration have a relatively low f0 maximum delta in another. The effects are weak, but similar to those we found in our study of lexical repair (Plug & Carter 2013). Therefore, we can conclude that it is at best a simplification to suggest that repair prosody is ‘orthogonal to the time course of error detection and correction’ (Cutler 1983: 81).

5. Acknowledgements

This work was supported by ESRC grant RES-061-25-0417. We thank Christina Englert for her contribution.

6. References


Abstract

Joint speaking, in which many people say the same thing at the same time, is a common vocal practice found in situations of heightened collective significance. In the wild, prosodic stylization is common. In the laboratory, we show that this stylization is not a necessary consequence of the requirement to speak in unison. Speech obtained from groups of 2, 4, 6 and 8 speakers remains relatively unaltered. But if the speech is unremarkable, the act of speaking is clearly not, and there is some behavioral and neuroscientific evidence for emergent phenomena arising in joint speaking that are not present in the speech of single individuals. Consideration of the status of joint speech, and its remarkable absence from contemporary linguistics, suggests that structuralist approaches to language that inform most of modern linguistics oversee much of that which is important about vocal communication.

Keywords: joint speech, choral speech, synchronous speech, unison

1. Introduction

Joint speech is an umbrella term that covers those forms of speaking behavior when many people say the same thing at the same time. Joint speech practices are found in every culture, and the study of joint speech production cannot overlook the fact that most situations in which joint speaking occurs appear as overt manifestations of collective intentionality, in which group purposes, group sentiments and group intentions are made vocal.

Collective prayer is the most common form of joint speaking, found in all major religions. Prayers are often short, and repeated many times over, often with the aid of prayer beads. They may have a call and response form, and prayers that are frequently repeated often exhibit highly stylized prosodic forms in which the segmental details are distorted or even radically altered. Fig. 1 shows excerpts from a recording of the recitation of the Catholic Rosary. Four successive iterations are shown of the underlying phrase “Blessed is the fruit of thy womb, Jesus” which is, for this speaker, reliably and invariantly produced as (approximately) /bles.froU.ThaIm.˚˚˚. Despite the substantial alteration to the segmental composition of the phrase, it can be seen that the four iterations are produced with considerable stability and consistency.

Prayer and protest make odd bedfellows, but in the chants of protesters we find many similar features: the expression of a collective intentionality through the production in unison of short and frequently repeated phrases. Along with repetition, many protest chants are structured as call and response, not unlike the structure, for example of the Catholic Hail Mary prayer. In protest too we again find the emergence of stereotypical prosodic forms. Perhaps the best known is the sequence of accents which we might characterize as in Fig. 2. This is the accentual structure of the famous chant “El pueblo, unido, jamás será vencido” (“The people, united, will never be defeated!”), which became globally known after the CIA-led coup that overthrew the government of Salvador Allende in Chile in 1973. What had been a chant associated with his election campaign became a global template for protest. Today the same basic pattern is found in many cultures and languages. A cursory survey of amateur videos from protests throws up examples in English, Greek, Portuguese and Arabic, at least. It is also the basic template for the widespread chant that has come to represent the uprisings across North Africa and the Middle East in the so-called “Arab Spring”, where the call of “Ash-sha’b yurid isqāt an-nizām”, or “The people demand the fall of the regime” has been used and adapted in Lybia, Tunisia, Egypt, Sudan, Syria and beyond.

Figure 1: Four successive iterations of the phrase “Blessed is the fruit of thy womb, Jesus”.

Figure 2: Common accent pattern found in protest chants in many languages.

The basic elements of repetition, prosodic stylization, (sometimes) call and response, and the vocal demonstration of...
a collective identity are also found in the chants of sports fans. Beyond these domains, joint speech is used in educational settings for diverse purposes such as rote learning, pronunciation training, and performance. In the latter case, the term “choral speaking” is conventionally used.

Despite its ubiquity, and despite the embedding of joint speech practices into situations of heightened collective significance, there has been little or no scientific treatment of joint speech practices to date. A laboratory variant of joint speaking, dubbed Synchronous Speech, has been studied in some detail (Cummins, 2003; Cummins, 2009). In this paradigm, pairs of subjects are presented with unseen texts to be read in synchrony with one another on a go-signal from the experimenter. This task is surprisingly easy for subjects, and the temporal alignment attained has been estimated to be characterized by a mean asynchrony of approximately 40 ms (Cummins, 2003). Anecdotal observations of synchrony in collective prayer suggest that synchrony in a synchronous speech task is much tighter (less asynchrony) than that commonly found in the wild. Unlike the ritual repetition of prayer or protest chant, utterances produced in a synchronous speech experiment are not prosodically marked, and indeed speaking in synchrony has become an experimental constraint that can be used to reduce inter-subject variability in phonetic experiments (Cummins and Roy, 2001; Cummins, 2004; Krivokapi´c, 2007; Kim and Nam, 2008).

In a recent study, some prosodic alteration, in the form of an exaggerated syllable-timing, was observed when Mandarin speakers read in synchrony (Cummins et al., 2013). In follow up work, we have been unable to reliably replicate this effect. We now suspect that the apparent change was a perceptual effect that is properly attributable to the known slower speech rate of synchronous speech, rather than any substantial prosodic reorganization. And so we now confront the observation that both prayer and protest chants reliably exhibit highly stylized prosodic forms, while synchronous speech in the laboratory apparently does not. We therefore conducted a small experiment to see whether collective speaking in groups larger than the dyad resulted in substantial prosodic reorganization. Representative results will be provided here.

2. Synchronous speech is unremarkable

2.1. Methods

Eight native English speakers took part in a single recording session. Five texts with very different accentual structures were employed, including a list of 8 trochees, a poem with some lines in a duple and some in a triple meter, and the second part of the Hail Mary prayer. Subjects stood in a single room with approximately one meter distance between neighbouring individuals. On the instruction of the experimenter, all texts were read in varying combinations of 1, 2, 4, 6 and 8 speakers at a time, with group membership chosen randomly. Each subject wore a head-mounted microphone to minimize cross-channel bleed.

In order to examine the coarse temporal structure of utterances, we look at the intervals between prominent onsets. For each utterance, the onset of prominent syllables was calculated using the algorithm introduced in Cummins & Port (1998), which provides a working estimate of the perceived onset, or P-center of the syllable (Morton et al., 1976; Scott, 1993). Synchrony among more than two speakers has never been quantitatively examined, and so a quantitative estimate of pairwise asynchrony was also computed for each speaking pair, by estimating the degree of temporal warping necessary to map one utterance onto the other, using the computational method introduced in Cummins (2009). For dyadic readings this generates a single score, while for readings with 8 speakers, this produces 28 dyadic scores. In each case, a high score indicates lack of synchrony, and a score of zero would imply perfect synchrony.

2.2. Results

In what follows, we have very different numbers of observations for the different conditions of \(n=1, 2, 4, 6, \text{ and } 8\). We therefore eschew analysis by ANOVA and look instead for obvious qualitative features that might index the number of speakers. Our principal target is any evidence of wholesale changes to macroscopic temporal structure as a direct result of the synchronization task. Beyond that, we can ask whether adding more speakers alters pairwise synchrony in any obvious fashion.

![Figure 3: Inter-onset intervals for lists of 8 trochees.](image-url)

We first consider timing measurements based on the list of 8 trochees ("Borrow, Dancer, Butter, Dagger, Boiler, Doggie, Body, Deeper"). Fig. 3 plots the median interval duration observed for the seven intervals defined by the eight word onsets, with separate plots for different numbers of speakers. No time normalization has been applied. Because only a single set of words was employed, there is little to be deduced from the slight pattern of alternation from one interval to the next, which is heavily influenced by the contingent segmental content of these specific words. For \(n = 1\), the final interval is shorter than for all other values of \(n\). But there is no visible effect whatsoever as one goes from 2 to 4, 6, and then 8 speakers. This is evidence that no substantial reorganization of macroscopic interval timing arises as a function of the number of speakers for this series of maximally regular words.

Turning now to a text with considerably more complex metrical structure, we examine the intervals between stressed syllable onsets in the short poem, the text of which reads (with apologies to cat lovers): *Kill a cat, kill a cat, Bash its brains in with a bat. Its nine lives expire, When tossed in a fire, So kill a cat today.* Fig. 4 shows the distribution of interval durations between stressed syllable onsets (underlined) for one and eight speakers. The shift from duple to triple meter between the seventh and eighth intervals is very obvious, and the metrical regularity is augmented by a considerable degree of lexical and segmental variability. However, when we examine the succession of median interval durations, there is no apparent qualitative difference observable as we move from one to 8 speakers. Examination of \(n=2, 4 \text{ and } 6\) confirms the absence of any sub-
Figure 4: Inter-onset intervals for the Kill a Cat poem. Onsets correspond to underlined letters in the text of the poem.

The substantial effect of group synchronization on macroscopic interval durations.

Figure 5: Inter-onset intervals for the second half of the Hail Mary prayer. Onsets correspond to underlined letters in the text.

Finally, we consider a text with less pronounced metrical structure, but whose inclusion is warranted as it is frequently recited collectively (Fig. 5). We use the second half of the Hail Mary prayer: Holy Mary, Mother of God, Pray for us Sinners, Now, and at the Hour of our Death, Amen. As with the trochees, we plot median interval durations as a function of the number of speakers. In this case, there is a clear rate effect, as recitations with 4, 6 and 8 speakers are all slower than those with one or two speakers. With the exception of the sixth interval, the pattern of successive durations appears relatively invariant. The sixth interval is that between Sinners, and Now, which includes a major syntactic break. Previous work has established that speaking in synchrony greatly reduces inter-subject variability in pause placement (Cummins and Roy, 2001), though no work has identified either lengthening or shortening as an overall effect of synchronization.

We estimated the asynchrony among all possible pairs of speakers. Asynchrony estimates are in units of area under a time warping curve that maps one utterance of a pair onto the other, as reported in Cummins (2009). Estimates have been log transformed, which produces more nearly normal distributions, and have then been converted to standard scores. Fig. 6 shows the distribution of asynchrony scores as a function of the number of speakers for the Hail Mary text. Similar results were obtained for other texts. It is clearly not the case that adding more speakers leads to greater dyadic synchrony, nor asynchrony. There is thus no manifest effect of the number of speakers on either the macroscopic temporal structure of an utterance, nor on the synchrony obtaining among subjects, with the one exception found that a pause at a major syntactic break led to convergence upon a relatively long value for \( n > 1 \) speakers. In line with previous results, it seems then that synchronous speech is relatively unmarked and unremarkable, even as we increase the number of speakers from 2 to 8. The exaggerated prosodic stylization that we regularly find in collective prayer and protest is thus not due to the demands of synchronization among speakers.

3. But synchronous speaking is special

If synchronous speech is unremarkable, the same cannot be said for synchronous speaking. We have established two sources of evidence that suggest that the act of speaking together needs to be understood as a collective act, in which the speakers become mutually entangled, or coupled, in a manner analogous to the way in which two runners in a three-legged race are physically coupled.

Figure 6: Distribution of pairwise standardized long asynchrony scores for different number of speakers.

Figure 7: Waveforms illustrating synchrony interrupted by abrupt and simultaneous cessation of speech in mid-syllable. The arrow indicates the point of cessation.

The first source of evidence arises in the observation that there is a distinguished class of speech error, common when two people speak in synchrony, and unknown otherwise. This typi-
...ocially arises when one speaker either makes an error, or displays some degree of uncertainty. What is then observed is that the phonetic and phonological properties of joint speech are intact, but that something else is going on at the collective level that is unique to the realtime reciprocal linkages between joint speakers. An understanding of what this behavior is, why it occurs, and how it acquires such significance requires us to adopt a rather different perspective on speech and language.

The scientific study of language that has arisen since the structuralist approach of de Saussure has emphasized some aspects of language, notably the combinatorics of finite, discrete elements in symbolic structures, that must then be decoded by a listener. In this view, the roles of speaker and listener are utterly distinct, and much of the behavior we are familiar with that attends speaking is simply not addressed. This is a thoroughly conventional view of what “language” is, and of the kind of message-passing activity that it facilitates. More recent extrapolations of this approach have attempted to narrow, rather than enlarge, the set of “linguistic” phenomena proper, so that on one influential view, “language” is to be viewed as a modular faculty whose defining (perhaps only?) property is the support of recursion (Hauser et al., 2002). This extremely narrow focus will not serve to understand joint speech, nor will it serve to understand most human languaging behaviors.

The multifarious ways in which speaker/listeners become coupled during a conversational exchange, with the rich intertwining of facial and manual gestures, with backchannels that support, encourage and nudge the flow of speech, with the careful regulation of gaze, all these, because specific to face-to-face vocal communication, are excluded from the science of language so construed (Richardson et al., 2007; Wagner et al., 2014). Within the descendants of the structuralist tradition, even the sound itself is to be partitioned into those elements that are found to support the demarcation of discrete sound categories (phonemes) and everything else which is consigned en masse to the miscellaneous drawer of “prosody”. Prosodists, then, spend most of their time vainly attempting to demonstrate how non-segmental aspects of speech sounds can be retro-fitted into a symbolic framework that has acquired the status of an immovable authoritative object.

If we pull back our field of vision and examine what a science of language is asked to provide an account of, we find that there has been a veritable industry born of the construction and defense of theories of how language has evolved, how it gives rise to the construction of a shared world, enables the development of the whole of human culture and technology, how it facilitates complex cognitive processes, and more (Deacon, 1997). All these great feats are heathenishly attributed to something called “language”, and the scientific position on what that is finds its acknowledged authority in the fields of syntax and semantics. Remarkably, theories of syntax and morphology, along with the whole of formal semantics, can all be constructed, tested, and established without distinguishing in any meaningful way between written and (transcribed) spoken utterances at all. From some perspectives, it appears as if the object of academic linguistics might be better viewed as the code underlying writing.

If we recognize this, we might begin to reassert and recognize some of the power of the voice (Connor, 2000). Writing is much younger than language. The introduction of writing introduced wholesale changes to how we think, how we communicate, how we argue, reason, and situate ourselves in a shared
world (Olson, 1996; Ong, 1982). It is an extension of the vocal tradition that preceded it by dozens of millennia, and the regularities of the written code are elaborations of, and transformations of, the regularities that are to be found in vocal utterances. To dismiss the productions of the voice as mere performance, and insist that they derive from an underlying formal system of rule based symbolic manipulation, is to deny the power of the voice, to ignore its position as the principal form of linguistic behavior, or languaging, which gave rise to what we recognize as modern humanity. It is to miss the very business of languaging.

For writing is not the only descendent of the much older phenomenon of vocal behavior. Many of the attributes of joint speech have likewise become codified, elaborated and transformed, but this has typically happened in the development of liturgy and ritual. The improvised and repeated gestures of spontaneous chant become encoded in practices of kneeling, solemn walking, head bowing, bead twirling, marching, and more. Hidden away in rituals consigned to the sphere of contingent cultural practice, such codifications have been overlooked by the sciences of language. Here, then, is the origin of the phonetic distortions, the prosodic stylizations, and the recurring patterns that arise in practices of prayer and protest. The characteristics of joint speech are invisible if we accept the received view of language, but if we can recognize the remarkable breadth of language behaviors, and the efficacy and power of the voice in structuring our collective practices, we can see so much more.

5. References


Vowel coarticulation and undershoot in prosodically weakened positions

Conceição Cunha, Jonathan Harrington, Ulrich Reubold1

1Institute of Phonetics and Speech processing, LMU Munich, Germany

cunha|jmh|reubold@phonetik.uni-muenchen.de

Abstract

This paper focuses on the influence of vocalic context and prosodic weakening on the production and perception of the German fricatives /s/ and /ʃ/. Many studies have shown coarticulatory influences of vocal quality on the preceding fricative and perceptual compensation for such coarticulatory relationships (Nittrouer & Studdert-Kennedy, 1987; Mann & Repp, 1980). However, very little is known about the perceptual parsing of coarticulated speech in prosodically weakened positions. The working hypothesis is that the degree of coarticulation may be similar in accented and deaccented positions, that the variability associated with undershoot is greater in deaccented words, causing listener errors in attributing coarticulation to its source. In order to explore this issue, we analyse the coarticulatory influences of German /i/, /a/ on /s/, /ʃ/. The present investigation involves physiological analysis of tongue position and movement. It addresses how this is related to perceptual judgments, and whether this relationship changes between accented and deaccented words.

Keywords: coarticulation, prosodic weakening, undershoot, German, prosody, perception

1. Introduction

The main aim of this study was to determine the extent to which vowel coarticulation and prosodic weakening affects the production and perception of /s/ and /ʃ/ in German. More specifically, the study investigates the articulatory implementation of coarticulation for /i/, /a/ on /s/, /ʃ/ in accented and deaccented positions and its parsing in perceptual judgements, with a special focus on perceptual compensation for prosodically weakened tokens.

The coarticulatory influence of vowel context on the preceding fricative has been extensively investigated in adult and child speech over the past decades (Recasens & Espinosa, 2007; Katz et al., 1991; Mann & Soli, 1991; Nittrouer et al., 1989; Nittrouer & Studdert-Kennedy, 1987; Repp, 1986; Mann & Repp, 1980). The raised tongue position in high vowels is assumed to facilitate the assimilation of the front-back location of the fricative’s constriction, resulting in a more frontal constriction for fricatives in /i/ than in /a/ contexts (Soli, 1981). This anticipatory lingual coarticulation is acoustically detectable by a shift of F2 loci. Lip rounding for /u/ has also been associated with global shifts in the spectrum during the fricative (Nittrouer et al., 1989; Repp, 1986; Soli, 1981). The realization of lip rounding in post-alveolar fricatives – considered to be an enhancement strategy (Stevens and Keyser, 2010) – has been shown to be subject to great inter-speaker variability, since some speakers realise /s/ and /ʃ/ with the same lip configuration and absolutely no rounding even in /ʃ/ context (Proctor et al. 2006, Fig. 6).

Many studies have shown perceptual compensation for anticipatory coarticulation. For example, the participants in Mann & Repp (1980) reported more /s/-responses before /u/ than in the unrounded context of /i/, meaning that listeners factored out the anticipatory spectrum-lowering of the lip rounding for /u/ on the preceding fricative (compensation).

The second aim of this study was to assess the influence of prosodic weakening on the degree of coarticulation of initial fricative vowel sequences and its influence on speech perception. More coarticulation has been attested in unaccented than in accented positions (Cho, 2004; Lindblom et al., 2009). In an acoustic and perceptual analysis of VCV-coarticulation in German, Harrington et al. (2013) suggested that the magnitude of coarticulation may be similar in both prosodic positions, but that the degree of variability resulting from target undershoot may be much greater in deaccented position. However, very little is known about the parsing of coarticulation under prosodic weakening in speech perception. Listeners may have more difficulty parsing coarticulatory relationships in prosodically weak constituents (Harrington et al. 2013), not due to greater coarticulation but rather due to increased variability, which might mask coarticulation causing listener errors in attributing the coarticulation to the source that gives rise to it.

In order to explore this issue, we analyse the influence of German /i/, /a/ on /s/, /ʃ/ fricatives, building on earlier studies investigating similar materials (Fowler, 2006; Mann & Repp, 1980; Mann & Soli, 1991; Nittrouer & Studdert-Kennedy, 1987; Whalen, 1981). Assuming lip rounding in the production of the post-alveolar fricative (at least in some speakers) and no lip rounding in the alveolar counterpart, we predict the greatest amount of lip rounding and/or backmost tongue constriction for /ʃ/ in the /u/ context, since both segments are produced with rounding of the lips. Conversely, no lip rounding and the most frontal tongue position are expected in the production of the alveolar fricative with the front vowel (/s/). In perception, we expect to replicate Mann & Repp’s (1980) results: the greater the vowel’s influence, the more /s/-responses in an /u/-context and conversely more /ʃ/ in the context of /i/. Following Harrington et al. (2013) we expect a similar degree of coarticulation in both prosodic positions, but undershoot and increased variability and less perceptual normalisation for context in deaccented positions.

2. Production Experiment

2.1. Methods

2.1.1. Experimental set-up

Physiological EMA data were recorded from eight speakers of southern German (4 male, 4 female) using the 3D articulograph CARSTENS AG501. Two sensors were placed on the tongue: one on the midline 1 cm behind the tongue tip (TT) and the other on a level with the molar teeth at the tongue back (TB). Two sensors were placed on the upper and lower lip (the latter henceforth LL). Four additional sensors were fixed to the maxilla, the nose bridge, as well as to the left and right mastoid bones: these served as reference sensors to correct for head movement.

The speech material consisted of initial fricative-vowel
sequences followed by /ɪ, ʊ/ in the 4 German lexical words Suppen ‘soups’, Schuppen ‘dandruff, hovel’, Sippen ‘clans’, Schuppen ‘ scoops’, supplemented with 14 dislocator words. The target words were embedded in phrase-final position in the carrier sentence Maria mag [target word] (eng, ‘Maria likes [target word]’). Two of the target words contain voiced fricatives which usually become devoiced in Southern German when following a voiceless/devoiced context (as in [məiːma:kʊɪp]). In order to elicit either accented or deaccented position by shifting the focus between the initial and the target word in the carrier phrase, the participants were presented with questions designed to elicit a narrow focus on the target word for the accented context and a broad focus for the deaccented context: either WAS mag Maria? (“WHAT does Maria like?”) or WER mag [target word]? (“WHO likes [target word]?”). Thereafter, the stimulus was presented with the word carrying the nuclear accent in capital letters (e.g. MARIA mag Schuppen vs. Maria mag SCHUPPEN). If subjects made a mistake, they were instructed to repeat the sentence. In total each speaker produced 80 utterances containing one of the target words (2 accentuation conditions x 4 target words (<2 fricatives x 2 vowels) x 10 repetitions).

2.1.2. Data analysis

The acoustic data were digitized at 16 kHz and automatically segmented and labeled using the Munich Automatic Segmentation tool (MAuS, Schiel 2011). The segment boundaries of the target words’ fricatives and the following vowels were manually corrected.

Post-processing of the physiological raw data was done semi-automatically in Matlab, whereas labeling and subsequent analyses of the physiological data were conducted using EMU and EMU/R (Harrington, 2010). The physiological annotation of the three sibilants was based on the vertical movement of the TT (in mm) and the TT tangential velocity (in mm/s). Our articulatory analyses were all based on the same time-frame which was derived from the gesture trajectories of the vertical movement of TT measured between the velocity peak of the fricative closing gesture (v0) and the acoustical vowel onset. We extracted in this time-frame the horizontal movement of TT and the horizontal movement of LL. In order to quantify the articulatory trajectories of the horizontal TT and LL movement, we used discrete cosine transform (DCT) to reduce the trajectories and the spectral slices to a set of coefficients. The nth DCT-coefficient Cn (m = 0, 1, 2) was calculated with the formula in (1):

\[ C_n = \frac{2k}{N} \sum_{n=0}^{N-1} x(n) \cos \left( \frac{(2n+1)m\pi}{2N} \right) \]  

These three coefficients Cn (m = 0, 1, 2) encode the mean, the slope, and curvature respectively of the signal to which the DCT transformation was applied (Harrington, 2010).

We used a relative measure, the log. Euclidean distance ratio, to quantify relative positions of tokens in relation to anchors. These anchors varied as follows: we averaged the corresponding parameters over all Schuppen and Sippen tokens per speaker and accentuation condition, as we expected the distance between these fricatives to be maximally distributed. To calculate a measure for both undershoot and coarticulation differences, we used only those tokens from the accented condition as anchors, to factor out undershoot we applied the same methodology, but using for each accentuation condition separate anchors, i.e. accented Schuppen and Sippen token for the accented, unaccented Schuppen and Sippen tokens for the unaccented condition. By doing so, only the effects of accentuation differences on the amount of coarticulation remain.

The Euclidean distances ESippen and ESchuppen were calculated in a space build up by the three DCT coefficients separately for each fricative token. The log-Euclidean distance ratio d was then calculated for each fricative, from (2):

\[ d = \log(E_{Schuppen}) \]  

The log-Euclidean distance ratio d was calculated in order to obtain one value per frricative which is a relative measure: negative values denote a closer distance to the Schuppen centroid, whereas positive values are associated with distances nearer to the Sippen centroid, while a value of zero denotes that a given fricative is equidistant between the centroids.

We applied repeated measures ANOVAs with d as dependent variable, and with the within-subject factors ACCENTUATION (accented vs. unaccented), WORD (so, si, fo, and fi).

2.2. Results

The effect of accentuation on the fundamental frequency of the target vowels was verified by conducting a repeated measures ANOVA with F0 as dependent variable and accentuation (accented vs. deaccented) and vowel (/i, /o/) as within-speaker factors. The results showed a significant effect only for accentuation (F(1,7) = 6.8, p< 0.05).

2.2.1. Horizontal tongue tip

For each speaker the average of the horizontal movement of the TT sensor, measured between the velocity peak of the closing gesture of the fricative and the vowel offset was calculated. Fig.1 show a clear degree of separation of the trajectories for /s/ and /ʃ/ with the alveolar trajectories of /s/ and /ʃ/ presenting a lower value for anteriority (i.e. a more fronted position) relative to the post-alveolar counterparts. However, the effect of vowel context is visible in the second half of the fricative in the trajectories of some speakers, in which the trajectory of /s/ won on posteriority and the /ʃ/ trajectory instead became more anterior.

As evident in Fig. 2a, Log. Euclidean distance ratios are greater in /s/ compared to /ʃ/, and in /ʃ/ compared to /s/, confirming the impression given by the trajectories in Fig 1. The RM-ANOVA with Log. Euclidean distance ratios as dependent variable showed a significant effect of word
(F(3,21) = 30.4, p < 0.001) and a small effect of accentuation (F(1,7) = 6.0, p < 0.05), in which deaccented tokens are closer to zero than accented ones. Recall that this measure quantifies differences that come about mainly because of undershoot. Deaccented tokens show smaller distances than accented ones.

However, on the relative distances presented in Fig. 2b, we still find a main effect of word (F(1,7) = 122.5, p < 0.001), but no influence of accentuation. We interpreted this as being a consequence of the similar degree of coarticulation in both prosodic conditions (after the undershoot was factored out).

### 2.2.2. Lower lip

Fig. 3 displays the mean trajectories per speaker of the horizontal movement of the lower lip for the same time interval as in Fig 1.

On the lower lip trajectories the degree of separation between fricatives is less evident. However, /ʃ/ showed as expected the more fronted position, which corresponds to the greatest amount of lip rounding. In /s/ the lower lip showed the backmost position and /ʃ/ displayed intermediate positions. The same trends could be confirmed in Fig. 4a, in which /s/ displayed the greatest positive d-values and /ʃ/ displayed the most extreme negative values. Moreover, as expected, the mean distances of /s/ and /ʃ/ present an intermediate value close to zero. This is the reason for the main effect of word on the d-values (F(3,21) = 29.8, p < 0.001).

Regarding accentuation, the trend for smaller distances in deaccented position visible in Fig. 4 was not statistically significant. An ANOVA (on which factors?) showed no effect of accentuation nor a significant interaction. For the relative distances shown in Fig. 4b we did not find any influence of word nor accentuation, meaning that the degree of coarticulation of the lower lip is also quite similar in both prosodic conditions.

### 3. Perception Experiment

#### 3.1. Method

For the perception experiment, we synthesized a 10-step continuum between /s/ and /ʃ/ by filtering white noise and prepended the resulting sounds to /nas/ in the same words which we recorded in the production study. The target words were spoken in a carrier sentence (Maria mag [target word]) by a trained male phonetician with slight Southern German characteristics. The vowel transitions of the resulting stimuli were either appropriate to /s/ or /ʃ/ (see also Nittrouer & Studdert-Kennedy, 1987 for English). We manipulated the accentuation pattern by means of PSOLA in Praat by shifting the nuclear accent between Maria and the target word. Therefore, four different continua in two accentuation conditions could be tested:

- /s/-ʃ continuum followed by /s/-transition and /a/: ʃi.
- /s/-ʃ continuum followed by /s/-transition and /o/: so.
- /s/-ʃ continuum followed by /ʃ/-transition and /a/: sɨ.
- /s/-ʃ continuum followed by /ʃ/-transition and /o/: ʃo.

These materials (10 step continuum x 2 vowels x 2 transitions x 2 prosodic conditions x 6 repetitions = 480 stimuli) were presented to 19 students of southern standard German aged between 21 and 29 years. Eight of them had taken part in the production experiment.

We ran a two alternative forced choice experiment in which participants were asked to choose between /s/ and /ʃ/ by clicking on buttons labelled with the corresponding German orthography for the sounds (either <S> or <Sch>).

For the statistics we ran repeated measures ANOVAs on the category boundary and slope with TRANSITION (s1, s2, f1, f0) and ACCENTUATION (accented vs. deaccented) as within-listener factors and, in case of an interaction, post hoc t-tests on each combination of vowel and accentuation.

#### 3.3.3 Results

Fig. 5 shows mean psychometric functions fitted to the distributions of listeners’ responses to the eight continua. The vertical lines correspond to the 50% cross over points.
The psychometric curves presented in Fig. 6 and the corresponding 50% cross over boundaries showed a clear main effect of vowel on the category boundaries (F(1,18)= 6.9, p< 0.001): that is, listeners perceived more /s/ in the /a/ context and more /ʃ/ in the /i/ context, so that the vowels had the major influence on the category boundary. Accentuation had no effect on the cross over point nor on the slope.

4. Discussion and conclusion

This study analysed the coarticulatory influence of /ɔ, u/ on /ʃ/, /s/- fricatives in two prosodic conditions (accented vs. deaccented). The analysis of the horizontal tongue tip (TT) and lower lip (LL) movement could confirm anticipatory lingual and labial coarticulation (Soli, 1981; Katz et al. 1991). As Fig. 3 suggests, there seems to be a considerable amount of inter-speaker variation in lip-rounding in the post-alveolar /ʃ/, irrespective of vowel context as in Proctor et al. (2006). As can be seen in a comparison of Figs 2 and 4, the coarticulatory effect on the LL gesture was much more pronounced than the effect of anticipatory lingual coarticulation, which is reflected by the fact that the post-alveolar preceding unrounded /i/ and the alveolar preceding rounding /a/ share comparable amounts of lip rounding.

Our results concerning the influence of accentuation on the degree of coarticulation were consistent with Harrington et al. (2013) who had found a similar degree of coarticulation in both accented and deaccented positions: Our present results from the production data suggest that there is undershoot in deaccented position, but little change on the degree of vowel-on-fricative coarticulation in speech production across the prosodic conditions. However, the perceptual category boundaries were not flatter in deaccented position as expected, but listeners were found to compensate for coarticulation to about the same degree across the two prosodic contexts. These results suggest that listeners are highly attuned not only to coarticulatory variation, but also to the degree to which such variation is affected by prosodic context.

5. Acknowledgements

We would like to thank all participants of the experiments. This research was supported by the German Research Council grant number HA3512/9-1, awarded to the second author.

6. References


A Study of Nasal Coupling for Non-Nasalized Sounds
Jianwu Dang¹,², Chen Zhao¹, Jianguo Wei¹, Kiyoshi Honda¹
¹ Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University
92 Weijin Road, Nankai District, Tianjin 300072, China
² School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
jdang@jaist.ac.jp, jianguo@tju.edu.cn, khonda@sannet.ne.jp

Abstract
Past studies have found that functions of the velum in speech production may not only be a binary switch with on-off states for nasal and non-nasal sounds, but also be involved in speech production of non-nasalized sounds. In this study, we investigate the unique functions of the velum involved in speech production of non-nasalized sounds and estimate biomechanical properties of the velum based on MRI observations. To do so, sound radiation from the lips, sound pressures in the oral and nasal cavities, and soft tissue vibrations of speech organs were measured during producing speech sound. According to the analyses of the multimodal data, it was found that the thickness of the velum is dependent on vowels, but not on the consonants even for the voice bar of voiced stops. The velum seems to be controlled actively for generating the voice bar of voiced stops. The voice bar of the stops is likely to be caused by the nostril radiation rather than by the laryngeal wall vibration.

Keywords: speech production, voiced stops, velum vibration, transvelar coupling

1. Introduction
Nasal and nasalized sounds play important roles in speech by means of the coupling between the nasal and oral cavities via the velopharyngeal port. A number of studies have been conducted to investigate morphological and acoustical details of the nasal and paranasal cavities [1-3]. Among them, Dang, et al. used MRI to investigate relations of the morphology and acoustics of the nasal and paranasal cavities [2, 3]. Sundberg, et al. used CAT scan to investigate the effect of the velopharyngeal opening on vowel formants during singing [4]. Feng and Castelli investigated the coupling properties of the nasal and oral cavities for nasal and nasalized vowels in French [5].

The coupling degree of the nasal and oral cavities is controlled by the velum [6, 7]. In general, functions of the velum are not only a binary switch with two states of on and off during speech production. Suzuki, et al. and Dang, et al. noticed that the amplitude of acoustic radiation from the nostrils changes with non-nasalized vowels [8, 9]. Dang and Honda used multimodal measurement data to investigate functions of the velum in vowel production [10]. Their results showed the nasal cavity is involved in speech production of the non-nasalized vowels via vibrations of the velum. Dang, et al. measured intraoral pressure and sound radiations at stop consonants, and suggested that the nasal radiation may contribute to generation of the voice bar in stop consonants [11]. This observation is somewhat inconsistent with the general account that the voice bar is considered to reflect acoustic radiation caused by the vibration of the pharyngeal wall with the mouth being closed. However, there is no empirical support to this conjecture.

In speech utterances, nasal sound, nasalized sound, and non-nasalized sound are mixed together. The velum plays an important role in the transition from one kind of sounds to the others. To investigate the functions of the velum in speech, this study uses magnetic resonance imaging (MRI) to investigate biomechanical changes of the velum in the production of non-nasalized sounds. We also use multimodal observations to investigate the behaviors of the velum, which include speech sound radiations, intraoral and intranasal sound pressures, and wall vibrations of speech organs. Acoustic measurements are employed in examining the contribution of the nasal cavity to acoustic output via the transvelar coupling, while the wall vibrations are used to estimate the components of sound radiation from the yielding wall during speech.

2. MRI-Based Experiments
In general, it is considered that the transvelar coupling is concerned with physiological and mechanical changes in the velum during speech. In this section, we measure the physiological changes using MRI observation, while mechanical changes are investigated by its surface vibration in the latter sections.

2.1. Experimental setup
A cine MRI experiment was conducted to investigate the velum’s behaviors in continuous utterances with non-nasalized sounds using the Synchronized Sampling Method (SSM) [12]. The MRI system used in this study was a Shimazu-Marchon ECLIPSE 1.5T Power-Drive 250 scanner. Three male adult Japanese subjects, S1, S2, and S3, participated in this experiment. The speech material was Japanese meaningless words consisting of vowels with stop consonants, including /baba/, /bibii/, /bubu/, /bebe/, /bobu/, /gaga/, and /gigi/. The frame rate of the MRI movies was 64 fps.

2.2. Measurement of the velum’s thickness
Figure 1 shows two frames in the movies for /i/ in the utterance /bibii/ (left) and /i/ in /bobob/ (right). From the figure, one can see that the orientation and thickness of the velum are changing to a certain extent. To investigate the mechanism of the velum in speech, we measured the angle and thickness of the velum, together with the pharyngeal width of the vocal tract. The angle of the velum is related to the effort of the speaker in controlling the velum, since the velum elevation is driven by muscle activation. The pharyngeal width reflects changes in the pharyngeal cavity. The thickness of the velum determines the mass per unit area, which is concerned with the...
degree of the velum vibration. The definitions of those measures are shown in Figure 2.

As shown in the left panel of Fig. 2, the velum angle is defined as the angle between the upper surface of the velum and the extended line of the hard palate. The pharyngeal width of the vocal tract is defined as the distance between the back surface of the tongue and the pharyngeal wall at the level between the second and three cervical vertebrae. The right panel illustrates the calculation of the thickness of the velum. h1 represents the distance between the lower surface and upper surface of the velum at the junction of the hard and soft palates. h2 is the height from the zero plane to the highest portion of the upper surface, where the extended line of the lower surface of the anterior part of the velum is treated as the zero plane. The thickness of the velum is defined as the average value of h1 and h2.

In this paper, we only focus on analysis of the thickness of the velum.

The thickness of velum is calculated for the periods of the consonants and vowels. The results for vowels and consonants are shown in Figure 3 for three subjects. The thickness of the velum increases monotonically in the order of /i/, /u/, /e/, /o/, and /a/. The increase of thickness was more than 13% comparing /a/ to /i/. In consonant periods, the thickness of the velum shows about the same values as that of the following vowels. The same tendency is seen for the other consonants. It implies that the thickness is dependent on the vowels, but not on the consonants. The velum is thinner for the front vowels and thicker for the back vowels. Since the thinner velum has less mass per unit area, it would vibrate in larger amplitude for the same driving force. Accordingly, the thickness is also one of the causes for that the front vowels have larger nasal radiation than back vowels. In the case of majority, the thickness of the velum for the second consonants increases up to 3% by comparing with that of the initial one. The second vowels have the same tendency of thickness increase as that of the consonants.

As suggested in the previous studies [8], [11], velum vibration is a possible cause of oral-nasal coupling in non-nasalized vowels. To confirm this speculation, in this section, we measure the velum vibration with other multimodal channels including speech sound radiations, intraoral and intranasal sound pressures, and vibrations of the larynx wall. This experiment is aimed at measuring the nasal coupling between the nasal and oral cavities for vowels and the voice bar of stop consonants.

3. Acoustic Experiments

As suggested in the previous studies [8], [11], velum vibration is a possible cause of oral-nasal coupling in non-nasalized vowels. To confirm this speculation, in this section, we measure the velum vibration with other multimodal channels including speech sound radiations, intraoral and intranasal sound pressures, and vibrations of the larynx wall. This experiment is aimed at measuring the nasal coupling between the nasal and oral cavities for vowels and the voice bar of stop consonants.

3.1. Experimental setup

The experimental setup is shown in Figure 4, where three microphones were used to record sound pressure signals, and three accelerometers were employed to pick up wall vibration. For acoustic measurements, a B&K 4003 microphone (M1) was placed in front of the subject (15 cm apart) to record the radiated sound pressures. Two B&K 4182 probe microphones (M2 and M3) were used to measure the intraoral and intranasal sound pressures via two identical flexible tubes of 30-cm length. The probe tubes had a 0.165-cm outer diameter and a 0.076-cm inner diameter with the matching impedance to the microphones. The probe tube of M2 was inserted into the oral cavity and glued onto the hard palate, where its tip was placed beneath the velum. The probe tube of M3 was inserted along the nasal floor through one nostril into the nasopharynx about 7.5 cm back from the nostrils. Mucous clogging of the tube was a potential factor to interfere with the pressure recording. To prevent this, the microphone signals were continuously monitored throughout the experiments so that mucous clogging when occurred could be quickly removed by injecting air into the probe tube.
An accelerometer (A1) was used to measure the vibration of the velum, which was placed in contact with the velum surface at about the central position on the nasal side by utilizing its wire's stiffness. Another accelerometer (A2) was set to measure the vibration of the laryngeal wall, which was placed on the surface of the laryngeal wall. The other accelerometer (A3) was attached on the lateral surface of the nostrils to evaluate the sound radiation from the nostrils. The accelerometers were ENDEVCO Model-22 devices with a weight of 0.14 grams.

Three male Japanese subjects, the same as in MRI experiment, participated in this experiment. The speech material was Japanese vowels, including /a/, /i/, /u/, /e/, /o/, and meaningless words consisting of the vowels with stop consonants of /b/, /d/, and /g/. For nasal coupling of the vowels, the analysis was focused on vowel segments, while for stop consonants the analysis was focused on the voice bar.

3.2. Nasal coupling in vowel periods

When modeling the transvelar coupling, the velum can be treated as a section in the vocal tract, where three physical quantities are configured: input sound pressure to the velum, output sound pressure from the velum, and volume velocity passed through the velum in velopharyngeal closure. The intraoral sound pressure is considered to be the input sound pressure, and the intranasal pressure is used to approximate the output of the velum, that is, the product of the volume velocity and the driving impedance of the nasal cavity looking from the velum. The volume velocity of the velum can be obtained by integrating the acceleration signal of the velum over time.

Figure 5 shows the spectra of the intraoral and intranasal sound pressures and the acceleration of the velum averaged over all vowel segments obtained from one of the subjects. As seen in the figure, the intraoral sound pressure can be considered as an energy source without any large resonance peaks, while the acceleration signal, as the input to the nasal cavity, shows a rapid slope down to 1 kHz. In contrast, the intranasal channel shows a resonance peak between 2 kHz to 3 kHz, which is formed by the input impedance of the nasal cavity.

Figure 6 shows the amplitudes of the speech sound (SS) recorded by M1, intraoral sound pressure (OP) recorded by M2, intranasal sound pressure (NP) recorded by M3, and the velum vibration (VV) recorded by A1 for three subjects. One can see that SS increases monotonically in the given order for all the subjects, while the OP and VV are decreasing. Variations of the NP show no obvious tendency. Amplitude of each channel is almost the same for S1 and S2, while for S3 the SS is higher and the other channels are lower than the other subjects.

3.3. Nasal coupling in voiced stop periods

In speech, there is a phenomenon that the voicing (seen as voice bar) starts before the release of closure in voiced stop consonants. The voice bar is conventionally considered to be radiated mainly from the larynx wall that is driven by vocal fold vibration, where the pharyngeal cavity actively enlarges for maintaining the difference of sound pressures above and below the vocal folds [13]. In this study, we testify this hypothesis by analyzing the multimodal data.

We employed the same approach as used in Section 3.2 to analyze the multimodal data during the voice bar. Figure 7 shows the SS, OP, NP, and VV for the three subjects during the voice bars. Comparing with the vowel period, the amplitudes of the SS and OP during the voice bar are independent of vowels. In vowel period, the VV depends on both the intraoral sound pressure and vowels-dependent thickness of the velum, in which velum vibration seems to be passively changed by those two factors (see Fig. 6 for details). In contrast, the velum maintains its vibration (VV) almost constantly for all vowels. In vowel periods, the amplitude of the VV was equal to or smaller than that of the intraoral sound pressure by a relative comparison among the modalities, while the amplitude is about 10dB larger than that of OP for Subjects S1 and S2. Although it is not clear why the ratio of VV to OP for the voice bar is much larger than that for the vowels, the large ratio implies that the speaker may actively control the condition of the velum for generating the voice bar. In the figure, however, S3 shows stronger voice bar but less velum vibration. We found that this subject has a difficulty to produce a voiced stop when his nostrils were obliterated. It can be considered that for him the velopharyngeal port is adjusted to open slightly, so that air flow was taken place the opening velopharyngeal port [7]. That is, S3 have a nasalized voice bar. Even in this situation, one can also see that the velum is controlled actively by comparing with velum vibrations during its vocal generation.

3.4. Source of the voice bar in voiced stops

In this section, we further estimate what the source of the voice bar is. Figure 8 shows the relation between speech sound (SS), intranasal sound pressure (NP), and larynx wall vibration (LV) recorded by A2 during the voice bar of the stops /b/, /d/ and /g/ for Subjects S1 and S2. This figure used linear scale in order to show the characteristics clearly for the small amplitude. One can see that the relation between the SS and NP is almost linear (the left panel), while the linearity between the SS and LV is less evident (the right panel). The
correlation coefficient between the SS and NP is 0.86, while the correlation between SS and LV is 0.37. It is indicated that the sound during the voice bar mainly depends on the nasal output, but not on the larynx wall vibration. The main component of the voice bar results from the nostril radiation due to the velum vibration.

Figure 7: The speech sound (SS), intraoral sound pressure (OP), intranasal sound pressure (NP), and velum vibration (VV) for three subjects during voice bar. S1, S2, and S3 denote the three subjects.

4. Discussion and conclusion

The purpose of this study is to testify the conjecture that the nasal cavity is involved in the production of non-nasalized sounds such as vowels and voiced consonants via the transvelar coupling. We used physiological, mechanical and acoustical measurements to investigate the transvelar coupling in the production of non-nasalized sounds. According to MRI movie data in producing CVCCV sequences, we found that the thickness of the velum varies with vowels. The velum thickness is dependent on the vowels no matter what the consonant is. It is thicker for back vowels and thinner for front vowel. The amplitude of the velum vibration during vowels is determined by both the velum thickness and the amplitude of the driving source, namely the intraoral sound pressure. The velum, however, seems to be actively controlled by speakers in generating the voice bar of the voiced stops. According to the correlation between speech sound, intranasal sound, and larynx wall vibration, it is found that the sound of the voice bar mainly radiated from the nostrils, which results from the velum vibration, but not from the larynx wall vibration. This finding may renew the conventional hypothesis that the voice bar is considered to reflect acoustic radiation caused by the vibration of the larynx wall when the vocal tract is completely closed at a certain location. In further study, we further investigate details involved in the mechanisms using numerical simulation.

5. Acknowledgements

This work is supported in part by the National Basic Research Program of China (No. 2013CB329301), and in part by the National Natural Science Foundation of China under contract No. 61233009 and No. 61175016. This study was supported in part by JSPS KAKHI Grant No. 25330190.

6. References

Assessing the individual ability to reproduce non-familiar speech sounds

Véronique Delvaux1*, Kathy Huet1, Myriam Piccaluga1, Bernard Harmegnies1

1 Institut de Recherche en Sciences et Technologies du Langage, Université de Mons, Mons, Belgium
delvaux@umons.ac.be

Abstract

This paper is concerned with the speaker-specific ability to process nonnative speech sounds, which we consider as a systematic source of variance that needs to be assessed if one wants to improve the control over the independent variables manipulated in SLA experimental studies. We adopt a pragmatic, task-oriented approach, based (i) on our participants’ performances in two reproduction tasks involving respectively a vowel- and a consonant set of stimuli, and (ii) on L1-typical realizations. Data are analyzed using a variety of quantification techniques, resulting in the development of 12 complementary indicators that allow to compare individuals in terms of overall, and specific aspects of their performances in the vowel and consonant reproduction tasks.

Discussion focuses on the possible mathematical refinements of the quantification techniques used in this paper, as well as the usefulness of the methods developed here to investigate the individual ability to reproduce non-familiar speech sounds.

Keywords: L2 learning, individual skills, phonetic talent

1. Introduction

This paper investigates speaker-specific abilities to acquire the phonetics and phonology of a foreign language. Performances in second language (L2) phonetic learning exhibit a large amount of interindividual variation (for a review, see Piske et al., 2001; Ehrman et al., 2003), which causes can be either related to contextual factors affecting each individual’s linguistic experience (e.g., “age” factors, Flege & MacKay, 2011) or to “internal factors”, i.e. factors that are inherent to the learner, such as cognitive abilities (Golestani & Zatorre, 2009), psychological aspects (Dörnyei, 2009), and linguistic abilities (Dogil & Reiterer, 2009). Contextual and internal factors are considered as interacting with each other in a complex way, as well as interacting with the specificities of the achievements that are required from the learner in a given situation. The latest, interactionist, models of second language learning aptitude propose a componential framework for mapping the interactions between tasks demands, context properties and learner factors (Robinson, 2009).

Recently, Jilka and collaborators proposed a comprehensive approach for testing “phonetic talent”, which assesses phonetic abilities of adult experienced L2 learners with a special emphasis on pronunciation (Jilka, 2009). Phonetic talent denotes a largely innate, neurobiologically grounded, individual skill which is part of general language aptitude, but may be separated from other specific linguistic skills such as grammatical talent in L2. Although phonetic talent is an appealing concept, its objective assessment is hindered by the difficulty to distinguish initial predisposition from the other interacting variables that have presided to each individual’s language development and may still influence his/her productions in a specific task, independently of his proficiency in any L2.

In this paper, we adopt a pragmatic, task-oriented view, with no strong hypothesis about innateness nor any attempt at enlightening the complex etiology of phonetic proficiency. We thus focus on the end result of the process, i.e. on the empirically observed spontaneous ability of adult speakers to accurately reproduce non-familiar speech sounds they are faced with, irrespective of the speakers’ language proficiency (either innate or acquired through development), and notwithstanding the differences between the sounds they are used to process and the sounds to be reproduced. This paper is an exploratory methodological account aimed at developing assessment tools for distinguishing individuals in terms of this specific ability.

2. Material and methods

2.1 Participants and stimuli

The participants were 10 native French speakers from Belgium, 5 female (S1-S5), 5 male (S6-S10), aged 24 to 42.

Two stimuli sets, a “vowel set” and a “consonant set” have been built using Klatt’s synthesizer (Klatt, 1980). The vowel stimuli set was made of 94 synthesized vowels that were evenly distributed over a mel scale F1*F2*F3 acoustic space. Total vowel duration (200ms) and F0 contour (from 110Hz to 90Hz) were kept constant across stimuli. The consonant stimuli set was made of 9 [ka] synthetic syllables, where [k] varied in voice onset time (VOT) from short VOT [k] to extra-long positive VOT [kʰ].

The [a] vowel was the same across stimuli, with a duration of 160ms and an appropriate F0 contour.

2.2 Experimental paradigm and sessions

Data were collected in two sessions: the “main” session and the “complementary” session. In the main session, the experimental paradigm comprised five successive parts: (i) production of L1 sounds (10 repetitions) to be used as control sounds in data analysis; (ii) training phase (vowel stimuli); (iii) test phase (3 repetitions of the 94 vowel stimuli); (iv) training phase (consonant stimuli); (v) test phase (5 repetitions of the 9 consonant stimuli). For the (ii) to (v) phases, the instructions were to “repeat the sound/syllable as faithfully as possible ‘as if it came from a foreign language’”. The complementary session was carried out a few weeks later in order to gather the additional data that were necessary to compute refined indices (Ind2 and Ind3, see below) of the participants’ behavior in the vowel reproduction task. This second session allowed us to collect 3 more repetitions of the vowel stimuli (4 speakers only).

2.3 Measures

The speech productions from the participants were segmented manually. For consonant productions, VOT was measured as the duration of the interval between burst onset and vowel
onset. For vowel productions, raw data consisted in the first three formant frequencies that were first automatically detected in the middle of the vowel using Praat (with adapted parameters for female speakers), then manually verified by two trained phoneticians examining spectrograms.

2.4 Data analysis

Data analysis consisted in (group and individual) comparisons between the acoustic properties of the stimuli and those measured in the reproductions. First, comparisons were carried out through linear regression analysis. Coefficients of determination (R-square) values were used to assess the contribution of the stimulus variance to the response variance, and slopes of the regression lines were compared to the 1-value signaling a perfect match between stimulus and response acoustic properties.

Second, we compared the acoustic properties of the stimuli and the reproductions using distance-based measurements, i.e. the absolute difference (in ms) between Stimulus VOT and Response VOT for consonant productions and, for vowel productions, the average Euclidean distance between the stimulus (s) and the response production (p) in the three-dimensional F1*F2*F3 acoustic space defined in mels (“Ind1”):

\[ \text{Ind1} = \frac{\sum_{s=1}^{S} \sum_{p=1}^{P} \left( \frac{1}{3} \sum_{i=1}^{3} (F_{sp} - F_{si})^2 \right)^{1/2}}{S \times P} \]  

Third, we developed two additional indices of the speakers’ performances in the vowel reproduction task. Ind2 is also based upon the Euclidean distance between stimulus and production, except that in this case, the inverse of the distance (-1/2 exponent) was taken into account, in order to obtain a number with variations positively correlated with performance in the task. Furthermore, in this case, the speaker was “calibrated” using his/her realizations of L1 vowels. Given that L1 has V vocalic phonemes and the speaker has realized V tokens of each, it is possible to identify zones of the vowel space corresponding to usual productions of the speakers, and zones where he/she is not used to produce vocalic sounds. The idea in Ind2 is to give higher reward to the success in imitating when imitation takes place in a region of the vocalic space the speakers do not spontaneously use in their usual practice of their L1. This is the reason for the weighting by the multiplicative term in the equation (2). It consists in the logarithm of the product of all the distances between a given production and each L1 vowel’s cluster centroid: the multiplicative term tends toward zero when at least one distance production/centroid tends toward zero. Thus, for a given realization, the product is large if the production resembles the target and if it is produced in a zone far from the ones corresponding with the speaker’s L1. Ind2 represents the average product over all the speaker’s productions:

\[ \text{Ind2} = \frac{\sum_{s=1}^{S} \sum_{p=1}^{P} \left( \frac{1}{V} \sum_{v=1}^{V} \log \left( \frac{1}{3} \sum_{i=1}^{3} (F_{sp} - F_{vi})^2 \right)^{1/2} \right)^{1/2}}{S \times P} \]  

In Ind3 (see (3), where “var” stands for “variance”), the similarity between targets and productions is no more the main point, and the approach is more statistical: it is based upon the analysis of variability in the reproduction task. When a speaker tries to attain a target, he/she produces realizations that fall around it in the reference space. If the speaker’s ability to reproduce the stimuli is high, his/her variability around the target in the reference space is random, and if no other source of variance is active, the variability is constant whatever the stimulus. On the other hand, if the speaker is strongly influenced by his/her L1, one can suppose that his/her variability will vary from one stimulus to another, depending on whether the stimulus is close or not to a region of the vowel space present in L1. Ind3 should therefore tend toward zero (all variances equal) for a speaker who performs well in the task.

\[ \text{Ind3} = \text{var}_p \left( \frac{1}{P} \sum_{p=1}^{P} \left( \frac{1}{3} \sum_{i=1}^{3} (F_{ps} - F_{si})^2 \right)^{1/2} \right)^{1/2} \]  

Two advantages of Ind3 are noteworthy. First, Ind3 implements the assessment of the distribution of the responses over the acoustic space in a continuous manner. Second, to a certain extent Ind3 circumvents the problem raised by the fact that, for some speakers more than for other speakers, the (anatomically or articulatory-induced) limits of their individual acoustic space diverge from the limits of the stimulus space. Since here the similarity between targets and responses is not taken into account, speakers who need to apply a large transformation to the targets’ formant values in order to repeat them faithfully are not at a disadvantage when their performances in the reproduction task are assessed.

3. Results

3.1 Group tendencies

The results of the linear regression analyses carried out on the reproductions from the main session (10 speakers) are illustrated in Fig.1 (vowel stimuli) and Fig.2 (consonant stimuli).

In the case of vowel stimuli, three linear regression analyses were carried out, respectively for F1, F2 and F3 (not shown here), pooling together data from all 10 speakers and 3 repetitions. These analyses revealed that as a group, the 10 speakers achieve a fair relation between their responses and the stimuli they were asked to repeat as faithfully as possible, at least in terms of F1 (R-square=0.75; \( F_{1,102} = 3084.84; p < 0.001 \)) and F2 (R-square=0.78; \( F_{1,102} = 3598.61; p < 0.001 \)) formant values. This result in turn indicates that the vowel task is appropriate in that it is neither too difficult (overall, the 10 speakers perform fairly), nor too easy (there is no ceiling effect). Moreover, given the overall high variability, the task can potentially elicit some inter-individual variation signaling individual differences in the ability to reproduce speech sounds. The slopes of the regression lines are respectively of 0.72 for F1 and 0.94 for F2, i.e. the responses formant values are overall closer to those of the targets in the case of F2 (a slope of 1 reflects a perfect match), whereas in the case of F1 the frequency span of the responses is shorter than that of the stimuli.

In the case of consonant stimuli, a single regression analysis was performed comparing the variances of response VOT and stimulus VOT (Fig.2). The results (R-square=0.14; slope of the regression line: 0.85) indicate that the 10 speakers’ VOT productions are moderately influenced by the stimulii VOT, or at the very least, that the speakers’ productions are less influenced by the stimuli properties in the case of consonants than in the case of vowels. Although the VOT measure is only one-dimensional (whereas vowel responses are assessed in a bi-dimensional formant space), one should not overlook the
fact that achieving long VOT for native French speakers involves acquiring new timing patterns between laryngeal and supra-laryngeal gestures.

3.2 Individual performances

3.2.1 Consonant stimuli

The results of the 10 separate linear regression analyses (one per speaker) from the consonant reproduction task are summarized in Table 1 in terms of slopes of the regression line and R-square coefficients.

Table 1. R-squares and slopes of the regression lines from the consonant reproduction task

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Slope</th>
<th>R-square</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.84</td>
<td>0.28</td>
</tr>
<tr>
<td>S2</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>S3</td>
<td>1.3</td>
<td>0.67</td>
</tr>
<tr>
<td>S4</td>
<td>0.41</td>
<td>0.03</td>
</tr>
<tr>
<td>S5</td>
<td>0.45</td>
<td>0.22</td>
</tr>
<tr>
<td>S6</td>
<td>1.72</td>
<td>0.32</td>
</tr>
<tr>
<td>S8</td>
<td>1.26</td>
<td>0.38</td>
</tr>
<tr>
<td>S9</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>S10</td>
<td>1.28</td>
<td>0.37</td>
</tr>
</tbody>
</table>

As illustrated in Table 1, results exhibit a large amount of inter-individual variation. Some speakers achieve virtually zero R-square values (S2, S7, S4) with varying intra-speaker responses variability (not shown here). For other speakers there is a moderate correlation between response VOT and stimulus VOT (e.g. S1, S6, S9, S10), sometimes with large intra-speaker variability. Finally, S3 performs the better. Note that for the 6 speakers exhibiting R-square values within the range of .2 to .4, slopes largely vary, i.e. from .45 to 1.72.

Moving now to the distance between the targets and the responses, an ANOVA was carried out with Distance (computed as the absolute difference, in ms, between stimulus VOT and response VOT) as dependent variable, and Speaker and Repetition as independent variables. Neither Repetition, nor the interaction between Speaker and Repetition yielded significant variations in Distance, whereas Speaker did (F(8,80), 21.73; p<0.001). However, this raw distance index is less efficient in terms of assessing the speaker-specific ability to reproduce long VOTs than the summary provided in Table 1 because this distance index remains short for the speakers who consistently produces short, L1-typical VOTs in response to all (long- and short- VOTs) stimuli.

3.2.2 Vowel stimuli

Three customized indices were computed in order to assess individual performances in the vowel reproduction tasks, i.e. on the data collected in the main session: Ind1 (10 speakers, 3 repetitions); on the data collected in the main and complementary session: Ind1, Ind2, Ind3 (4 speakers, 6 repetitions). By lack of space, only the latter results are presented below.

Results indicate that, for the three indices, S6 and S10 are the highest ranked, and S3 and S4 are the lowest ranked (Table 2). The indices mainly differ in how they assess the performances of S4. S4 performs particularly poorly in Ind3. In fact, in the case of S4, there is a large variance of the responses with regards to some of the targets whereas the variance is considerably smaller regarding other targets, so that the variance of variances is particularly high. In contrast, S10 is the highest ranked based on Ind3, not so much so because on average variances around targets are low, but because they are all very similar, so that the variance of variances is low.

Table 2. Ind1, Ind2, Ind3 scores and ranking orders with respect to performances (‘1’=best performer) in the vowel reproduction task (4 speakers, 6 repetitions)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Ind1 Score</th>
<th>Ind2 Score</th>
<th>Ind3 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>200Hz</td>
<td>62</td>
<td>3872</td>
</tr>
<tr>
<td>S4</td>
<td>216Hz</td>
<td>75</td>
<td>7475</td>
</tr>
<tr>
<td>S6</td>
<td>148Hz</td>
<td>80</td>
<td>2457</td>
</tr>
<tr>
<td>S10</td>
<td>137Hz</td>
<td>87</td>
<td>1552</td>
</tr>
</tbody>
</table>

Only Ind2 ranks S4 over S3, indicating that when performing the task S4 resorts less than S3 to her L1-typical realizations. If one compares the values obtained by S3 and S4 regarding the three indices, a possible interpretation is that S4 takes more risks when she attempts at faithfully repeating the stimuli, so that she is more prone to move out of her L1 routines (and is rewarded through Ind2). However, these out-of-habits excursions increase the variance of variances measured by Ind3. By contrast, S3 may be considered as a female speaker who is overall more in control of her productions (Ind3), and more efficient (Ind1), but who accomplishes the task by keeping closer to her phonetic routines in L1 (Ind2).

4. Discussion

The aim of this paper was to contribute to the assessment of the intrinsic speaker-specific ability to appropriately mobilize perceptual and productive processes in order to accurately reproduce non-familiar speech sounds. Although individual phonetic abilities are usually considered as playing an influential role in L2 acquisition (e.g. Jilka, 2009), we found no previous attempt in the literature to investigate this individual ability as itself, independently of the speakers’ proficiency in a specific L2, and based on direct measurements on speech production data.

We first investigated intrinsic speaker-dependent variability in two reproduction tasks on different speech materials, using regression analyses and stimuli-response distance measurements. Since the 10 participants had to reproduce oral vowels as well as voiceless plosive consonants, they had, on the one hand, to attain new, non L1-typical, targets in an acoustic/timbre space they already mastered, and on the other hand, to reorganize their timing patterns between already mastered laryngeal and supra-laryngeal gestures. Group analyses revealed that the two reproduction tasks elicited a great diversity in the subjects’ performances with neither ceiling nor floor effect.

A complementary data collection session was carried out in order to refine the quantification techniques appropriate to assess individual performances in a vowel reproduction task. Ind1 consists in a global assessment of the overall distance between the stimuli and the responses in the 3-formant space. As such, Ind1 represents an appropriate way of assessing the performance component of the ability to reproduce a large diversity of (including non-familiar) vowels, in purely acoustic terms. Ind2 and Ind3, on the contrary, take into account the fact that each speaker is equipped with an L1-dependent phonological system structuring his/her speech production and speech perception experiences. By addressing the way in which speakers are able to cope with the routinized dynamics of their own system, Ind2 and Ind3 are oriented towards the
competence component of the speakers’ ability to reproduce non-familiar vowels. The 3 indices have demonstrated a good ability in discriminating between the subjects. They are nevertheless quite different from one another, both in terms of underlying cognitive hypotheses (strong in Ind2 and Ind3, absent in Ind1) and of workload requirements (Ind2 requires prior calibrating of the subject, Ind1 and Ind3 do not). They also result in partially different outcomes, in that speakers are not ranked in exactly the same order according to the different indices.

Given the complementarity of the different indicators developed in this paper to assess the speakers’ performances in the reproduction tasks, we claim that it would be inappropriate to select one indicator over the others. Rather, we propose to operationalize the notion of ‘individual ability to reproduce non-familiar sounds’ as a compound of the most meaningful indicators arising from the analyses reported here. Fig.3 illustrates this approach for S3, S4, S6 and S10 using a radar chart. It provides a useful visualization of inter-individual differences in the various aspects composing their ability to reproduce non-familiar speech sounds. Fig.3 shows that S4 performs the poorest, with below average performances for all indicators but two. S3 and S10 are the best, S3 performing better in the consonant task whereas S10 performs better in the vowel task.

As a conclusion, we reported in this paper a first attempt to develop tools for assessing the individual ability to reproduce non-familiar speech sounds. Although promising, this first proposal clearly requires further mathematical refinements, theoretical developments and experimental work in the near future.

5. References


Fig.2. Linear regression analysis for consonant stimuli (10 speakers): response VOT as a function of stimulus VOT

Fig.3. Radar chart summarizing the performances of 4 speakers along 12 indicators: ‘Ind1’, ‘Ind2’, ‘Ind3’, ‘VOT distance’; from linear regression analysis: F1, F2, F3 and VOT coefficient of determination (‘R-square’) and difference-to-45° of the regression line slope (‘Angle’). All values expressed in z-scores, the higher the value, the better the performances
Formant strategies of professional female singers at high fundamental frequencies

Andrea Deme

Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary

Eötvös Loránd University, Hungary
deme.andrea@nytud.mta.hu

Abstract

When the soprano raises the fundamental frequency above the first formant of a vowel, a remarkable loss of acoustic energy and linguistic information occurs along with an abrupt change in the voice timbre. To avoid these effects, sopranos are assumed to tune their first formant to the raised fundamental frequency. The support for this claim is mostly based on formant data provided by indirect measurement methods and articulatory data, since direct acoustic data becomes more difficult (or even impossible) to obtain as the fundamental frequency gets higher. In the present study a new combination of measurement methods is introduced. The aim was to extract formant data of three sopranos in the entire set of the Hungarian vowel inventory in a wide pitch range. The results provide evidence for the technique of tuning the first formant to the raised high fundamental frequency in a substantial amount of data.

Keywords: formant tuning in soprano, inverse filtering, EGG, external excitation

1. Introduction

In high-pitched singing the fundamental frequency (f0 or f1) often exceeds the typical frequency region of the first formant (F1) of vowels. In these cases unaltered articulation of vowels would result in a remarkable loss of acoustic energy (i.e. sound level) and the loss of an acoustic cue (i.e. F1) which is regarded to be important for defining and identifying vowel quality. Singers of the Western operatic style who are required to sing loudly (without any amplification) supposedly tend to compensate for these losses by changing the articulatory configuration of vowels while producing them at higher pitches. This way they can enhance the efficiency of resonance utilization. Several investigations have been conducted to provide data on the techniques singers use at high pitch regions to avoid the negative acoustic consequences of raising f0 above the F1 of the vowels in speech (F1speech).

With respect to articulatory maneuvers Sundberg (1975), and Sundberg and Skoog (1995) concluded that pitch raising is accompanied by gradually increasing jaw opening in sopranos when f0 approaches the region of F1. Based on the traditional interpretation of the interrelation between articulatory movements and acoustics, the authors suggest that the articulatory data they obtained reflect the acoustic event of F1 “tuning” in an indirect way: they propose that F1 is tuned to f0 when f0 reaches the region of F1speech during pitch raising. Sundberg and Skoog (1995), however, also noted that the modification of jaw opening starts only at a higher f0 in the case of the close-mid vowels /e/ or /ɛ/ whereas (in some singers) it is not clearly present in the close vowels /i/ or /u/. The authors claim that in the case of close vowels singers tend to tune their F1 as well as in all other cases, but this tuning is probably achieved by decreasing the degree of the tongue constriction (and not by modifying the jaw opening) for reasons of articulatory convenience. As opposed to Sundberg (1975), and Sundberg and Skoog (1995), a study by Bresch and Narayanan (2010) found that the dependency of F1 on f0 is only validated in the case of the close vowels /i/ or /u/. They also conclude that this dependency is probably more singer-dependent than it was suggested before. The authors base their claims on articulatory and formant data obtained from MRI-based vocal tract area function calculations.

Direct resonance data, on the other hand, are much more difficult to obtain for high-pitched singing. Due to wide harmonic spacing, the raised f0 causes undersampling of the vocal tract (VT) transfer function in the output sound. Consequently, analyzing the spectrum of the output signal is an inefficient way to determine formant frequencies (see e.g. Deme 2012). Therefore, the investigation of the acoustic characteristics of high-pitched sung vowels requires novel methodology to extract formant data in order to by-pass the problem of low resolution of the VT transfer function. The studies in this field concentrate mostly on the open vowel /a/ at particular pitch ranges. Sundberg (1975) used an external vibrator applied at the larynx for measuring five vowels at four fundamental frequencies below 700 Hz. Hertegärd and Gauffin (1993) investigated the production of /a/ at 250, 390 and 750 Hz by inverse filtering the flow signal recorded with a Rothenberg mask. Joliveau et al. (2004) applied external excitation at the mouth for measuring four vowels in a pitch range below 1100 Hz. Garnier et al. (2010) used the same technique for investigating the vowel /a/ above 440 Hz. With regard to the cases when f0 > F1speech the above mentioned studies agreed on the tendency of F1: f0 tuning. However, Joliveau et al. (2004) also noted the lack of this tendency in vowels that use lip rounding above approximately 900 Hz. Some of these studies have also reported that the methods used might have also biased the data to some extent or required some modification of the singing technique intended to be observed: Hertegärd & Gauffin (1993) noted that their subject complained that the mask hampered her in opening the mouth, while Joliveau et al. (2004) and Garnier et al. (2010) instructed their singer to sing very softly and without vibrato even at high pitches which might have also affected the control of formant tuning strategies.

The aim of the present study was to extend on the previous research by investigating the entire vowel set of a particular language in a wide pitch range (covering most of a soprano’s range), thus to discover the formant value changes accompanying pitch raising. This aim was achieved by the use of a new combination of some of the previously introduced methods, in a way that the problem of measuring formant frequencies at high f0s is resolved while the influence of the measurement method on the data obtained is minimalized. It was hypothesized that F1: f0 tuning occurs in each vowel in the cases of f0 > F1speech regardless of the degree of closeness or lip rounding. However, it was also suggested that F1: f0 tuning appears in
correspondence with the closeness of the vowel during pitch-raising since the starting point of the tuning depends on the value of $\text{F}_1$.

2. Subjects, material, method

Three professional Western operatic female soprano singers were recorded producing nine Hungarian vowels /$a$ / $e$ / $i$: /$i$ / $o$ / $u$ / $y$/ at six fundamental frequencies in three octaves from 175 Hz to 988 Hz (musical notes: F3, B3, F4, B4, F5, B5) and in speech. Although short counterparts of long vowels (/$i$ / $o$ / $u$ / $y$/) were not involved, and length is phonologically distinctive in Hungarian, the vowel set investigated here can be regarded as a representation of the entire vowel inventory. Since the duration of vowels is determined by the musical notes, length and quality distinction between short and long vowels become hard to interpret in singing. Each combination of vowels and fundamental frequencies was recorded twice in each singer’s production which resulted in $(3 \times 9 \times 7 + 2 = 378$ stimuli. Before recording one set of vowels, reference pitch was provided to the singer through headphones.

Two signals were recorded simultaneously. On the first channel the audio signal was captured by an omnidirectional microphone at 30 cm distance. On the second channel the signal of an electroglottograph (EGG) attached to the singer’s neck was recorded. To resolve the problem that high-pitched sung vowels become undersampled due to wide harmonic spacing, low frequency external excitation (electrolarynx) was used during the recordings; while singing the vowel, the singers were instructed to phonate for a few seconds than turn the electrolarynx on, freeze the articulation and hold their breath (with closed glottis). This way at the end of each vowel the buzzer signal substituted the voice source and resampled the VT configuration. The output of this filtered buzz sound was also recorded with the microphone. An example of this “double sampling” is shown in Figure 1. The speech and EGG signals were time-aligned to compensate for the transit time of sound (from the larynx to the mouth) which was determined to be approximately 0.8 ms.

![Figure 1: Narrow-band spectrogram of the vowel /$a$/ in speech sampled by the vibrating vocal folds (1) and the electrolarynx (3). In (2) the singer was phonating with the electrolarynx already switched on.](image)

The first (F1) and second (F2) formants of the vowels were determined by manual inverse filtering of the glottal flow signal (the pressure signal integrated) by means of the custom-made DeCap software (Svante Granqvist, KTH). The derivative of the EGG signal (dEGG) was also fed into DeCap and used to support the measurements (see Figure 2). The principle of the method is to compensate for the filter function of the VT, thus to restore the spectrum and waveform of the voice source by means of manually adjustable filters corresponding to the formants of the sound. In the case of correct adjustment the result is a smooth spectrum envelope and a smooth flow spectrogram (with a ripple-free closed phase) that would be characteristic of the source signal (if it could have been captured at the glottis) (see Hertegård & Gauffin 1993, Sundberg et al. 2013). The accuracy of the analysis was increased by the use of dEGG which reflects the moment of vocal fold contact (as a peak in the signal), thus allowing to designate the maximum declination rate of the transglottal airflow and the starting point of the closed phase needed for adjusting the filters (see Henrich et al. 2004).

To verify formant value estimations the electrolarynx recordings were also filtered (in phase 3 displayed on Figure 1) and the results were compared with the formant and bandwidth frequencies obtained by the filtering of the flow. Therefore, the reliability of the measurements was enhanced. The two repetitions of each stimulus were measured separately and averaged in the analysis.

![Figure 2: Inverse filtering with DeCap (after Sundberg et al. 2013). The upper panel shows the inverse filtered signal and the dEGG (dark and light colors, respectively); the lower panel shows the audio spectrum, and the spectrum of the filtered flow (light and dark colors, respectively). The arrows with the labels F1, F2, F3 show the first three formants adjusted for inverse filtering.](image)

3. Results

One of the three sopranos did not manage to perform each of the stimuli at the highest fundamental frequency $B_1$ (988 Hz). Therefore, some data points are missing in Figure 3 and the displacement of vowels in the acoustic vowel space accompanying pitch raising was evaluated based on the comparison between speech and the second highest $f_0$, i.e. $F_0$ (698 Hz) where no missing data were present.

3.1. Vowel formants in singing: F1

Figure 3 shows the first and second formants of the nine vowels as a function of the fundamental frequency for each singer separately and the harmonics of the $f_0$ (dashed lines). On this figure, apart from moderate variability among the singers, clear-cut tendencies can be observed. When $f_0$ approached (but not necessarily reached) the region of $F_1$ the singers tuned $F_1$ to the raised $f_0$ in each vowel. From that point on, the tuning of the $F_1$ was parallel to the raising of the $f_0$. In other words, no exception of $F_1$ tuning was found at high $f_0$s, and the lower limit of the tuning was dependent on closeness: it was the highest (698 Hz) for the open and open-mid vowels /$a$ / $e$ / $i$ /, lower (494 Hz) for the close-mid vowels /$o$ / $e$ / $\ddot{e}$ /, and the lowest (349 Hz) for the close vowels /$u$ / $y$ / $i$ /. Below these critical $f_0$ values $F_1$ was more independent of $f_0$ in each vowel and it was realized in the vicinity of the value of $F_1$.

The results seem to be partly inconsistent with the findings of Joliveau et al. (2004), as they noted the failure of the $F_1$: $f_0$ tuning in the case of vowels produced with lip rounding. However, they observed this tendency around and above 1 kHz, whereas in the present study only 988 Hz was reached.

91
3.2. Vowel formants in singing: F2

The value of F2 was more independent of pitch, especially in the case of front vowels /a, e/. In the front vowels /æ, y: a slight decrease of F2 was observed starting at f0 where F1: f0 tuning also begins. In the back vowels /b, o: u/, however, a stronger decrease of the F2 values was observed.

3.3. The changes of the acoustic vowel space

In terms of the acoustic vowel space, the centralization tendency described in 3.1 and 3.2 results in the collapse of the front–back distinction and in a shift towards the position of /æ/ as seen in Figure 4. According to A of Figure 4, vowels in speech are well-separated along the two axes (corresponding to the front–back and close–open dimensions, respectively). In singing at F3 (698 Hz), however, only the front–back distinction is preserved to some extent.

Figure 3: Formant values (F1, F2) of the three sopranos as a function of the fundamental frequency. Each marker corresponds to a formant frequency of one singer at a particular fundamental frequency. The grey dashed lines represent the first three harmonics of the voice source (h1, h2, h3).

These tendencies mean that front and back vowels start to converge at the higher f0s (namely from about F4 or B4, or approximately 350–500Hz), and that the degree of the change in F2 is dependent on the degree of backness: among the front vowels it is /æ: i/, while among the back vowels it is /o: u/ that is affected the most.

4. Discussion and conclusions

The presented study focused on formant tuning strategies of sopranos trained according to the Western operatic singing technique. The singers of this technique are required to provide a homogenous timbre through their entire pitch range, and a loud voice without amplification. When f0 is higher than F1 of a vowel in speech (as often happens in sopranos), the acoustic energy of the radiated sound decreases remarkably, thus even more vocal effort is being needed to achieve the desired sound level. Increasing the subglottal pressure alone to increase sound level would strain the vocal folds, hence it must be avoided. Effective utilization of resonances, however, can help the singer to supplement vocal effort. Consequently, resonance strategies are very important in sopranos’ practice.

Based on the results, we can conclude that all professional sopranos in our study employed generalizable strategies for resonance tuning in a way that is consistent with the description in prior literature. It was demonstrated in a substantial amount of material (the entire set of the Hungarian vowels, in a wide pitch range, in 3 singers’ production) that professional sopranos tend to tune their F1 to or slightly above the frequency of f0 if the raised f0 approaches the F1 of the vowels in speech. It should be emphasized, however, that the articulatory strategies singers might have applied to achieve this goal are not extractable from these acoustic data alone. In other words, the present data are only indicative of resonance strategies and not the articulatory strategies that the singers use in their practice. Accordingly, the debated issue whether singers lowered the tongue or increased the jaw opening to raise F1 in close vowels (see further in Sundberg and Skoog 1995) cannot be answered here.

The lower limit of F1: f0 tuning was dependent on the closeness of the vowel in accordance with the assumption that F1: f0 tuning starts when f0 approaches the region of F1 typical of the vowel in speech. Indeed, in close-mid and close vowels F1 tended to be increased steeply through the studied pitch range, parallel to pitch raising. In the present study one of the results of Joliveau et al. (2004), namely the lack of F1: f0 tuning in rounded vowels at high f0s was not replicated. This inconsistency might shed light on the connection between formant tuning and the pitch range of singers. In the study of Joliveau et al. the failure of the tuning was observed above 900 Hz, while most of the singers reached 1042 Hz during the experiment which means that the upper boundary of the singers (most probably) lied above the critical 900 Hz. By contrast, the singers in the present study reported that the upper
boundary of their ranges was reached at the highest fundamental frequency studied (988 Hz). Garnier et al. (2010) noted that there is an upper limit of F1: f0 tuning close to 1100 Hz, above which the tuning of F2: f0 starts, if this region still belongs to the singer’s pitch range. On this basis, two explanations can be suggested. One possibility is that the singers of the present study reached the top of their ranges at 988 Hz, thus they did not need any other strategy to reach higher pitches with the required voice quality and loudness; however, they managed to expand the limit of F1: f0 tuning in vowels with lip rounding to a small extent. The second explanation is that the singers reached the top of their ranges because they did not have any other strategies which would have been able to substitute F1:f0 tuning. The adequacy of these suggestions needs further investigation.

The changes of F2 differ systematically according to backness. Also, they were found to accompany F1: f0 tuning. While front vowels’ F2 decreases slightly, F2 of back vowels increases to a greater extent. It can be suggested that the changes of F2 are the concomitant result of the increased jaw opening (as Joliveau et al. 2004 suggest). As the mandible is lowered, the constriction caused by the tongue is less and less narrow. To some extent this “loss” can be compensated by increasing the height of the dorsum. However, in their articulatory modeling study Lindblom and Sundberg (1971) concluded that at a sufficient jaw opening (at 23 mm in their data) this loss of constriction cannot be compensated for anymore; consequently, at this degree of opening no velar articulation is possible either. The authors claim that front vowels are also affected by jaw opening, but only to a lesser extent. The widening of the constriction results in higher F2 in back vowels and a slight lowering of F1 of the front vowels. Therefore, the decrease of the acoustic separation between front and back vowels is expected, as demonstrated in the present data. The changes of F1 and F2 resulted in reduced acoustic vowel space and a shift towards the position of /æ/ as the f0 increased.

What might be the consequences of these acoustic changes regarding perception? Although in the present data vowels converged and overlapped more and more with pitch raising, it was also observed that the front–back distinction was preserved at even moderately high f0s (i.e. 698 Hz). Accordingly, it can be expected that the distinction of the back and front vowels will uphold longer with pitch-raising also in perception (thus confusions in the identification task would include mainly back–back and front–front pairs), as demonstrated already in Deme (2012) and Deme (in press). Nevertheless, relating the exact formant data of the present study to perceptual tendencies in further detail may easily be misleading. We should bear in mind that the researcher faces a very difficult task in extracting formant data from the output sound in singing (particularly above about 400-500 Hz). The undersampling of the vocal tract transfer function is, however, not only influencing the efficiency of measurement techniques, but also speech perception. That is, at high pitch no exact formant frequencies are extractable for the auditory perception either: at a particular f0 only the same harmonics with different amplitude are available. For that reason, the author’s ongoing work also includes vowel identification tests on the present material which correlates the perceptual tendencies to the raw acoustic output, as well as to the formant data obtained here. These two approaches are supposed to clarify jointly which of the acoustic parameters may account for the perceptual tendencies accompanying pitch raising. The preliminary results suggest that listeners might be able to extract formant values effectively (and rely on them in vowel identification) even at the fundamental frequency of F3 (698 Hz). At B3 (988 Hz), on the other hand, they simply seem to identify the first two harmonics as F1 and F2, respectively, in each of the vowel qualities produced by the singers.

Manual inverse filtering is a method that has already proved to be successful in formant value estimation in male singing voice on fundamental frequencies below 500 Hz (see Sundberg et al. 2013) and also in sopranos for the vowel /a/ (Hertegård & Gauflin 1993). In the present study inverse filtering was supported by the derivative of the electroglottograph signal (dEGG) according to the suggestion of Henrich et al. (2004) and Sundberg et al. (2013). The reliability of the measurements was enhanced by the use of external excitation (electrolarynx) which allows for measuring formant frequencies independently of the fundamental frequency of the voice. The combination of these methods appeared to be successful in the formant value estimation in singing, even at very high pitches. The present study reports the first attempt to describe the acoustic modifications caused by high pitched singing on the entire vowel set of a particular language.

5. Acknowledgements

The help and support of the following personnel and institutes is gratefully acknowledged. The recordings were made in the Kemptelen Farkas Speech Research Lab of the Research Institute for Linguistics, Hungarian Academy of Sciences. The acoustic measurements were carried out in the Department of Speech, Music and Hearing of KTH Royal Institute of Technology, Sweden, with the help of prof. emer. Johan Sundberg, prof. Sten Ternström, and Svante Granqvist, PhD on a study visit funded by the Campus Hungary program of the Balassi Institute, Hungarian Scholarship Board Office.

6. References


Coordination of tongue tip and body in place differences among English coronal obstruents

Donald Derrick1,2, Romain Fiasson2, Catherine T. Best1
1 MARCS Institute, University of Western Sydney, Australia
2 New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, NZ
D.Derrick@uws.edu.au, romain.fiasson@canterbury.ac.nz, c.best@uws.edu.au

Abstract

Using electromagnetometry tracking of the tongue, Best et al. (2010, 2014) have demonstrated that Wubuy, an Australian language with four coronal stop places, shows significant differences in tongue tip vs. tongue body motion range and motion coordination contrasting apicals and laminals. Here we continue this line of inquiry with three coronal obstruents in English, the apical alveolar stop /d/ and alveo-palatal affricate /dʒ/ vs. the laminal dental fricative /θ/. The results show support for tongue tip/body motion range differences between /d/ and /θ/ across vowel contexts. They also showed a tongue tip/body motion coordination distinction between the apical /d/ and laminal /θ/, which was significant for /i/ and /u/ but not /ə/ contexts. Results are consistent with the Wubuy findings (Best et al., 2010, 2014) despite the differences in the coronal obstruent contrasts of the two languages, suggesting an apical/laminal distinction in tongue tip/body coordination.

Keywords: articulatory phonology, coronal obstruents, apical-laminal distinction, tongue tip/body coordination

1. Introduction

Of all the speech articulators, the tongue tip is the most flexible and versatile for production of obstruct consonants, that is, stops, affricates and fricatives. This flexibility is evidenced by the huge variety of tongue-tip, or coronal, consonants that exist cross-linguistically. Importantly, however, languages also vary widely in which coronal obstruents they use contrastively, as well as in the phonetic realizations of those coronals, which can vary notably even within a language (e.g., among regional accents or specific talkers or contexts). A fundamental issue, then, is the interplay between the universal and language-specific forces that underlie the way coronals are produced. One well-accepted characteristic of coronal obstruents that would lend itself to examining universal vs. language-specific properties is the distinction between apicals, with tongue tip constriction, and laminals, with anterior tongue blade constriction (Butcher & Tabain, 2004; Flemming, 2003). It has also been proposed that the apicals are distinguished from the laminals by tongue tip orientation – up for apicals, and down for laminals (Brownman & Goldstein 1989). Both analyses focus on tongue tip only.

But despite the versatility of the tongue tip, its spatial and temporal motion is necessarily constrained by the positioning of the tongue body because the tip (including the blade just behind the tip) is of course attached to the body, and thus its motion is at least partially determined by the posture and motion of the tongue body. Therefore, an important and possibly universal characteristic of coronal consonants should be the dynamic coordination between tongue tip and tongue body movements.

Here we examine possible universal vs. language-specific characteristics of the time-varying coordination between the tongue tip and the tongue body during coronal obstruent production. At one extreme are languages such as Wubuy (Nunggubuyu), an Australian language that uses a very rare four-way coronal place distinction within the stop manner class: two apicals (alveolar [t], retroflex [ʈ]), and two laminals (dental [θ], postalveolar [ʃ]). In an electromagnetic articulometry (EMA) study of Wubuy coronal stop production, which tracked mid-sagittal flesh-points on tongue tip and tongue body, Best et al. (2009; 2010; 2014) found that tongue tip/tongue body motions are more tightly coupled for laminals than for apicals, which use a more stabilized TB to support lever-like actions by the TT.

1.1. The Present Study

But few languages constrain coronal production within a single manner class in such a crowded place of articulation space. Therefore we decided to test whether this pattern holds for English, which has only a three-way (voiced) coronal obstruent contrast, as represented by the alveo-palatal affricate /dʒ/, alveolar stop /d/, and dental fricative /θ/. (see Figure 1).

Figure 1: English coronal obstruents in a 2 x 2 matrix contrasting apical vs. laminal closure by anterior vs. posterior release. NOTE: we schematize the alveo-palatal affricate as having apical anterior closure (solid line) but laminal posterior release (dashed); closure and release are congruent for /d/ and /t/ (solid line only). Figures based on those in the Proctor et al. (2010) modeling study of Wubuy coronal stops.

English has many other coronal obstruents: /h/ and /s/ are anterior apicals like /d/; /θ/ is an anterior laminal like /ð/; /ʃ/; /ʒ/ are posterior laminals; and /tʃ/ has apical anterior closure with laminal posterior release like /ɔtʃ/. However, none of them display the posterior apical or laminal closure+release of

94
of the Wubuy retroflex [ʈ] and post-alveolar [c], respectively...

More importantly, the place of articulation differences among the three English coronal consonants are confounded with manner of articulation differences, specifically, apical is confounded with stop manner and laminal with fricative manner. Thus, the same type or degree of apical/laminal organization of tongue tip and tongue body might not be as clear in English as it seems to be in Wubuy because of the added production/perceptual cues that result from these different manners of production. Therefore, we ask whether anything like the tongue tip/body coordination that appears to define the apical vs. laminal distinctions observed in Wubuy is also seen in production of these English coronals.

1.1.1. Hypothesis 1

If the apical/laminal difference in tongue tip-tongue body coordination is universal, then it should differentiate English apical /d/ and possibly /ð/ from laminal /ð/, analogous to Wubuy apical versus laminal coronal stops.

1.1.2. Hypothesis 2

But if contrastive tongue tip-body coordination is language-specific (e.g., distinguishing coronal places within a manner class), then it may not reliably differentiate among English /d/-/ð/.

2. Methods

Nine native English speakers participated, aged between 20 to 61 years. There were 4 females and 5 males. There were 4 American and 5 Australian English speakers.

Participants were tested at the speech production lab of the University of Western Sydney’s MARCS Institute. They were seated in a non-metallic chair sitting beside the NDI WAVE articulometer, where wired magnetic sensor coils were taped over their left and right mastoids and nasion. Sensors were glued along the mid-sagittal line to their tongue body about 2 cm away from the circumvallate papillae (to avoid the gag reflex), about 1 cm from their tongue tip, and in between the two at the tongue blade. Sensors were also glued to the gum below their lower incisors, and on the border at the centers of their upper and lower lips.

Participants read aloud 10 blocks of sentences, presented on a computer screen via DMDX. Each block contained /d/, /ð/ and /ð/ in / Cu, /Ci/ and /Cu/ contexts (Table 1), in the carrier phrase “Now I want a _____ around her.” The target items and carrier were designed to be phonetically comparable to the Wubuy target items (Best et al., 2010; 2014). Occlusal plane and palate traces were then collected.

2.1. Data analysis

We corrected for head motion using the data from the nasion and mastoid reference censors, and rotated the dataset

<table>
<thead>
<tr>
<th>1.1.1. Hypothesis 1</th>
<th>1.1.2. Hypothesis 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>If the apical/laminal difference in tongue tip-tongue body coordination is universal, then it should differentiate English apical /d/ and possibly /ð/ from laminal /ð/, analogous to Wubuy apical versus laminal coronal stops.</td>
<td></td>
</tr>
<tr>
<td>But if contrastive tongue tip-body coordination is language-specific (e.g., distinguishing coronal places within a manner class), then it may not reliably differentiate among English /d/-/ð/</td>
<td></td>
</tr>
</tbody>
</table>

2.1.1. Locating Tongue Tip Gestures

For both measures, we located the tongue tip gestures for the coronal consonant of each token, using MVIEW’s find gest routine (Tiede, 2005, 2010, Matlab, 2012). The onset of the consonant constriction gesture is identified by a critical threshold of increase in velocity (GONS), which is defined conventionally as 20% of the maximum or peak velocity (PVEL1) prior to constriction. The articulator slows down as the constriction target is achieved. The onset of the consonant closure is arbitrarily defined as 20% of the maximum velocity, named the nucleus onset (NONS) of the consonant closure period. The offset of the closure period, or nucleus offset (NOFF), is reached when the articulator reaches 20% of its maximum velocity during the release of the constriction (PVEL2). The velocity of the articulator then slows down again, with the gesture offset (GOFF) defined as a decline to 20% of PVEL2. The center point of the NONS to NOFF closure period is Mid-C (see Figure 2).

2.1.2. Short-term velocity correlation

Once the tongue tip gestures were located, the tangential velocity was computed for each sample of the tongue tip and tongue body sensors’ motion trajectories using Euclidean distances between samples and the central difference method. A rectangular sliding window of 7 samples (~30 milliseconds) was shifted sample-wise over the tongue tip and tongue body velocity signals, beginning at the tongue tip gesture’s GONS and ending at GOFF.
The static, or near-zero velocity sections in both the tongue tip and tongue body trajectories can result in high correlations. However these correlations are spurious since the hypotheses concern simultaneous vs. not simultaneous movements. To avoid these false correlations, each value was multiplied by the average tongue tip and tongue body velocity within the sliding window.

The procedure results in a series of correlation values over time for each target consonant, themselves based on the average (mean) velocities within their window. The series of correlation values are then averaged again for a single mean velocity correlation value per token. The correlation values were not rectified, so values before averaging can be positive or negative. For our purpose this is preferable since a negative correlation means that either the tongue tip or tongue body is accelerating while the other is decelerating. This outcome strongly hints at phase differences and so should have an opposite influence to positive correlations, indicating in-phase movements (see Figure 3).

![Figure 3: Schematic of the Short-term Velocity Correlation Measurement.](image)

The prediction is that laminals should have higher values (more synchronous movements of tongue tip and tongue body) than apicals.

### 2.1.3. Quad ratio

In addition, the spatial 2D (midsagittal) position of the tongue tip and tongue body sensors were taken at the time points of GONS (1), GOFF (2), NONS (3), and NOFF (4). The area of the resulting quadrilateral was then computed for the tongue tip gesture. Similar measures were taken for the tongue body gesture based on the above four tongue tip gesture times. This resulted in a single value representing the area of motion transcribed by the tongue tip gesture and a second value for the area transcribed by the tongue body for the same points in time. A schematic of this technique can be seen in Figure 4:

According to the hypothesis that for laminals the tongue tip and body move in tandem, while for apical the tongue body is relatively more stabilized during tongue tip motion, the tongue body gesture area should be significantly larger for laminals than for apicals, while the tongue tip gesture area should be the similar for both. However, given that speaker variability due to different palate and tongue shapes and sizes, the self-normalising ratio of tongue tip gesture area divided by tongue body gesture area was chosen for analysis. To make results easier to handle numerically and to generate a statistically normal distribution, a logarithmic transformation was applied to these ratios. The result is that positive values indicate that the tongue tip gesture area is larger than the tongue body gesture area. A negative value indicates the opposite. The prediction follows that the normalized tongue tip/body area ratio should be higher for apicals than it is for laminals, even for English coronal obstruents, if the universal hypothesis about tongue tip and body coordination for coronals is correct.

![Figure 4: Schematic of the QuadRatio measurement.](image)

### 2.1.4. Statistics

Tokens where gestures could not be found or where gestures did not line up appropriately with the consonant (270 of 641 observations, or 42.1% of the tokens) were excluded in order to avoid conflating vocalic gestures with the intended consonantal gestures. The numbers were this high due to the difficulty participants experienced in repetitively producing coronals with a pellet glued to the tongue tip. To compensate for this, statistical analysis was conducted on the two measures using generalized linear mixed effects models in R (R Core Team, 2013) as they are extremely robust against imbalances in datasets. For both QuadRatio and Velocity Correlation, the first test checked for interactions between the consonants and their vocalic contexts. Assuming no interactions, a second test compared the results of the measure against the consonants alone.

### 3. Results

#### 3.1.1. Short-term velocity correlation

For the short-term velocity correlation, there was a significant main effect of vowel context such that the correlation was lower for /u/ and /i/ than for /a/, and a significant pain effect such that the correlation was lower for /ð/ than /d/, as seen in Table 2.
There was also an interaction between consonant and vocalic context such that the short-term velocity correlation was significantly higher for /ð/ than /d/ in the /uCu/ and /iCi/ context, and significantly higher for /d/ than /dʒ/ in the /uCu/ context, as seen in Table 2 and Figure 6.

### 4. Discussion and conclusion

Both the quad ratio and velocity correlation results show a distinction between the laminal /ð/ and the apicals /d/ and /dʒ/. The quad ratio result is consistent with that seen in Wubuy. The short-term velocity correlation’s consistency varied significantly by vocalic context: laminal /ð/ showed a higher velocity correlation than the apicals /d/ and /dʒ/ in the /uCu/ context, and a higher correlation than /d/ in the /iCi/ context. However, in the /uCu/ context this result vanished for English.

The alveo-palatal affricate /dʒ/ trended with apical /d/, indicating that the apical vs. laminal distinction appears based on onset, and anterior vs. posterior appears based on release.

Thus, we found robust support for Hypothesis 1, i.e., that tongue tip/tongue body coordination differences between apicals and laminals are universal. However, the specific nature of tongue tip/tongue body coordination in apicals vs. laminals is partially language-specific as vocalic context strongly influenced velocity correlation.

### 5. Acknowledgements

Many thanks to the reviewers for their helpful critiques.

### 6. References


Articulatory Effects of Prediction During Comprehension: An Ultrasound Tongue Imaging Approach
Eleanor Drake¹, Sonja Schaeffler², Martin Corley¹
¹Department of Psychology, University of Edinburgh, Edinburgh, UK
²Clinical Audiology, Speech and Language Research Centre, Queen Margaret University, Edinburgh, UK
E.K.E.Drake@sms.ed.ac.uk, sschaeffler@qmu.ac.uk, Martin.Corley@ed.ac.uk

Abstract
We investigated whether effects of prediction during spoken language comprehension are observable in speech-motor output recorded via ultrasound tongue imaging: Predicted words can be specified at a phonological level during reading comprehension, and listening to speech activates speech-motor regions. It has been suggested that speech-motor activation may occur during prediction of upcoming material (Pickering & Garrod, 2007). Speakers model their own upcoming speech, with the effects being observable at an articulatory level in the form of anticipatory co-articulation. We investigated whether the effects of prediction as a listener can also be observed at an articulatory level. We auditorily presented high-cloze sentence-stems, immediately followed by presentation of a picture for naming. Picture names either fully matched the omitted sentence-cloze item or mismatched it at onset (e.g., TAP-“cap”). By-condition differences in picture-name articulation indicated that prediction of upcoming material during speech listening can engage speech-motor processes.

Keywords: ultrasound tongue imaging, prediction, Delta-technique, motor-speech

1. Introduction
Language comprehension involves the prediction of upcoming material, in addition to the processing of perceptually available input (e.g., Altmann & Kamide, 1999; DeLong et al., 2005; Federmeier, 2007). Listening to speech activates neural regions associated with motor-speech planning and execution, and modulates tongue muscle excitability (Fadiga et al. 2002; Pulvermüller et al., 2006; Watkins & Paus, 2004; Wilson et al., 2004). Such communicative resonance may reflect involvement of the speech-motor regions in prediction during comprehension: It has been suggested that the synthesis process involved in predicting another’s upcoming speech may recruit mechanisms more typically associated with speech production (e.g., Pickering & Garrod, 2007).

ERP evidence demonstrates that, for written input, predicted linguistic material can be specified at a phonological level during reading comprehension (DeLong et al., 2005). However, it is unclear if, and at what levels, such prediction might involve the infrastructure and mechanisms of speech production. Prediction during language comprehension is performed incrementally. Therefore, when investigating prediction of spoken language, some spoken language must be presented to the listener prior to the critical manipulation point. Under such circumstances it appears not to be possible to distinguish the neural activation posited to be associated with the synthesis of upcoming input (i.e. prediction) from that associated with the analysis of perceptual input.

In the current study we instead employ an articulatory imaging technique to investigate the effects of prediction as a listener on motor-speech activity itself. When a speaker predicts their own upcoming speech output this is observable at an articulatory level in the phenomenon of anticipatory co-articulation (e.g., Farnetani & Recasens, 1997). Anticipatory co-articulation can be observed via ultrasound tongue imaging (e.g., Zharkova & Hewlett, 2011). We investigated whether ultrasound tongue imaging (henceforth UTI) would reveal articulatory-level effects of prediction during comprehension. We reasoned that if the speech-motor system is activated during prediction as a listener, effects of prediction might be observable in articulation when the listener speaks.

Participants named pictures in three contexts whilst their articulatory movements were recorded via UTI. In the Control condition a visual fixation point was presented for 3 seconds prior to picture presentation; in the Match condition, a sentence-stem predicting the upcoming picture-name was auditorily presented immediately prior to picture presentation; in the Mismatch condition a sentence-stem predicting a rime-partner of the upcoming picture-name was auditorily presented immediately prior to picture presentation. In this way the experimental conditions differ only in whether the auditory linguistic context predicted the target picture-name or an alternative picture-name. Any by-condition differences in picture-name realizations (articulation) must therefore reflect an effect of prediction as a listener.

2. Method
2.1. Participants
Participants (1 male, 7 female) were monolingual speakers of English, had no phonetic training, reported normal hearing and visual acuity, and ranged in age from 22 to 40 years. All gave informed written consent in line with British Psychological Society guidelines. The study was granted ethical approval by the Psychology Research Ethics Committee of the University of Edinburgh.

2.2. Materials
The picture-name set was created by pairing the consonantal onsets /k/ and /t/ with 6 VC rimes (e.g., /k/ + /æp/ → CAP; /t/ + /æp/ → TAP). Each of the 12 words generated in this way was represented by a colour picture selected from an online database (online pre-test mean picture-name agreement = .76, range = .3 to 1). All picture names were concrete nouns of medium lexical frequency (mean log10CD = 2.93, SD = 0.41, range = 2.07 -3.91; SUBTLEX-US database, Brysbaert & New, 2009). For each picture name, 3 sentence stems were generated that strongly predicted the picture name as their final word (online pre-test minimum cloze probability > .8). The 36 sentence stems were designed each to end in a vowel or semi-vowel in order to allow audio to be cut at a
comparable and non-informative point across all stimuli. Sentence-stems were recorded as spoken by a native female speaker of British English, at a mean rate of 3.92 syllables per second (mean sentence stem duration = 3.10 seconds, range = 1.90 – 5.29 seconds).

2.3. Procedure
The experiment was run at the Ultrasound Tongue Imaging suite at Queen Margaret University. The full experiment was presented on a Dell XPS 1702 laptop using DMDX presentation software (Forster & Forster, 2003). The presentation software fully randomized item presentation within blocks. Participants were familiarized with the 12 picture names prior to the beginning of the experiment, in order to ensure that they would be able to correctly name pictures during the experimental phase. During the familiarization phase each picture was presented once in each of three blocks. All participants used target names for pictures 100% accurately by the third familiarization block.

Following the third familiarization phase participants were fitted with the ultrasound helmet (used to maintain probe position throughout the experimental procedure; Scobbie et al., 2008). Participants then named pictures once more as they had in the third familiarization block, in order to acquire experience of speaking whilst wearing the ultrasound device prior to commencing the experiment proper. The pictures were then presented for naming in 3 conditions: In the Control condition pictures were presented with no auditory context, following presentation of a fixation point. In the experimental conditions (Match and Mismatch) pictures were presented immediately following auditory presentation of a high-cloze sentence-stem: In the Match condition the picture-to-be-named matched the predicted (but missing) sentence cloze word (e.g., “Jimmy fixed the drip from the old leaky” … TAP); In the Mismatch condition the picture-to-be-named differed in onset phoneme from the predicted word (e.g., “On his head he wore the school” … TAP, where the predicted word would be “cap”).

Control-trial blocks were presented at the beginning and end of the experiment. Trials in the experimental conditions were presented between the two Control blocks. Each sentence-stem was presented once in the Match condition and once in the Mismatch condition (i.e. paired once with the picture it predicted and once with the rime-pair of that picture). Each picture was presented twice in the Control condition, three times in the Match condition and three times in the Mismatch condition. The experimental design was therefore fully within-participant and within-items. Whether a given sentence-stem was first heard in a match or a non-match context was balanced across participants.

2.4. Data Capture and Processing
Using AAA software (Scobbie & Wrench, 2008) we recorded acoustic and ultrasound data for each trial: Recording started at the onset of the sentence-stem stimulus and ended once the participant had named the picture. Ultrasound data was captured via an Ultrasoundix device used in conjunction with a headmounted micro-convex probe, with depth set at 80mm and angle at 150°, capturing a mid-sagittal tongue image at a rate of 100fps. Data was exported from AAA in AVI format at a rate of 30fps, following which an audio-video synchronization check was performed in VirtualDub (http://www.virtualdub.org/).

2.4.1. Audio data processing
We manually performed acoustic landmark labelling via visual inspection of the spectral signal in Audacity (http://audacity.sourceforge.net/). For each trial we identified the off-set of the sentence-stem audio, the acoustic onset of the picture name, the acoustic onset of the vowel, and the acoustic offset of the steady-state vowel. The time-points of each trial’s landmarks were recorded in .csv format, allowing this information to be made available to the ultrasound video-processing software.

2.4.2. Video data processing
Each frame of ultrasound video constituted a 512 x 277 grid of pixels. Pixels ranged in luminance from 000 (black) to 255 (white). In order to achieve data tractability, we processed each frame so that luminance was averaged over 8 x 8 contiguous pixels (see McMillan & Corley, 2010). A vector was generated from each frame, with each 8 x 8 pixel block assigned a specific position in the vector. Each vector ran from bottom left to top right of the AVI frame and each pixel block was recorded by its luminance (0 to 255). Vectors formed the basis for analyses, which were performed by calculating and comparing “Delta scores”; i.e., the Euclidean distances between individual vectors (frames).

3. Analysis
In order to minimise the effects of noise in the ultrasound images (see Scobbie & Wrench, 2008) we performed a preparatory analysis in order to determine the quality of the data acquired from each participant for each CV onset. This analysis was performed on ultrasound data acquired between the acoustic burst and the end of the steady-state vowel for each token; subsequent analyses were performed on data acquired prior to the acoustic release of the onset consonant. We used multidimensional scaling (Mardia, 1978) to calculate how well the Delta scores distinguished tokens of a given CVC word from tokens of all other CVC words produced by that participant: This was achieved by determining the mean Euclidean distance of a given vector (i.e. articulation) from: (i) all vectors representing different words; (ii) all vectors representing the same word. The Discrimination score for each onset for each participant was equal to (i)/(ii). Therefore the higher the score the better the data discriminated between a given CV onset and others in the picture-name set, and the less “noisy” the data. This information was used to geometrically weight the contribution of each participant’s data to subsequent analyses (Carroll & Ruppert, 1988). In this way we were able to avoid arbitrarily discarding “poor quality” data, whilst accounting for the great by-participant variability known to be associated with ultrasound articulatory data.

We used a linear mixed-modelling approach, implemented in R 3.0.2 via the lme4 package, version 0.999999-4 (Bates, Maechler, & Bolker, 2013; R Core Team, 2013). Data were weighted as described above. Condition (Match/Mismatch) and Onset Consonant (/k/ or /θ/) were included as fixed effects, and Participant and Picture-name as random effects. Because this approach provides estimated, rather than exact, effect sizes it was not appropriate to calculate associated p-values exactly. We therefore treat \(|t| > 2\) as indicating a statistically significant effect.
3.1. Location of articulation analysis
The first analysis investigated by-condition differences in data topography. Articulatory data acquired between -500 ms and 0 ms of the acoustic burst were collapsed to produce one average-luminance vector per token. This allowed each articulatory token to be compared to a reference vector for that item (picture-name). The reference vector for each item represented the mean of the vectors for that picture-name as produced in the Control condition. Vectors for all individual articulations in the Match and Mismatch conditions were compared to the relevant Control reference vector. This produced a Delta-score for each articulation (token), which indicated the distance in multi-dimensional space between that token and the participant’s mean Control articulation of the relevant picture-name.

The Delta-scores were then modelled as the outcome variable in a linear mixed effects model (as detailed above). Inspection of the model indicated that Delta scores in the Mismatch condition were significantly greater than those in the Match condition (β = 10.89, t = 2.15). This indicates that pre-acoustic articulation was less similar to the Control condition in the Mismatch condition than in the Match condition.

3.2. Time-course analysis

![Time-course Data: Frame-to-Frame Change in Ultrasound Tongue Image During Pre-acoustic Articulation](image)

**Figure 1:** Frame-to-frame change in ultrasound tongue image during pre-acoustic articulation. Faint lines indicate 95% confidence intervals. Circles at top of plot indicate inter-frame intervals. Circles are filled where change differs significantly by condition (Match v. Mismatch).

The second analysis investigated by-condition differences in the articulatory time-course of the pre-acoustic articulations. Articulatory data for each token constituted all ultrasound video frames recorded from 1000 ms prior to the acoustic burst until the acoustic burst (i.e., 31 frames per token). We calculated a Delta score for each inter-frame interval of each token: In this way each Delta score indicated the Euclidean distance between the current frame and that immediately preceding it. Higher Delta scores therefore indicated greater frame-to-frame change, associated with greater change in the configuration of the tongue as indicated by the ultrasound image. Each articulatory token was represented by a series of Delta scores indicating successive frame-to-frame change within that token’s data. This output was automatically averaged and plotted by-condition (Match v. Mismatch; see Fig. 1) and by onset-consonant (/k/ v/ /h/).

We investigated the time-course data by performing a mixed effect model analysis at each time-point (inter-frame interval). Inter-frame change (Delta) was treated as the outcome variable. In order to account for the increased risk of Type I errors associated with the use of multiple comparisons (for discussion see Lage-Castellanos et al., 2009) we treated effects as significant only when they were clustered across three or more consecutive inter-frame intervals. Effects of condition were found to be statistically significant (i.e., |t| > 2) at all intervals from -483 to -283 ms, and consistently indicated greater frame-to-frame movement in the Mismatch condition than in the Match condition.

4. Discussion and conclusion
We reported a study in which we adapted an automated UTI analysis technique in order to investigate the effect on speech production of prediction during speech comprehension. Participants named pictures in a control condition and in two experimental conditions. The experimental conditions differed only in whether the picture name Matched or Mismatched the predicted word. Predictions were elicited via presentation of a high-cloze auditory sentence stem (i.e., via spoken language comprehension).

We applied two analysis approaches in order to investigate lingual motor-activity in the period immediately prior to the onset of acoustic information associated with picture-naming. The first approach collapsed information about the location of the tongue across time, and compared both experimental conditions to the control condition. The second approach provided information about the degree of movement observable at each inter-frame interval, and compared the two experimental conditions directly. Both approaches revealed by-condition differences in speech-motor activity, indicating that prediction during speech-comprehension produces both spatially and temporally observable effects on motor-speech output. Productions in the Mismatch condition appear to be less “canonical” than those in the Match condition (i.e., differed more from Control productions than did productions in the Match condition).

The current study demonstrates that a Delta-approach to ultrasound tongue image analysis can be adapted to be applicable beyond the paradigm for which it was initially developed (McMillan & Corley, 2010). The automated nature of the approach makes it appropriate for use in psycholinguistically-motivated studies because it reduces demands on researcher time and expertise (compared to a typical tongue-tracing approach), and allows meaningful averaging across differing items. Data-quality weighting provides a non-arbitrary approach to handling between-participant differences in noise-signal ratios, thereby extending the proportion of useable data.

The findings of the current study are novel in that they demonstrate online adaptations to motor-speech realizations arising of prediction during comprehension: Prediction-elicited representations cascaded to directly affect speech-motor production itself, rather than simply affecting the time-point or moment at which motor-execution began. Further investigation will be required to determine more exactly the nature of the information that cascades to a speech motor-execution level: If the predicted onset item itself were
activated at a motor-execution level we might expect to find that tokens produced in the Mismatch condition were more similar to their rime-partner than were tokens produced in the Match condition (e.g., articulation of TAP in the Mismatch condition would be more similar to articulation of CAP in the Control condition than would articulation of TAP in the Match condition). We did not find this to be the case for the time-frame analyzed in this study, although just such an effect has been demonstrated in tongue-twister data when applying the Delta-technique (McMillan & Corley, 2010).

It should be noted that in our analyses all data was time-locked to the acoustic onset of speech production. We adopted this approach in order to avoid finding by-condition differences simply as a function of when articulation commenced. That situation might occur under a stimulus-locked approach if motor-execution were identical across conditions but commenced later (i.e., longer after stimulus presentation) in the Mismatch condition. However, we agree with a reviewer who commented that, when exploring effects of prediction on articulation, it would be valuable to study speech-motor behaviour at the point of stimulus presentation (i.e., time-locked to picture presentation, in the case of the current study). This is an area for further development of the Delta-technique, and a spatial analysis of data time-locked to stimulus presentation might well provide valuable information regarding the exact nature of the prediction-effect demonstrated in the current study.

For the purposes of the current paper we note that although we do not report stimulus-locked data, the time-frame used in the second analysis approach includes data acquired at the point of stimulus presentation, and indeed extends further back in time to include articulatory data acquired whilst listening to the auditory material. Differences in speech-motor activity become observable as a consequence of whether or not a comprehension-elicited prediction is met. Given the nature of our stimuli, this confirms that: (i) listeners produce predictions during comprehension of spoken language presented at a typical conversational speech-rate, and; (ii) the effect of such predictions is observable in the listeners’ own speech productions.

5. Acknowledgements

We thank Professor Alan Wrench (Articulate Instruments) and Steve Cowen (QMU) for technical advice and assistance. We would also like to thank two anonymous reviewers.

6. References


Calculating articulatory syllable duration and prosodic boundaries
Donna Erickson1, 3, Shigeto Kawahara2, Jeff Moore3, Caroline Menezes4, Atsu Suemitsu5, Jangwon Kim6, Yoshiho Shibuya1

1 Kanazawa Medical University, Japan
2 The Keio Institute of Language and Cultural Studies, Japan, 3 Sophia University, Japan, 4 University of Toledo, USA, 5 Japan Advanced Institute of Science and Technology, Japan, 6 University of Southern California, USA.
EricksonDonna2000@gmail.com, kawahara@icl.keio.ac.jp, jeffmoore.personal@gmail.com, caroline.menezes@utoledo.edu, sue@jaist.ac.jp, jangwon@usc.edu, yshibuya@kanazawa-med.ac.jp

Abstract
Articulatory duration of vowels is often measured as the duration from jaw closure maximum to jaw closure maximum, flanking the production of the syllable nucleus. However, this method may not necessarily represent articulatory syllable duration, since the actual onset/offset of the syllable depends on the specific articulator (lip, tongue) that implements the articulation of the syllable onset/coda, and which does not strictly synchronize with timing of jaw movements nor with acoustically measured syllable durations (e.g., [Menezes, 2004]). We propose that by using a different approach, that suggested by the C/D model [Fujimura, 2000], it is possible to compute quantitatively the time values of prosodic boundaries from articulatory dynamics data. The algorithmic output, the “syllable pulse train”, is the phonetic realization of the utterance’s rhythmic structure (e.g., [Bonaventura and Fujimura, 2007]), which in turn reflects the phonologically derived metrical structure of the utterance (e.g., [Erickson et al., 2012]). Our small study presented here using the C/D model indeed revealed systematic articulatory patterns across speakers.

Keywords: C/D model, articulatory duration, prosodic boundaries, metrical structure

1. Introduction
Acoustic syllable duration is traditionally measured as the temporal interval between variations in the acoustic waveform, e.g., [1]. However, it is possible that the articulation of the syllable may not be directly inferable from the acoustic output, e.g., [1, 2]. The articulatory duration of vowels can be measured as the duration from maximum jaw closure to maximum jaw closure, as the speaker articulates the nucleus of the syllable, e.g., [1, 3]. This approach may not always represent articulatory syllable duration accurately, since the onset/offset of the syllable depends on the specific articulator (lip, tongue) that implements the articulation of the syllable onset/coda. These movements do not strictly synchronize with the timing of jaw movements nor with acoustically identified syllable boundaries, e.g., [2].

We propose a different approach for defining articulatory syllable duration, based on the C/D model proposed by Fujimura [4]. Essentially, the model was proposed to account for the mismatch between acoustics and articulation, and was thus accordingly named, the Converter/Distributor model, in that it converts/distributes the phonological information to acoustic signals. The model is innovative in recognizing syllables of varying magnitudes as the basic structure of an utterance. Syllable magnitudes are calculated based on jaw displacement for each syllable. A relatively prominent syllable has a larger jaw displacement than a relatively less prominent syllable. In the C/D model, an utterance is made up of a train of pulses that vary in height based on the syllables’ magnitudes, and the syllable pulse height (syllable stress level) is commensurate with the articulatory syllable duration. This is interesting because recent studies show no consistent relationship between the syllable acoustic duration and stress level [2, 6]. Ongoing work by Erickson and colleagues suggests that the pattern of syllable pulse magnitudes within an utterance corresponds to the rhythm of the utterance, which in turn reflects the phonological metrical structure of the utterance, e.g., [7].

While the magnitude of the syllable is determined by maximum jaw displacement, the consonantal gestures of the syllable influence the timing of the syllable pulse. According to the C/D model, the timing of the syllable pulse is centered relative to the speed (maximum and minimum velocity) of the crucial articulators of the onset and coda gestures. Crucial articulator (CA) refers to that articulator (tongue tip, tongue blade, tongue dorsum, lip) that articulates the onsets and coda of the syllable. For example, the CA for [n] is tongue tip, for [p] is lower lip, and for [k] is tongue dorsum. Based on observation of an “iceberg” point (point with smallest mean invariance) in the overlaid demisyllabic velocity time function, the center of the syllable is defined as the midpoint between the syllable onset “iceberg” to the syllable coda “iceberg” [4, 8, 9]. However, in this paper we calculate the center of the syllable as the midpoint between the maximum speed of the crucial articulators following [10], described in more detail in Methods.

The C/D model describes not just syllable strength patterns, but also phrasing patterns [4]. If we know (i) the syllable strength (from the amount of jaw displacement), represented as various-sized syllable pulses, and (ii) the timing of the syllable pulse as described above, then we can calculate the articulatory duration of the syllable as well as the prosodic boundaries of the utterance.

In this paper, we use the C/D model to calculate prosodic boundaries and utterance rhythm from articulatory recordings of an English phrase, as spoken by four North American

1 The intonation pattern (or melody) of an utterance is an aspect of utterance prosody, produced by laryngeal changes, described in part by its F0 contour, and is handled in the C/D model in a separate tier (see e.g., [5]). Therefore it is not addressed per se in this paper.
2 There may be certain articulatory gestures, as in an interdental consonant, that prevent complete closure. This issue needs to be examined, but it is outside the scope of this paper.
English speakers. Previous studies have calculated prosodic boundaries from syllable triangles to examine a phrase with three monosyllabic digits [2, 9]. The current work uses the syllable triangle algorithm to examine the metrical structure of a four-word phrase: nine tight night pipes. The hypothesis is that by using the C/D model, systematic prosodic boundaries will be revealed across speakers, and these boundaries will be the same as those predicted by metrical theory. Moreover, their location (and possibly size) will match perception by listeners.

2. Method

2.1. Articulatory recordings and analysis

Acoustic and articulatory recordings were done using 3-D EMA (Carstens AG500 Electromagnetic Articulograph) at the Japan Advanced Institute of Science and Technology (JAIST), and at Haskins Laboratories, New Haven, Connecticut. One sensor was placed on the lower medial incisors to track jaw motion, one sensor each was placed on the tip of the tongue (TT), the mid of the tongue (TB) and the back of the tongue (TD). A sensor was placed on the lower lip (LL) and on the upper lip (UL). Four additional sensors (upper incisors, bridge of the nose, left and right mastoid processes behind the ears) were used as references to correct for head movement. The articulatory and acoustic data were digitized at sampling rates of 200 Hz and 16kHz, respectively. The occlusal plane was estimated using a biteplate with three additional sensors. In post processing, the articulatory data were rotated to the occlusal plane and corrected for head movement using the reference sensors after low-pass filtering at 20 Hz. The lowest vertical position (maximum displacement) of the jaw with respect to the bite plane was located for each target syllable of the utterance using the MATLAB-based custom software mview (Haskins Laboratories); this measure was used to indicate the height of each syllable pulse in the utterance. The position of the syllable pulse in the syllable was set at the midpoint between the maximum speed of the crucial articulator of the syllable onset and that of the syllable offset, also determined by a function in mview. The speakers were two male and two female North American English speakers. The utterance examined was, Yes, I saw nine tight night pipes in the sky tonight, adapted from [7,11]. Analysis is done for the phrase nine tight night pipes, which contains closed syllables with [ai] vowels. Before the data collection, the speakers had a chance to look at a picture illustrating the sentence, and could practice the sentence until they felt comfortable with it. The utterance was part of a larger corpus, presented to the speakers in randomized order, with five repetitions. The second or third utterance of each speaker is analyzed in this paper. The crucial articulators for the target syllables in this sentence are tongue tip (nine, tight, night), and lower lip (pipe).

Figure 2 shows jaw displacement tracings for the phrase nine tight night pipes, the measurements of which are the syllable pulse height. It also shows displacement and velocity tracings for each of the Crucial Articulators.

Figure 2 shows the same utterance; the arrows mark the maximum speed of the Crucial Articulators of the syllables. Notice these are different from the time of maximum jaw displacement.
2.2. Metrical organization of the phrase

The metrical organization postulated for the four-word phrase is one major phrase (nine tight night pipes), and two recursive minor phrases (tight/night pipes) [12,13], as indicated by the smooth brackets in Figure 5. According to metrical theory, one word in each phrase receives the largest stress; the major phrase stress can fall on either one of the minor phrase stresses. The minor phrase stress for this phrase will be on nine and on pipes; the major phrase stress will be on one and only one of these, i.e., nine or pipes, depending on the utterance conditions of the speaker. Here we suggest that pipes has the major phrase stress (based on work reported in [11]), and nine, the minor phrase stress. Numerical values of stress are arrived at by counting up the number of metrical grids (x’s): the stress pattern hypothesized for this phrase is ((3)(1)(1 4)), where smooth brackets indicate phrasing, and, also shown in Figure 5. We assume that a boundary of a higher level is more likely to be realized, and will be longer in duration. Thus, the largest boundary will be after pipes, then after night, and then after tight. No boundary will appear in the two-word phrase, night pipes.

2.3. Perception tests

A small-scale perception test was done with fifteen American college listeners for the four phrases analyzed in this study. Participants listened to the phrases and were instructed to mark phrasal structure, putting commas where small gaps/groupings are heard, following the methodology used by [14]. Before the actual testing, they were given a sample sentence, and asked to write commas where they heard a gap. They heard each utterance 8 times, in non-randomized order.

3. Results and discussion

3.1. Syllable pulse height and articulatory syllable duration

<table>
<thead>
<tr>
<th>Subj</th>
<th>Syllable magnitude</th>
<th>Articulatory syllable duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>A02</td>
<td>24.2 22.5 23.3 24.2</td>
<td>245.8 22.8 238.8 245.8</td>
</tr>
<tr>
<td>A03</td>
<td>23.7 23.3 22.7 23.9</td>
<td>266.5 262 255.3 268.7</td>
</tr>
<tr>
<td>A09</td>
<td>45.8 45.8 46.8 46.7</td>
<td>271.8 271.8 277.7 277.1</td>
</tr>
<tr>
<td>A10</td>
<td>23.1 21.6 22.4 23.5</td>
<td>285.5 296.3 310.8</td>
</tr>
</tbody>
</table>

Notice the positive relationship between syllable magnitude (syllable pulse height) and articulatory syllable duration. For instance, the syllable magnitude for A02 for both nine and pipes is 24.2 mm, the articulatory syllable duration for both is 245.8 ms. A positive relationship between syllable magnitude and boundary duration is interesting, especially in view of the finding by [6] of no significant correlation between acoustic syllable duration and jaw displacement. Moreover, using articulatory information as a means for determining syllable duration has advantages since sometimes it is impossible to measure duration from acoustic signals, as for instance, one does not know exactly where [t] for night ends and the initial [p] for pipes begins. Also notice that three of the four speakers show the largest amount of jaw opening for pipes. In the next section are shown the syllable triangles, followed with a discussion of the metrical structure and phrasing.

3.2. Syllable Triangles

Figures 6–9 show the results of applying the algorithmically-objective method for deriving articulatory syllable duration and prosodic boundaries, for the phrase nine tight night pipes, as spoken by four North American speakers. In the figures below, the height of each triangle represents the amount of jaw displacement/the amount of syllable magnitude (stress), the base of each triangle is the articulatory syllable duration, and the spaces between each triangle show the prosodic boundaries.
3.3. Syllable triangles and metrical organization
In terms of syllable triangle height, the figures (also see Table 1) show that for three of the four speakers, *pipes* has the largest amount of jaw displacement, followed by *nine*. This matches the metrical structure depicted in Figure 5.

In terms of prosodic boundaries, as was predicted by Figure 5, we see for all speakers, two boundaries, one after *nine* and one after *tight*, but none between *night* and *pipes*. Moreover, as predicted by Figure 5, three of the four speakers show a bigger boundary after *nine* than after *tight*, as shown by the durations in Table 2 (left-most data columns).

Table 2. Prosodic boundary durations (left-most 2 data columns) and perceived boundaries (right-most 2 columns). The numbers in bold type indicate the larger of the two boundaries.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Measured boundary after <em>nine</em></th>
<th>Measured boundary after <em>tight</em></th>
<th>Perceived boundary after <em>nine</em></th>
<th>Perceived boundary after <em>tight</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>A02</td>
<td>99.3 ms</td>
<td>81.5 ms</td>
<td>53%</td>
<td>47%</td>
</tr>
<tr>
<td>A03</td>
<td>55.2 ms</td>
<td>65.8 ms</td>
<td>57%</td>
<td>43%</td>
</tr>
<tr>
<td>A09</td>
<td>19 ms</td>
<td>12 ms</td>
<td>55%</td>
<td>41%</td>
</tr>
<tr>
<td>A10</td>
<td>80.4 ms</td>
<td>10.6 ms</td>
<td>39%</td>
<td>61%</td>
</tr>
</tbody>
</table>

3.3. Comparison of syllable triangle boundaries with listeners’ perceptions
Table 2 also shows the perceived boundary strengths, calculated in percentages as the number of commas placed at a given juncture compared to the total number of commas for a given phrase. For Speaker A09, the total percentage does not add up to 100%, because some listeners put commas after the final word, *pipes*, even though that was not the assigned task. The results of the very preliminary tests showed that listeners perception of location of boundaries agreed with the algorithmically calculated boundaries, that is, they heard breaks after *nine* and *tight*, and not between *night* and *pipes*. However, the strength of listeners’ perceptions matched the numerical calculated boundaries for only two of the speakers. Perhaps this is because the perceptual task of evaluating pauses was a difficult one. Listeners often varied their placement of commas when presented with the same utterance eight times in a row even though they knew it was the same utterance. There are a number of reasons for the difficulty in perceptually evaluating strength of prosodic boundaries, including the fact that notation of boundary strengths of non-major phrases is not part of our writing system. Or it may be that a larger number of listeners are needed to elucidate minor prosodic boundaries, as was done in [15]. Listeners’ perception of boundary strengths is part of our on-going investigation into the efficacy of the C/D model for understanding utterance prosody.

4. Conclusion
This paper is a preliminary study to apply the C/D model to calculate articulatory syllable duration and prosodic boundaries. The results of this small study suggest that analyzing articulatory data within the framework of the C/D model can provide prosodic information about stress and boundaries. Moreover, this approach is algorithmically very objective, and as such, produces replicable results. It is hoped that the type of information generated by calculating articulatory syllable duration and prosodic boundaries, using the C/D model, will lead to future, more comprehensive and objective studies of metrical structure and phrasing.

5. Acknowledgements
The impetus for this approach to the phonology-phonetics interface of language and its metrical structure comes from the insights of Osamu Fujimura. This work was supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C)#22520412 and (C)#25370444. A special acknowledgement to Mark Tiede for help with mview, to Bryan Pardo for UBEDIT, and to Mark and to Ian Wilson, for participating in the experiment. Also, we thank Sungbok Lee for discussions about this paper.

6. References
Abstract

Lule Saami presents a phonological three-way consonantal quantity system - a typologically rare phenomenon. This study examines the acoustic implementation of this contrast based on acoustic data from 7 native speakers, recorded in Divtasvuodna (Tysfjord) in Norway. Results show that the difference between three degrees of phonological quantity is primarily manifested through a substantial and systematic consonant duration increase from one quantity to another. There is furthermore a slight readjustment on adjacent vowel durations, particularly on the vowel following the target consonant. In addition to these temporal cues, other correlates are involved in the implementation of this contrast, including the intensity of the adjacent vowel. Implications of the phonetic results are discussed, with particular attention to the issue of (super)geminate representation within moraic theory.

Keywords: Saami language, Lule Saami, phonological quantity, consonant length, experimental phonetics

1. Introduction

Lule Saami, a severely endangered, indigenous language spoken by some 650 speakers (Mørën-Duolljá 2013) in northern Norway and Sweden, presents a phonological three-way consonantal quantity system. Cross-linguistically rare, this ternary system encodes both lexical and morphological contrast. Lule Saami uses consonant gradation (i.e. morpheme-edge modifications of consonant quantity and/or quality) to express lexical and grammatical distinctions. Word-medial non-cluster consonants can occur with three different degrees of length. These are sometimes referred to as quantity (Q) 1, 2 and 3, or as singletons, geminates and supergeminates. We will transcribe these consonants here as [C], [C:], and [C:C], respectively. The Lule Saami phonological system has ternary quantity opposition for most consonants but only unary quantity for vowels (short). Certain vowel sequences are allowed in this language, including identical non-high vowels, but most of these are assimilated non-identical adjacent vowels.

Lule Saami has lexical three-way contrast as shown in example (1). In this example, the words in quantity 2; “reason” and “squirrels” are homophones.

\[
\begin{align*}
\text{Quantity 1} & \quad \text{Quantity 2} & \quad \text{Quantity 3} \\
\text{oare [əoarːe]} & \quad \text{oare [əoarːe]} & \quad \text{oar’e [t̚əarːe]} \\
\text{“reasons” “reason” OR “squirrels” “squirrel”}
\end{align*}
\]

The language also uses ternary quantity morphologically as demonstrated in example (2).

\[
\begin{align*}
\text{Quantity 1} & \quad \text{Quantity 2} & \quad \text{Quantity 3} \\
\text{(dån) baså [pasoa]} & \quad \text{(sán) basså [pas:aа]} & \quad \text{(sán) bas’så [pas:oа]} \\
\text{“(you) clean” “(he/she) cleans” “(he/she) starts to clean”}
\end{align*}
\]

Systems with three degrees of quantity are controversial within the field of phonology, as well as in phonetics. From a phonological viewpoint, many authors argue that quantity oppositions are maximally binary (e.g. Chomsky and Halle 1986, Odden 1997). McCarthy and Prince (1986) posit that the syllable is maximally bimoraic and thus reject the existence of supergeminates. From a phonetic perspective, the debate concerns the phonetic implementation of the contrast; to establish whether a three-way distinction can be purely durational, or whether other, qualitative correlates must be involved.

1.1. Quantity in Finno-Ugric languages

An existing acoustic study by Engstrand (1987), which investigated phonetic correlates of phonological quantity in Lule Saami, showed that there were significant consonant durational differences across three quantities for the consonant [n]. These results are, however, based on a limited corpus. Only one speaker participated in the experiment testing durational differences on one consonant only. A new acoustic investigation of Lule Saami quantity is clearly called for.

Other Finno-Ugric languages such as Inari Saami, North Saami and Estonian also display phonological three-way length contrast in consonants and/or in vowels. In Estonian, consonants and vowels can be short, long or overlong, the ternary difference is phonetically implemented via an interplay of prosodic parameters (i.e. stress, pitch and duration) and is dependent on word-structure (Lehiste 1997, Ehala 2003, Spahr 2012). Perception studies show that the ratio between the first and the second syllable of a dissyllabic sequence seems to be the most important indicator of quantity in Estonian. Listeners were able to discern Q1 from Q2 when having access to only the first syllable of a dissyllabic sequence, but both syllables were needed in order to perceive a difference between Q2 and Q3 (Eek and Meister 1997). The temporal relation between the two syllables in a dissyllabic sequence has been shown to have ratios of 2:3 for Q1, 3:2 for Q2, and 2:1 for Q3 (Lehiste 1965). Another important clue for discriminating quantity in Estonian is fundamental frequency. Words have a F0 peak towards the end of the first syllable in Q2, whereas the peak is close to the beginning of the first syllable in Q3 (Lehiste 1997). Lippus et al. (2007) found that for native Estonian speakers, tonal cues were important for the perception of Q1/Q2/Q3 distinction in the case of vowel quantity, whereas quantity in voiceless stops was distinguished based on temporal cues alone.

In Inari Saami, duration differences among three consonant lengths are statistically significant and correlate negatively with V1 duration (Bye et al. 2010). In Northern Saami, a ternary consonant duration contrast is claimed to be a result of an underlying binary contrast surfacing as a three-way surface distinction due to mora-mapping rules (Bals et al. 2007; 2012). The primary purpose of this study is to determine the way the three-way quantity opposition is acoustically implemented in Lule Saami. Specifically, we wanted to examine whether phonological quantity is associated with other cues other than the temporal ones, as suggested by earlier evidence for some Finno-Ugric languages and other unrelated languages.
displaying singleton/geminate contrast (Ridouane 2010). Implications of the results for the general issue of geminate representation are discussed, with particular attention to the way moraic theory may account for the ternary system of Saami quantity.

2. Methods

We recorded seven native speakers of Lule Saami in the Lule Saami community in Divtasvuodna (Tysfjord), Norway. The participants were four men and three women aged 26 to 81 (the mean age was 49.9 years). All of the speakers are multilingual and speak at least Norwegian in addition to Lule Saami. The corpus consisted of six verbs in three different forms (i.e. six minimal triples) that allowed us to measure alternations across the three quantities (Q1 vs. Q2 vs. Q3) in an identical (C)V.CV context. The consonants tested were [l], [m], [n], [r], [s] and [v]. The target consonant is always preceded by a short vowel and followed by a sequence of identical vowels. [a] and [au] were used for all consonants except [r] where no verb containing these vowels in the same positions could be found, and thus [a] and [ɔ] was used instead. Speakers produced conjugated verb forms from a given infinitive form, repeating a total of 18 word tokens five times per token. Table 1 presents the list of word items used in this study:

<table>
<thead>
<tr>
<th>Infinitive word form</th>
<th>Q1 2p.SG</th>
<th>Q2 3p.SG</th>
<th>Q3 Incept. 3p.SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>“ballat” (to fear)</td>
<td>[palæ]</td>
<td>[pal:æ]</td>
<td>[pal:la]</td>
</tr>
<tr>
<td>“njammat” (to suckle)</td>
<td>[nam:oa]</td>
<td>[nam:oa]</td>
<td>[nam:moa]</td>
</tr>
<tr>
<td>“mannat” (to leave)</td>
<td>[man:oa]</td>
<td>[man:oa]</td>
<td>[man:na]</td>
</tr>
<tr>
<td>“bår threatened)</td>
<td>[pær:ɔ]</td>
<td>[pær:ɔ]</td>
<td>[pær:ro]</td>
</tr>
<tr>
<td>“tjavvat” (to leave)</td>
<td>[fæv:oa]</td>
<td>[fæv:oa]</td>
<td>[fæv:oa]</td>
</tr>
<tr>
<td>“bassat” (to clean)</td>
<td>[pas:oa]</td>
<td>[pas:oa]</td>
<td>[pas:oa]</td>
</tr>
</tbody>
</table>

The target words were embedded in a carrier sentence designed to limit intonation and coarticulation effects. The target word was preceded by “dån” (you SING) or “sån”(he/she), and followed by “guokt gálmmdå”, /kuokta kəlmmə/, (two three). Each sentence was repeated four times, yielding a corpus consisting of 504 tokens (18 word items * 7 speakers * 4 repetitions). The acoustic data were segmented, annotated and analyzed using Praat (Boersma & Weenick 2012), and both temporal and non-temporal acoustic variables were analyzed. The non-temporal values extracted from the post word-initial onset material of each token were fundamental frequency, f1, f2, f3 (all in Hz) and intensity (dB). We extracted mean values from the first and last 30 ms of each segment and absolute values from the middle of each segment. Segment durations were measured in milliseconds.

3. Results

Among the speakers that participated in the study, all produced the contrast in three or more of the minimal triplets that we tested. Three speakers produced the contrast for all of the minimal triplets that they pronounced. We consider that the three-way opposition is categorical; it either exists for a given speaker and a given word, or the speaker does not have the contrast for that word. Based on this consideration, the word/speaker combinations where the contrast was not produced have been excluded from the analysis of durational results to avoid skewing the mean values. Two of the words were unknown to one speaker and two speakers had incomplete data for one word. The analyzed data subset thus includes the data from 28 of 38 possible speaker/word combinations.

3.1. Temporal variables

The Lule Saami three-way contrast is phonetically implemented through significant durational differences among singletons, geminates and supergeminates for all the consonants tested (using a two-way ANOVA, correlated samples: p < .0001). The average duration of a singleton (Q1) was 90 ms, a geminate (Q2) 215 ms and a supergeminate (Q3) 323 ms. There were also considerable differences among individual consonants. The two-way ANOVA also showed that consonant label had a significant effect on consonant duration (p < .0001). The largest durational contrast was found for [r] in the word bårat, where the mean duration of the target consonant was 69 ms in Q1, 223 ms in Q2 and 323 ms in Q3; an increase in duration by 221 % for the geminate and 366 % for the supergeminate compared to the singleton. The huge leap in duration is due to allophonic realizations of [r]. In Q1 it surfaces as a flap, which explains why it is only half as long as the short vowel preceding it. In Q2 and Q3, [r] is realized as a trill, allowing it to reach full geminate and supergeminate duration. The largest durational contrast between a Q2 and a Q3 was found for [s] in the word tjavvat; the consonant duration was 204 ms in Q2 and 319 ms in Q3, which equals an increase of 56%. The smallest Q2/Q3 durational contrast was found for [s] in the word bassat. Here, the duration only increased by 40%, from 253 ms in Q2 to 353 ms in Q3. Figure 1 illustrates the results and presents the mean durations for V1, C, and V2 for the three phonological quantities.

Figure 1: Mean durations of segments (V1, C, V2) in post word-initial onset material for six words and seven speakers in three quantities (Q3, Q2, Q1). Error bars represent standard deviation.

3.1.1. Vowel readjustment

We observed a slight readjustment effect on adjacent vowels. Within the words tested in this study, the preceding short vowel (V1) had an average duration of 109 ms in Q1, 104 ms in Q2 and 90 ms in Q3. The long vowel (V2) following the target consonant had an average duration of 286 ms in Q1, 250 ms in Q2 and 242 ms in Q3. The largest overall effect was on the vowel following the target consonant (V2), as can be seen in figures 1 and 2. Between Q1 and Q2, the target consonant increases on average by 148%, V1 simultaneously decreases on average by 5% while V2 decreases by 15 %. Between Q2 and Q3, the target consonant increases on average by 109 %,
V1 decreases on average by 15% whereas V2 decreases by 5%.

The effect of adjacent target consonant quantity was significant in the case of both V1 (p < .0001) and V2 (p < .0001). Tukey HSD post hoc tests showed that in each combination of quantities and vowels duration differences were significant, except for between Q2 and Q3 for V2 (p = .545).

The average duration of the total target VCV sequence was systematically longer from Q1 to Q2 and from Q2 to Q3, with an average total duration of 484 ms in Q1, 565 ms in Q2 and 651 ms in Q3. The relations between segments in the V1,C,V2 sequence for the three quantities is illustrated in figure 2 below:

![Figure 2: Relative mean durations of segments (V1, C, V2) in post word-initial onset material for six words and seven speakers in three quantities (Q3, Q2, Q1).](image)

It seems that, similarly to what Ham (1998) found for monosyllabic words in Hungarian where syllable duration was correlated with the number of morae in the syllable, in our case the post word-initial material of a disyllabic sequence; the number of morae within the sequence correlates with the sequence duration. As shown in the phonological analysis below, positing that the first syllable contains one mora in Q1, two morae in Q2 and three morae in Q3 would account for these data.

### 3.2. Non-temporal variables

#### 3.2.1. Fundamental frequency

Several speakers produced a high-low intonational pattern within the disyllabic sequence, with an F0 peak at the left-edge of the target consonant in all three quantities. For certain speakers, the peak was somewhat higher in Q2 and Q3 than in Q1. This F0 difference was not significant (p > .05). Based on our observations, we cannot say that there is a tonal component in Lule Saami that contributes to co-signaling quantitative distinctions. This is unlike the findings for Estonian (Lippus et al. 2007).

#### 3.2.2. Intensity

Quantity had a significant effect on adjacent vowel intensity (V1: p = .012, V2 p = .02) but not consonant intensity (p = .18). V1 intensity increased with quantity whereas V2 intensity decreased. The intensity of V1 was highest in Q3 and lowest in Q1. For V2, average intensity was highest in Q1 and it decreased to Q2 and further to Q3. For V2, intensity seems to be correlated with duration; the shorter the vowel, the lower the intensity.

#### 3.2.3. Adjacent vowel quality

The formant structure of the vowels preceding and following the target consonant did not change from one quantity to another. We did observe a qualitative difference between V1 and V2 that is related to the quantitative difference between these vowels. V1, which is short, is more closed and more anterior than the vowel sequence in V2, which is more open and more posterior. These differences do not, however, change according to the quantity of the target consonant. There were also coarticulation effects on the vowels depending on the surrounding consonantal environment. Formant values differ considerably between words but not between quantities.

### 4. Phonological representation

The phonological representation of supergeminates has been the subject of much debate in the literature. Some versions of Moraic theory (Hayes 1986; 1989) represent geminates and supergeminates as monosegmental consonants linked to one or two morae, respectively. If mora-sharing is not allowed, then supergeminates are problematic for the principle of Maximal Binarity (McCarthy & Prince 1986), which claims an upper limit of two morae per syllable. However, trimoraic syllables have been claimed to exist in several languages, including Hungarian, Estonian, Hindi, Persian and Kashmiri (Ham 1998, Hayes 1989, Morén 2000).

We present here a preliminary autosegmental analysis (Goldsmith 1976) using moraic theory (Hyman 1985; McCarthy and Prince 1986; Hayes 1989) that accounts for the phonetic data analyzed in this study. Several works argue that phonetic duration is organized around the mora (Port et al. 1987, Hubbard 1995, Broselow et al. 1997, Ham 1998). Our Lule Saami data showed that there were significant durational differences between the target consonant segments in Q1 (90 ms), Q2 (215 ms) and Q3 (323 ms). Assuming a tight relationship between phonetic and phonological representations (e.g. Pierrerehubert 1990, Broselow et al. 1997), and assuming that this closeness should be reflected in linguistic theory, the durational characteristics of Lule Saami geminates and supergeminates could be captured by treating these segments as heading one mora (for geminates) and two morae (for supergeminates). Adding morae from Q1 to Q2 and from Q2 to Q3 would be directly reflected in the extra duration displayed by these segments. Hence, moraic structure would be directly manifested in the phonetic duration of Lule Saami geminates and supergeminates. One could posit that the first syllable of the disyllabic sequence contains one mora in Q1, two morae in Q2 and three morae in Q3. The second syllable contains two morae in all three quantities due to the vowel sequence. Assuming, along with moraic theory that geminates are moraic, the three degrees of quantity duration observed in Lule Saami can be represented as shown in (3):

a. Singleton  b. Geminate  c. Supergeminate  (3)

The structure in (1a) is the representation of a single consonant. The consonant links directly to the following syllable node and constitutes a syllable onset. The representation in (1b) is that of a geminate. The consonant is moraic and anambisyllabic - i.e. it is a coda, as well as the onset
of a following syllable. The structure in (1c.) represents a supergeminate. It is bimoraic and ambisyllabic. In our data, a slight readjustment effect was found for the duration of adjacent vowels. The effect was greatest in the following vowels (V2). If syllable duration were not correlated with the number of morae within the syllable, i.e. that the syllable maintains a constant temporal frame regardless of the number of morae it contains, we would expect a larger readjustment effect (decrease in duration) on the first vowel to compensate for the increasing duration of the consonant. For future studies, it would be important to compare dissyllabic sequences where the first syllable contains vowel sequences in order to examine what differences (if any) that might cause on the readjustment effect.

5. Conclusion

Robust durational differences were found among the three quantities for all the consonants tested. We propose that although there were also slight non-temporal differences in F0 and intensity, these are phonetic enhancements and do not play a decisive role in signaling the ternary quantity contrast in Lule Saami. While some speakers afforded extra prominence to the test words in Q3, this added emphasis is not phonologically significant. A planned perceptual study will test the role of non-temporal acoustic cues in native speakers’ discrimination of consonantal quantity. While we also observed significant durational readjustments on vowels, those effects were miniscule in comparison with the durational differences among the consonants. Thus, we take them to be phonetic in nature.

A moraic analysis that adds a mora from Q1 to Q2 and from Q2 to Q3 fits nicely with the contrasts documented in our data. Assuming this analysis entails the presence of bimoraic consonants and trimoraic syllables in Q3. It is important to note, however, that this analysis fails to account for non-durational alternations also present in the Lule Saami consonant gradation system - including pre-stopping, spirantization and changes in laryngeal features. A full empirical investigation is necessary, including systematic collection of Lule Saami phonetic and phonological data, before one can propose a full-fledged analysis that will account for all Lule Saami alternations at work in the its consonant gradation system.

6. Acknowledgements

This study has been carried out in collaboration with the Mávståslj: Spoken Lule Saami Documentation Project that is funded by the Norwegian Research Council (Forskningsrådet). It is also part of the program Investissements d’Avenir, overseen by the French National Research Agency, ANR-10-LABX-0083, (Labex EFL).

7. References


A Spatiotemporal Study of Speech in Patients with Recurrent Nerve Paralysis after Thyroid Surgery. Focusing on VOT
Camille Fauth¹, Béatrice Vaxelaire², Jean-François Rodier³, Pierre-Philippe Volkmar³, Rudolph Sock⁴, François Rodier

¹ LORIA CNRS UMR 7503 – Nancy France
³ Centre Paul Strauss – Strasbourg – France, Département de Chirurgie Oncologique
⁴ Université Pavla Jozefa Safarikova, Faculté des Lettres Košice - Slovaquie

camille.fauth@gmail.com, vaxelair@unistra.fr

Abstract

This study analyses the temporal characteristics of the voice of patients with laryngeal post-thyroidecmy surgery immobility, using the VOT spatiotemporal cue. We examine the consequences of laryngeal paralysis on oro-laryngeal timing of 7 patients, recorded during three postoperative phases. Their voices are compared to those of 7 control speakers. Our corpus consists of twelve VCV sequences. The approach, an articulatory-to-acoustic one, allowed us to observe, in the acoustic speech signal, different cues that may allow tracing articulatory timing patterns and configurations. A statistically significant change in the duration of VOT appears only for voiced consonants, in the early postoperative phase. Time and voice therapy have a positive impact, as patients' values in late recording phases are close to those obtained for control speakers.

Keywords: speech production, speech disorder, clinical phonetics, thyroid gland, laryngeal immobility, perturbations, readjustments, VOT

1. Introduction

Recurrent nerve paralysis remains one of the most studied post-thyroidecmy complications that affect voice quality in the literature. The laryngeal immobility which may appear due to this operation is a classic complication of this highly codified surgery (Laccourreye, Malivaud, Ménard, & Bonfils, 2009). Thyroid gland surgery can lead to accidental injury when it is inevitable to injure one recurrent laryngeal nerve, even if the surgeon takes great care to preserve these nerves (Mohil et al., 2011). Fortunately, the incidence of post thyroidectomy unilateral laryngeal paralysis remains relatively low. In fact, according to studies published in recent years (see for example Benninger, et al. 1998 ; Lo, et al. 2000), the rate of laryngeal paralysis in the immediate postoperative period (one month after operation) following an operation of the thyroid gland is between 0.5% and 8.3%. There is no agreement on whether unilateral laryngeal paralysis consecutive to thyroid operation is temporary or permanent, but the final impact of laryngeal paralysis is estimated between 0 and 3.2% with a median of 1% (Trésallet, et al. 2006 ; Laccourreye, et al. 2008). The majority of cases of laryngeal immobility secondary to thyroid surgery present few symptoms, and about one third of the patients recover spontaneously (Laccourreye et al., 2009).

The aim of this study is to analyse temporal characteristics (VOT) of the speech of patients with recurrent paralysis. This longitudinal study falls within the perturbation and readjustment paradigm preoccupied mainly with evaluating the flexibility of the speech production and perception system, determining the range of linguistically tolerated deviations from speech “targets”, and ultimately elaborating viability constraints in articulatory and acoustic terms. Our investigation, although acoustic, aims to analyze the different articulatory-to-acoustic cues such as V.O.T which can be observed from a continuous signal, and may allow tracing articulatory timing patterns and configurations. (Abry et al., 1985).

2. Material and Method

2.1. Patients

7 patients have been recorded: 5 women (anonymised as follows: UPBAS, UPPHEI, UPPHAL, UPPWAL and UPPWAN) and 2 men (anonymised as follows: UPPLAT and UPPPRAI). Along with a short biography, pre- and post-surgery information has been kept in each patient’s medical file. All of the patients are native French speakers between 60 and 70 years old, and who underwent total thyroidectomy for benign nodules. In all cases, the pre-operative laryngeal examination showed no impairment of vocal fold mobility. All patients were seen and recorded once a month during their speech therapy sessions.

Data were obtained in 3 phases: 1) a post-surgery phase 1, 15 days after surgery when the patient’s voice is altered in varying degrees; 2) a post-surgery phase 2, one month after surgery; 3) a post-surgery phase 3, two month after surgery, which allows observing probable voice and speech recuperation. Due to medical constraints, the experimenter had unfortunately not been able to collect data from the patients before surgery. Data were acquired from a healthy control speaker, matched with age and gender with each patient. Such data served as reference values for the disordered subjects.

2.2. Corpus and Data

Acoustic data collection was performed using a Marantz PMD661 professional digital recorder connected to a Sennheiser e845s microphone. The recordings were stored directly onto a CompactFlash card as .wav files. The recordings took place in a hospital for the first recording session and for control speakers. Subsequent recordings took place in the speech therapist’s office. The subject was sitting comfortably in a quiet room about 50 cm away from the
microphone. S/he was instructed to speak normally, to move as little as possible and to stay, more or less, in the same position during the entire recording session. The corpus consisted in VCV nonsense words, where V1 was vowel /i/, V2 was vowel /u/, and where V1 was vowel /a/ and V2 was vowel /i/. The consonant was one of the 6 French oral plosives /p, t, k, b, d, g/. The combination of these sounds allowed creating 12 nonsense words that were inserted in the carrier sentence “Cet________ça”. The subject was asked to pronounce each sentence 10 times, in random order.

2.3 Measurements

Measurements were acquired using the software PRAAT (Boersma, 2001). Vowels /i/ and /u/ were specifically chosen in order to observe the acoustic consequences of larynx trajectory from a low position for /i/ to a high position for /u/, and vice versa. Consonants /b, d, g, p, t, k/ were selected in order to observe effects of place of articulation on intra-oral pressure in relation to timing between burst-release and onset of vocal fold activity (VOT as defined by Lisker and Ambramson, 1964), or the timing between burst-release and onset of a clear formant structure (VOT as specified by Klatt, 1975). These six consonants were also chosen as they allow comparing entirely voiced sequences V1 /b, d, g/ V2 with partially voiced sequences V1 /p, t, k/ V2. The latter sequences should impose more constraints on vocal fold activity, since they require alternating closed and open configurations of the glottis.

3. Hypotheses

Since laryngeal immobility provokes modifications in voice periodicity and difficulties in controlling the movement of the vocal folds and hence of voicing, the following hypotheses are formulated: difficulties in voicing control would have an impact on the duration of VOT. Indeed, this cue, considered as relevant in the characteristics of vocal fold vibrations, and potentially capable of contributing to distinguishing voiced from voiceless consonants, should be modified. Traditionally, Klatt’s (1975) VOT, is shorter for voiced plosives, compared with their voiceless counterparts. For voiced plosives, the CV transition only implies a change in the vocal tract state, from being obstructed to being sufficiently open so as to allow emergence of a clear formant structure. For voiceless plosives, on the contrary, the CV transition requires, besides change in the state of the vocal tract, a modification of the configuration of the glottis, from an open position to produce the consonant to a closed one in producing the vowel. Consequently, due to potential difficulties in maintaining laryngeal vibrations, VOT should thus be longer for voiced plosives (which would be more or less devoiced, and produced with difficulty by the patients), compared with voiced plosives produced by control speakers. Hence, VOT durational values for voiced plosives produced by the patients would resemble those obtained for their voiceless counterparts, thus rendering distinction between the two phonetic categories potentially critical, at least in a quantitative perspective.

4. Results

A three factor analyses of variance (ANOVA) indicates that the main effect for recording phase was significant (p<0.5) for the VOT variable.

4.1. [b] context

We choose to only present in details results obtained from the context [b], the same phenomena being observed for all the other French voiced plosives.

Regarding the voiced plosive [b], VOT duration of control speakers is 11.45 ms (sd = 3.34 ms) for [abi] and 12.17 ms (sd = 3.39 ms) for [iba], and these values are conventional regarding the literature (see eg Sock & Benoit, 1986.). Moreover, the low standard deviation values reveal a certain regularity in the production of these speakers. In contexts [a-i] and [i-a], VOT for patients is always longer than for control speakers (see Figures 1 and 2). This is particularly remarkable in the postop 2 phase, where the patients’ VOT is 32.51 ms (sd = 20.72 ms) for [abi] and 24.77 ms (sd = 10.04 ms) for [iba]. The duration of VOT is reduced from postop 3, where it is then 21.17 ms (sd = 12.41 ms) for [abi] and 16.90 ms (sd = 6.88 ms) for [iba]. This trend continued until postop 4, where VOT is then measured at 18.48 ms (sd = 6.67 ms) for [abi] and 16.01 ms (sd = 9.09 ms) for [iba].

Patients’ recording phases are characterised by higher standard deviations (see values in parentheses) than those of the controls, which may be the result of inter-and / or intra-individual variability.
4.2. Voiceless contexts

VOT duration of voiceless plosives failed to uncover significant differences between the productions of patients and control speakers, regardless of the recording phase. Laryngeal immobility therefore does not seem to perturb this temporal parameter.

4.3. VOT as an articulatory cue

In post-surgery phase 1, VOT duration is longer for the three plosives [b, d, g], compared with that of control speakers, in contexts [a-i] and [i-a]. This phase is also characterised by higher standard deviation values, thus revealing noticeable variability during this recording phase. It should be mentioned that differences in VOT values between patients and control subjects are more pronounced in the [a-i] context. This is due presumably to the [C=>i] transition which requires a more precise and problematic control than the [C=>a] transition. Indeed, producing vowel [i] imposes a reduced constriction in the palatal region, while avoiding contact between the tongue surface and the palate; patients seem to have more difficulty in executing this transition from the consonant to the target vowel [i]. This apparently is not the case for the [C=>a] transition, where jaw lowering, accompanied by that of the tongue body to which it is tightly coupled, creates without any particular difficulty the radico-pharyngeal constriction necessary for the acoustic emergence of this vowel.

Let us recall that VOT may vary with place of articulation. It is known that the further the occlusion is located in the back cavity of the vocal tract, the longer the VOT (Peterson & Lehiste, 1962). This means that the more extended the contact area, the higher the VOT. Hence, VOT is generally shorter for bilabial plosives and longer for velar plosives, apical plosives having intermediary values in most cases. Moreover, it has been found that VOT (Klatt, 1975) is longer for voiceless plosives than for voiced plosives in French (Sock and Benoît, 1986).

In French, it is usually the case that VOT duration for [b] is shorter than that of [d] which, in turn is shorter than that of [g]. VOT is also shorter for [p] than for [t] or [k].

If the abovementioned data apply to subjects without speech disorders, Figures 1 and 2 show that these findings are also
verified for patients with recurrent paralyses, in both vowel contexts and for all recording phases. If VOT duration is indeed increased in post-surgery 2 phase for production of voiced plosives by patients, they are however quantitatively distinct from those of their voiceless counterparts.

5. Conclusions

Difficulty encountered by patients in producing voiced sounds thus has an impact on VOT values of voiced plosives alone. Difficulty in maintaining laryngeal vibrations throughout the closure of the purportedly voiced plosives lead patients to produce longer VOTs. Patients are constrained to interrupt periodic laryngeal activity quite early, while producing the target voiced consonant; this therefore favours relatively precocious release, and hence an increase in VOT duration. With time and speech therapy, reduction of VOT values, as early as in post-surgery phase 2, is observed, thus suggesting enhancement in the control of laryngeal gestures. The phenomenon is pursued in post-surgery 3 phase, when VOT values are then comparable to those of control speakers, and specifically those of plosives [b] and [d].

Analyses of the productions of patients, taken separately, allowed to show that the remarkable variability observed in post-surgery 1 phase, was a reflection of inter and intra subject differences.

In general, this variability diminishes in late recording phases, thereby confirming better control of laryngeal activity in such stages: productions then become more regular. Analyses of VOT duration of voiceless plosives did not allow highlighting significant differences between productions of patients and those of control speakers, regardless of the recording phase. It therefore seems that laryngeal paralyses do not perturb this temporal parameter during the production of voiceless stops in [C>V] transitions. Let us recall that it is especially at the spectral level that the vowel is modified (Fauth, 2012) and not at the level of onset of laryngeal activity, relative to release.

Finally, let us also recall that VOT duration is indeed lengthened in post-surgery 1 phase in the production of voiced plosives in patients. Nonetheless, VOT values are quite distinct between voiced vs. voiceless plosives.

6. Limits and perspectives

Note that speaker variability is still very high in postoperative recording phases. It might be useful to increase the number of speakers in order to classify patients into subgroups according to the impact of surgery on the patient’s voice. Additional recordings are also underway to increase the number of subjects.

We are currently investigating modification of other intrasegmental spatiotemporal parameters like VTT or Voice Termination Time (Agnello, 1975), also called voiced closure or closure voicing, occlusion duration (closure duration excluding VOT), the acoustic silent phase, etc.

7. Acknowledgements

This work was supported by the program « Perturbations et Réajustements : parole normale vs parole pathologique » of the Maison Inter-Universitaire des Sciences de l’Homme Alsace (MISHA), 2008-2012, and also by an ANR Grant “DOCVACIM” to the Phonetics Institute of Strasbourg / U.R. LiLpa, E.R. Parole et Cognition and by the program of the Scientific Committee of the University of Strasbourg (UdS) Gutenberg-Strasbourg, 2009-2011.

8. References


Modeling individual variation in prosody: the case of Clitic Left-Dislocations in Spanish
Ingo Feldhausen
Institut für Romanische Sprachen und Literaturen, Goethe-Universität Frankfurt am Main, Germany
Labex EFL, UMR 7018-LPP, Université Paris-Sorbonne Nouvelle, France
ingo.feldhausen@gmx.de

Abstract
This paper offers new perspectives on how the phenomenon of variation in linguistic data can be integrated into formal grammatical theory. Drawing from data on the prosodic phrasing of sentences with embedded clitic left-dislocations in Spanish, a version of the Stochastic Optimality Theory is proposed which accounts for the attested inter-speaker variation in the data. The main idea of the proposal is that the degree of constraint overlap is not fixed for the entire population. Rather, each speaker exhibits a different degree of overlap, while the underlying constraint hierarchy is invariant. The proposed theoretical account differs from previous studies in two crucial ways. First, it deals with prosodic phrasing (while most studies have been conducted on segmental or syntactic issues). Second, it is not limited to variation within a population, but also accounts for inter-speaker variation.

Keywords: Prosody, Individual or Inter-Speaker Variation, Clitic Left-Dislocation, Spanish, Stochastic Optimality Theory

1. Introduction
While it is well known that prosodic characteristics such as (a) intonational phrasing, (b) the realization of prosodic boundaries and (c) the realization of pitch accents differ between languages (cf. Jun 2005, 2014), a growing amount of evidence questions whether these aspects are used homogenously within a language. Work such as Grabe (2002), Féry (2004), Feldhausen (2010, 2011, 2012), Myrberg (2010, 2013), Niebuhr et al. (2011) are some of the few studies addressing the issue of significant, systematic inter-speaker variation within a population. Since this variation cannot be attributed to factors such as style, genre, syntactic or prosodic complexity, we appear to be dealing with a case of free variation. Free variation is defined as the occurrence of two forms in the same context without changing the meaning of the utterance. While prosodic research typically concentrates on describing and modeling the results of the average speaker / listener, the aforementioned studies support the view that individual / inter-speaker differences should not be ignored. Following this vision, I propose a theoretical model for capturing individual variation in prosody. More concretely, I propose a modification of the Stochastic Optimality Theory (SOT, Boersma & Hayes 2001) in order to account for the free variation attested in the prosodic phrasing of sentences with embedded clitic left-dislocations in Spanish. The reason why individual differences have yet to be thoroughly considered in prosodic research might be due to the fact that in generative phonology, accounts modeling linguistic competence in the area of sound structure have previously been based on categorical claims: The models are non-probabilistic, and for this reason any given sequence is either grammatically or completely impossible (Pierrehumbert 2001: 195). As a consequence, variation has largely been ignored or eliminated: In these models, variation in observed data typical-
boundaries in which some boundaries are obligatory and others are optional (see §2). The fact that some boundaries are optional in this data set is in line with work such as Frazier et al. (2006), who show that the relation between prosodic boundaries (i.e. their position and size) is more important for the correct interpretation of a sentence than a strictly local match between syntactic and prosodic structure.

2. Experimental data

The material for the present study is taken from Feldhausen (2012), who is interested in the question as to whether CLLD constituents in Spanish display a prosodic phrase break at both their right and left edge. Prosodic phrases in speech are indicated by pauses, changes in pitch and amplitude, and/or the lengthening of the final syllable(s) within a phrase.

Feldhausen (2012) conducts a production experiment (based on scripted speech) in which a homogenous group of four native speakers of Peninsular Spanish uttered 144 sentences with embedded CLLD as in (1) and non-embedded CLLD (El águila, la vendió mi hermano ‘The eagle, my brother sold.’). CLLDs are syntactically characterized by the presence of a phrase in the left periphery of a clause (el águila de Málaga in (1)) which is connected with the clause by means of an anaphoric element (the accusative clitic lo in (1)).

His results show that CLLD constituents have an obligatory right boundary, both in matrix and embedded contexts, (2). The boundary is typically marked by high edge tones. In embedded contexts, the matrix clause is also obligatorily separated from the embedded clause containing the CLLD by a prosodic break (see (2); n.b. even though the boundary precedes the complementizer que ‘that’, one can say that the CLLD constituent has a left boundary, as que does not have word accent and prosodically counts as a clitic). The results also show that the prosodic break between the subject and the verb of the matrix clause is optional (marked by _ in (2)). Actually, it is realized in only 25% of the cases between the speakers.

( S _ V )(q CLLD )( ... )(2) Bárbara supone que el águila de Málaga, la vendió su hermano

Table 1: Percentages of individual realizations of the ‘S/(V’-boundary between target sentences with embedded CLLD in Feldhausen (2012).

<table>
<thead>
<tr>
<th>S/(V phrasing</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>A</td>
</tr>
<tr>
<td>33%</td>
<td>B</td>
</tr>
<tr>
<td>28%</td>
<td>C</td>
</tr>
<tr>
<td>39%</td>
<td>D</td>
</tr>
<tr>
<td>25%</td>
<td>Total</td>
</tr>
</tbody>
</table>

Even though the average percentages of S/(V –phrasing add up to 25%, there is great variation between speakers, as shown in Table 1. While speaker A never realizes a boundary between S and V, speaker D realizes a boundary in 39% of the cases. Speaker B and C fall somewhere in between. Proposing a theoretical approach for the phrasing of CLLD in Spanish by only considering the average percentages of the population crucially ignores the clear differences between speakers. The degree of optionality differs between speakers and can hardly be caught by the average percentage. In what follows, I concentrate on embedded CLLD, even though the analysis also captures non-embedded CLLD.

3. The Optimality Theoretic Analysis

3.1. Stochastic Optimality Theory (SOT)

In the traditional version of Optimality Theory (Prince & Smolensky 1993/2004), grammars of single languages are considered to consist of categorically ranked constraints C in a strict dominance hierarchy: C1 >> C2 >> C3. Due to the strict order, C1 can never dominate C2 and C2 never dominates C3. Boersma (1998) and Boersma & Hayes (2001) modify this view in SOT by assuming (i) a continuous ranking scale and (ii) a stochastic candidate evaluation.

Figure 1 illustrates a continuous ranking scale on which constraints have a certain ranking value. Higher values correspond to higher-ranked constraints; lower values correspond to lower-ranked constraints. A shorter distance between two constraints (e.g. C2 and C3 in Figure 1) implies that the relative ranking of the constraints is less fixed.

Figure 1: Continuous ranking scale.

Next, it is assumed that the candidate evaluation is stochastic. At evaluation time (i.e. the time at which the candidates are evaluated to determine a winner), each constraint position is perturbed by a random positive or negative value. As a consequence, a constraint is not a single point, but acts as if it is associated with a range of values. These ranges are interpreted as normal (= Gaussian) distributions, see Figure 2.

Figure 2: Overlapping constraints.

The permanent value of a constraint is called the ranking value and constitutes the center of the range. The concrete constraint value at evaluation time is called the selection point. The perturbation allows the selection point to fall anywhere in the range. When the distance between two constraints is relatively short (as for C2 and C3 in Figure 2), the constraint ranges overlap (black area in Figure 2). Due to the probabilistic distribution of the ranges, a selection point near the center is more probable than a selection point far from the center. The result is a ‘normal’ ranking in which C2 >> C3. However, if the selection point of C2 is very low, while the selection point of C3 is very high, the reverse order is the result (C3 >> C2). The more the constraints overlap, the more likely a reverse ranking is. This means that in a certain percentage of the evaluations (depending on the amount of overlap), a lower-ranked constraint is preferred over a higher-ranked constraint and the suboptimal candidate wins.

3.2. Accounting for variation within a population

In order to account for the prosodic phrasing of utterances with CLLD in Spanish, I rely on six constraints from three important families (alignment, markedness and STRUCTURE), and make use of the constraint ranking proposed by Prieto (2006) for Spanish SVO utterances. With Prieto (2006), I assume that the classical alignment constraint ALIGN-XP,R (Selkirk 1995) is active in Spanish. ALIGN-XP,R requires the right edge of a syntactic XP to be aligned with the right edge of a prosodic phrase. I also assume that two further alignment constraints, namely ALIGN-TO,R and ALIGN-CP,L - proposed
by Feldhausen (2010, 2011) for Catalan - are likewise active in Spanish. ALIGN-TOP,R requires a prosodic boundary at the right edge of CLLD constituents, while ALIGN-CP,L demands a boundary at the left edge of the embedded clause (CP). In concordance with Prieto (2006), I further assume that the structure-avoiding constraint MIN-N-PHRASES (minimize the number of prosodic phrases; see Truckenbrodt 1999:228) and the two markedness constraints on the size of prosodic phrases, MIN-BIN and MAX-BIN\((\text{IP HEAD})\), are active in Spanish. MIN-BIN requires that prosodic phrases consist of a minimum of two prosodic words \((\omega)\). MAX-BIN\((\text{IP HEAD})\) demands that the prosodic phrase bearing sentence stress consists of a maximum of two \(\omega\). I propose the constraint ranking in (3) (The constraints in bold letters represent the proposed order from Prieto 2006):

\[
\text{ALIGN-TOP,R} \gg \text{MAX-BIN(IP HEAD)} \gg \text{ALIGN-CP,L} \\
\gg \text{MIN-N-PHRASES} \gg \text{ALIGN-XP,R} \gg \text{MIN-BIN}
\]

(3)

Table 2: Normal ranking for embedded CLLD

<table>
<thead>
<tr>
<th>(SV) (CLLD) (VO)</th>
<th>ALIGN-TOP,R</th>
<th>MAX-BIN(IP HEAD)</th>
<th>ALIGN-CP,L</th>
<th>MIN-N-PHRASES</th>
<th>ALIGN-XP,R</th>
<th>MIN-BIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 3: Reverse ranking for embedded CLLD

<table>
<thead>
<tr>
<th>(SV) (CLLD) (VO)</th>
<th>ALIGN-TOP,R</th>
<th>MAX-BIN(IP HEAD)</th>
<th>ALIGN-CP,L</th>
<th>MIN-N-PHRASES</th>
<th>ALIGN-XP,R</th>
<th>MIN-BIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>(S V) (CLLD) (V O)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

The high-ranked constraints ALIGN-TOP,R and ALIGN-CP,L account for the fact that clitic left-dislocations always display a prosodic boundary at their right edge and that embedded clauses are always separated prosodically from the matrix clause (see Table 2 and Table 3). The constraint ranking in (3) accounts for the (S V)(q CLLD)(VO) pattern, i.e. the ‘normal’ phrasing pattern, which is realized in 75% of the cases (Table 2). Based on the idea of overlapping constraints, I assume that MIN-N-PHRASES and ALIGN-XP,R overlap (see blue box in Table 2 and Table 3). The overlap occurs to such a degree that the frequency-dependent variation (75% vs. 25%) within the population is accounted for (to be revised in § 3.3). Table 3 shows the ranking for the more rarely realized (S(V) (q CLLD)(VO) pattern (25%).

In sum, by using SOT, I can account for both the obligatory prosodic boundaries at the right and the left edge of Spanish CLLD constituents and the optional boundary between the matrix subject and verb. By assuming that MIN-N-PHRASES and ALIGN-XP,R overlap, the model explains the frequency-dependent distribution of the matrix SV-boundary. However, the model only accounts for the frequencies based on the entire population (75% vs. 25%). Up to now, the obvious differences between speakers are left out of consideration.

3.3. Accounting for individual variation

In order to account for the attested inter-speaker variation, I propose that the distance between the overlapping constraints differs between speakers. At the same time, the underlying constraint hierarchy of a grammar is invariant (exactly as in other OT accounts). As for the phrasing of matrix S and V, this means that the degree of overlap between the constraints MIN-N-PHRASES and ALIGN-XP,R is not constant for the whole population. Thus, the differences in the frequency of the output forms between speakers arise as the result of varying degrees of overlap in MIN-N-PHRASES and ALIGN-XP,R, while the underlying constraint hierarchy remains unchanged. Figure 3 shows the grammar of speaker A, in which MIN-N-PHRASES and ALIGN-XP,R do not overlap. Since the selection point of ALIGN-XP,R can never be higher than the selection point of MIN-N-PHRASES, speaker A does not show any variation and realizes matrix (SV) 100% of the time. Figure 4 shows the grammar of speaker D. Here, the two constraints overlap, as the center of the higher-ranked constraint MIN-N-PHRASES and the center of the lower-ranked constraint ALIGN-XP,R are located closer to one another. As a consequence, the lower-ranked constraint is preferred over the higher-ranked constraint by speaker D in 39% of the cases.

The results from Feldhausen (2012) show that speakers B and C also show variation, but to a lesser extent than speaker D. Accordingly, MIN-N-PHRASES and ALIGN-XP,R also overlap in the grammars of speakers B and C, but the distance between the constraints is longer than in the grammar of speaker D and as such, the constraints overlap to a lesser extent. Even though each speaker exhibits a different degree of overlap, they all share the same underlying constraint hierarchy. This idea is illustrated in Figure 5. The grammar of Spanish, i.e. the underlying constraint hierarchy given in (3), includes the grammars of the individual speakers, as exemplified here for the grammars of speakers A and D (i.e. Figure 3 and Figure 4). Accordingly, the figure representing the grammar of Spanish would also encompass the grammars of speakers B and C (as well as the grammars of every other speaker in that population).

Figure 3: Grammar of speaker A – (SV) phrasing, but no (S)V.

Figure 4: Grammar of speaker D – (SV) and (S)V phrasing.

The proposed model can explain variation within a population and variation between its speakers. A further advantage to the account is that it also explains intra- or within-speaker variation, since the speaker-specific degree of overlap explains why different speakers use different forms to varying extents (cf. speakers B, C, and D) and why some speakers show no variation (cf. speaker A).
Furthermore, the data show that some boundaries are obligatory, while others are optional. The fact that the boundaries in the vicinity of the CLLD are obligatory across speakers, while the one between matrix S and V is optional, suggests that the CLLD boundaries are more important for the interpretation of the sentence than the SV-boundary. This seems to be in line with the 'rational speaker hypothesis' (Clifton et al. 2002, Frazier et al. 2006), which states that a speaker uses prosody in a rational and internally consistent way, while the listener presumes such rationality during interpretation. The present approach reflects this hypothesis in the following way: Highly relevant prosodic patterns are expressed by higher-ranked constraints, which are not prone to overlap (here ALIGN-TOp,R and ALIGN-CP,L). In contrast, lower-ranked constraints, which are less relevant for the interpretation, might overlap (here MIN-N-PHRASES and ALIGN-XP,R). The CLLD boundaries are very important interpretatively, as they do not only align with syntactic edges (of an XP, which is highly marked due to its dislocation), they are also relevant in terms of information structure. By assuming together with Frazier et al. (2006:244) that the "impact of prosodic boundaries depends on the other prosodic choices a speaker has made", it becomes clear why the matrix SV-boundary is optional; it does not fulfill the interpretative requirements (to the same degree) and is as such optional between speakers (to varying degrees). This explains the sharp contrast compared to the phrasing pattern of simple SVO structures in Spanish, in which the boundary between S and V is obligatory (e.g. Prieto 2006, Feldhausen et al. 2010). In SVO structures, the SV-boundary is the only sentence-final boundary and is not influenced by other boundaries.

The present approach assumes that the underlying order of constraints is fixed for all speakers within a population. In other words, if speakers have different underlying grammars, they speak a different dialect or language. This assumption predicts that no speaker realizes the 'reverse' order more often than the 'normal' order, i.e. no speaker realizes matrix (S)(V) more often than matrix (S V). This observation would need to be confirmed empirically in future research. A study with only four speakers might not be robust enough to test such predictions. However, if a 'reverse' ranking is especially prevalent for one speaker, a closer look into the questionnaire on the personal background of the subject ('sociolinguistic questionnaire') might be fruitful. It is possible that other influential factors exist which have not been considered yet. Nevertheless, the assumption that individual grammars are built based on the grammar of a given language fits with the ideas in Bresnan et al. (2007). They conclude that an individual grammar leads "to isolated loci of variability in the grammar rather than complete alternations of paradigms". The isolated locus of variability is given here by the overlap of MIN-N-PHRASES & ALIGN-XP,R.

In summary, if Pierrehumbert (2001) and Bresnan et al. (2007) are right in assuming that variation is an idiosyncratic and intrinsic part of linguistic competence, a formal grammatical theory should be able to account for systematic individual differences, the extent to which has yet to be determined. The present proposal offers an important leap in this direction.

6. References


Kirchner, R. (2002). Preliminary thoughts on "phonologisation" within an exemplar-based speech processing model. UCLA WP 6.


Niebuhr, O., M. D’Imperio, B. Gili Fivela & F. Cangemi. (2011). Are factorial structures in Catalan linguistic competence, a formal grammatical theory should be able to account for systematic individual differences, the extent to which has yet to be determined. The present proposal offers an important leap in this direction.

5. Acknowledgements

My gratitude goes to Shinichiro Ishihara, Monika Rathert and the audience of the 2012 TIES Conference in Oxford, UK for their helpful comments and questions. Thanks also to Audrey MacDougall for her assistance with editing. This work was supported by the French Investissements d’Avenir - Labex EFL program (ANR-10-LABX-0083).
Intonation and preverbal subjects in Italian

Ingo Feldhausen
Institut für Romanische Sprachen und Literaturen, Goethe-Universität Frankfurt am Main, Germany
UMR 7018-LPP, Université Paris-Sorbonne Nouvelle, France
ingo.feldhausen@gmx.de

Abstract

This work constitutes the results of a case study showing how a detailed investigation of prosodic properties can shed new light on long debated syntactic questions. There is an ongoing discussion in syntactic literature on Italian as to whether preverbal subjects are structurally located in the C-domain (CP) or in the Infl-domain (TP). Since there are good syntactic arguments for both positions, a general lack of consensus remains concerning this issue and it seems unlikely that syntactic tests alone are able to solve the problem. The present paper, in contrast, provides intonational evidence that preverbal subjects in Italian Subject-Verb-Object (SVO) structures are in the C-domain.

Keywords: Italian, Intonation, Preverbal Subjects, Syntactic Position, Prosodic Phrasing, Left-Periphery

1. Introduction

In the syntactic literature on Italian (and other Romance pro drop languages) there is an ongoing discussion on whether preverbal subjects such as il popolo 'the subjects' in the SVO structure in (1) are structurally located in the C-domain or in the Infl-domain, see Fig. 1. With Italian being a pro drop language, pró may always occupy the Spec,TP position (Rizzi 1982).

Il popolo guardava il monumento. (1)

‘The subjects looked at the monument.’

Figure 1: C-domain, Infl-domain and potential subject positions.

Benincà & Cinque (1985), Giupponi (1988), Poletto (2000) and others put forth the hypothesis that S is located in the C-domain, i.e. the domain of the complementizer phrase CP (this assumption is called S-in-C Hypothesis here). In her thorough study, Poletto provides different kind of evidence in favor of this assumption. For example, in Piedmontese (Turin) as in (2), the subject gnun ‘nobody’ precedes the complementizer che ‘that’, which is the head of the CP. She therefore concludes that S has to be located in the C-domain.

Gnun che a s’bogia! (2)

‘Nobody moves’

Since non-specific quantifiers such as gnun ‘nobody’ cannot be left-dislocated (Rizzi 1982, Cinque 1990), (2) also presents evidence that S must not necessarily constitute an instance of (clitic) left-dislocation (CLLD; Poletto 2000:149). CLLD is generally assumed to be located in the C-domain (Rizzi 1997) and to represent given information. Cardinaletti (2004), on the contrary, claims that S is always located in the Infl-domain, i.e. the domain of the inflectional/tense phrase TP (this assumption is referred to as the S-in-Infl Hypothesis). In her careful study, she challenges the analyses made by Poletto (2000). As for (2), for example, she argues that che ‘that’ is not a complementizer, but rather a head of a functional projection for irrealis mood, which is located in the Infl-domain. As a consequence, the subject gnun in (2) can easily be located in Spec,TP. Despite these diverging analyses of the preverbal subject position, Poletto (2000) and Cardinaletti (2004) agree on the impossibility of (unspecific) quantifiers being left-dislocated.

This purely syntactic (and unresolved) discussion is complemented here by considering intonational aspects of SVO and CLLD structures in Italian. Actually, two characteristics are important for testing the subject position from an intonational point of view. First, broad focus SVO structures in Italian are phrased as single prosodic units without sentence-internal intonational breaks (D’Imperio et al. 2005). This means that (SVO) is the typical phrasing pattern of such structures in Italian (3c), while (S)(VO) is not (3d). Second, CLLD in Italian (and other Romance languages) is characterized by an obligatory prosodic break at its right edge (Gili Fivela 1999, Frascarelli 2000, Bocci 2007), see (3e). This boundary can be taken as evidence for the more general assumption that a prosodic break occurs between CP and TP, as proposed in works by Elordieta et al. (2005), Ishihara (2007), and Kratzer & Selkirk (2007), see (3a,b).

a. \([\text{CP} \ldots] \ [\text{TP} \ldots] \) Syntactic Structure (3)
b. ( \( \) ) Prosodic Structure
c. \((\text{SVO})\) S

d. \((\text{VO})\) C

e. \((\ldots)\) CLLD

The S-in-C Hypothesis and the S-in-Infl Hypothesis lead to different predictions with respect to the prosodic phrasing pattern of SVO structures. The former hypothesis predicts phrasing pattern (3d), since S is assumed to be located in the C-domain, while the S-in-Infl Hypothesis predicts pattern (3c), since S is assumed to be located in the Infl-domain. Based on the previous research described above, the following hypotheses can be established:

- Hypothesis 1: Clitic left-dislocations are located in the C-domain and as such are prosodically separated from the Infl-domain (see (3e); following Frascarelli 2000, Bocci 2007)
- Hypothesis 2: Quantifier subjects in all new contexts are not prosodically separated from the following...
material and as such are located in the Infl-domain (following Cardinali 2004, D’Imperio et al. 2005).

- **Hypothesis 3:** Preverbal, lexical subjects that are contextually given have the status of a CLLD and are thus prosodically separated from the Infl-domain.
- **Hypothesis 4:** Preverbal, lexical subjects in all new contexts are not prosodically separated from the following material and as such are located in the Infl-domain (following Cardinali 2004, D’Imperio et al. 2005).

### 2. Methodology

To test the hypotheses, a production experiment based on scripted speech was conducted, in which four native speakers of Northern Italian (3F = Venice / Verona; 1m = Siena; 26-29 years) read a total of 160 sentences (including filler sentences). The material was created based on the four following conditions C1-4 (equivalent to hypotheses 1-4):

- **Condition 1 (C1):** Sentences with (clitic) left-dislocated objects, (4).
- **Condition 2 (C2):** SVO structures with a quantifier subject (QP), (5).
- **Condition 3 (C3):** SVO structures with lexical subjects that are not part of the focus domain, (6).
- **Condition 4 (C4):** broad focus SVO structures with lexical subjects, (7).

The crucial condition is C4. If C4 shows the phrasing pattern (3c), the S-in-Infl Hypothesis is supported. But if C4 shows the phrasing pattern (3d), the S-in-C Hypothesis is supported.

Seven target sentences were created for each condition. Corresponding examples are presented in (4)-(7). Each target sentence (b. sentences / bold letters) was accompanied by an appropriate context question (a. sentences) in order to fulfill the information structural needs of the target sentences. While the given preverbal material is mentioned in the preceding context question (this is the case for LD objects (C1) and given subjects (C3)), the preverbal elements in C2 and C4 are not mentioned in the context.

#### Left-dislocated object, (C1):

- a. Che cosa succede con il popolo?
  "What happened with the subjects?"
- b. Il popolo l’hanno convocato ieri.
  'The subjects, (s/he) convened them yesterday.'

#### Quantifier subject, (C2):

- a. Cosa succedeva?
  "What happened?"
- b. Nessuno guardava il monumento.
  'No one looked at the monument.'

#### Given subject, (C3):

- a. Che cosa faceva il popolo?
  "What were the subjects doing?"
- b. Il popolo guardava il monumento.
  'The subjects, they looked at the monument.'

#### Broad focus subject, (C4):

- a. Cosa succedeva?
  "What happened?"
- b. Il popolo guardava il monumento.
  'The subjects looked at the monument.'

The data were directly recorded as .wav-files on a Samsung SIII mobile phone by means of the Tape-a-Talk software (16 kHz, 16 bit). The recordings took place in a quiet room at the University of Frankfurt. The acoustic and instrumental analysis was conducted using the latest It_ToBI annotation system (Gili Fivela et al. 2013) by considering three tonal boundary cues: (a) continuation rise (H-), (b) sustained pitch ('H-'), and (c) complex boundary tone (LH-); see Frota et al. (2007) for details. Further cues such as the lengthening of preboundary syllables (Nespor & Vogel 1986: 176, Dehé & Samek-Lodovici 2009) or the application of phonological rules such as Raddoppiamento Sintattico (RS; Lepsky & Lepsky 1977, Nespor & Vogel 1986/2007:165, Frascarelli 2000:19) or the Rhythm Rule (RR, Frascarelli 2000:20) were not considered. For example, RS is characterized by the lengthening of consonants within the intermediate phrase. This rule only applies to some varieties (e.g. in the Toscana region, Rome, and in southern Italy), however, and not to the Northern ones. While RR, on the other hand, exists in Northern varieties, it is unclear whether the Verona/Venice region shows this ip-internal repairing strategy for stress clashes. For these reasons, I did not consider them in this work.

### 3. Results

The results show that a prosodic boundary is typically realized after the preverbal element in all conditions. As shown in Table 1, the total number of boundary realizations varies between the four conditions, from 60% in C4 to up to 75% in C1. The number of lacking realizations of a boundary is lower (from 15% in C1 up to 40% in C4). A chi-square test shows that the difference between the total number of boundaries across all conditions (N=54) and no boundaries (N=21) is significant ($\chi^2 = 14.52$, df = 1, $p < 0.01$ (p = 0.00014)). Further, a test of equal or given proportions (R) shows that the frequency differences between the conditions are not significant ($p = 0.78$). The boundary after S (or O) is realized most commonly by 'H- (56%), followed by H- (39%), Table 2. There are only few instances of the low boundary tone L% (3%).

<table>
<thead>
<tr>
<th>Boundary type</th>
<th>Boundary unclear</th>
<th>No boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>O_LD (C1)</td>
<td>15 (75%)</td>
<td>2 (15%)</td>
</tr>
<tr>
<td>S_quantu (C2)</td>
<td>14 (70%)</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>S_given (C3)</td>
<td>13 (65%)</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>S_new (C4)</td>
<td>12 (60%)</td>
<td>8 (40%)</td>
</tr>
</tbody>
</table>

Table 1: Total number and frequency of boundaries at target position per condition.

<table>
<thead>
<tr>
<th>Boundary type</th>
<th>Boundary unclear</th>
<th>No boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>''H-</td>
<td>30 (56%)</td>
<td></td>
</tr>
<tr>
<td>H-</td>
<td>21 (39%)</td>
<td></td>
</tr>
<tr>
<td>L-</td>
<td>3 (5%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 presents the tonal configurations on the preverbal element (i.e. target position). The most prevalent tonal configuration is L+H* 'H- (37%). A corresponding F0 contour is given in Figure 5 for C4. Further tonal configurations on the preverbal element are L* H- (as shown in Figure 4 for C3), L+H* 'H- (as in Figure 3 for C2), L+H* 'H-, and L*+H 'H- (as in Figure 2 for C1). No clear preference was attested between the tonal configurations and conditions.
While the sentence internal configuration may vary (Figure 3), no variation was observed with respect to nuclear configuration (i.e. the nuclear accent and the following IP edge tone). The sole nuclear configuration used was H+L* L-L%, as can be seen in Figures 2 to 5. This pattern confirms previous research on Italian declarative sentences (e.g. Gili Fivela et al. 2013).

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>L+&gt;H* !H-</td>
<td>20</td>
<td>37%</td>
</tr>
<tr>
<td>L* H-</td>
<td>9</td>
<td>17%</td>
</tr>
<tr>
<td>L+H* !H-</td>
<td>9</td>
<td>17%</td>
</tr>
<tr>
<td>L*+H* H-</td>
<td>8</td>
<td>15%</td>
</tr>
<tr>
<td>L*+H !H-</td>
<td>5</td>
<td>9%</td>
</tr>
</tbody>
</table>

Table 3: Total numbers of the most common tonal configurations in target position.

4. Discussion & Conclusion

In general, our results confirm the S-in-C Hypothesis and do not support the S-in-Infl Hypothesis. By discussing each hypothesis individually, I will demonstrate what led to this interpretation.

Hypothesis 1 could be verified, since clitic left-dislocated objects were found to be prosodically separated from the following material. This confirms previous research on CLLD in Italian (Gili Fivela 1999, Frascarelli 2000, Bocci 2007) and as such adheres to the previous understanding of the prosodic phrasing of CLLD in Romance languages (see Delais-Roussarie et al. 2004 for French, Feldhausen 2010 for Catalan, Feldhausen 2014 for Spanish, Frota & Vigário 2007 for Portuguese). Hypothesis 2 was falsified: Quantifier subjects in all new contexts are actually prosodically separated from the following material. As the boundary is assumed to be located between CP and TP, quantifier subjects cannot be taken as being located in the Infl-domain. This contradicts the claims made by Cardinaletti (2004) and supports the analysis of Polletto (2000). Hypothesis 3 could be confirmed in this study. Preverbal subjects that are contextually given are prosodically separated from the Infl-domain and behave prosodically as left-dislocated objects (in fact, they can be considered as instances of CLLD). The final and crucial hypothesis, H4, was falsified. Preverbal, lexical subjects in all new contexts display
a prosodic boundary at their right edge and are thus prosodically separated from the following material. This contradicts the phrasing pattern presented in D’Imperio et al. (2005). Furthermore, and more importantly, this finding challenges the predictions made using the S-in-Infl Hypothesis, which assumes that such subjects do not display a prosodic boundary at their right edge. This clearly supports the S-in-C Hypothesis.

A possible argument against this interpretation may be drawn from the test design; it is possible that the speakers did not understand the question-answer pair in its full dimension and thus interpreted the preverbal subjects as given. This apparent counter-argument can easily be ruled out, however. Quantifier subjects also display a prosodic boundary at their right edge and cannot be taken as instances of CLLD (a fact on which Cardinaletti 2004 and Polletto 2000 agree). It is rather unlikely that the speakers interpreted the contextually new subjects as given material while interpreting quantifier subjects as new within the same experiment.

The results (as given in Table 1) have shown several instances with an unclear boundary or no boundary at all, which one could claim to weaken the conclusions drawn in this study. Even though statistical tests have clearly shown that these instances are not significant, future research on that topic should include both a greater number of participating speakers and a greater number of boundary cues. Several cases of "unclear" and "no boundary" might very well display a boundary in the present study when considering further boundary cues such as preboundary lengthening or sandhi rules. This would mean that the actual number of instances displaying a boundary is even higher than presented here. A larger number of participants, on the other hand, would not only constitute a more representative group of speakers, it could also show whether the number of cases of "unclear" and "no boundary" in fact remains small and not significant within a larger population. I believe that the prevalence of these cases will in fact remain insignificant, as the present results are consistent between the speakers.

Finally, I would like to comment on the somewhat surprising differences observed here in the phrasing pattern of all new SVO structures, (namely (S)(VO), as opposed to (SVO), as reported in D’Imperio et al. 2005). Even though both studies considered all new sentences, the results clearly differ from one another. This might be due to the strong dialectal variation for which Italy is known (see, e.g. Berrato 1995, Gili Fivela et al. 2013). While D’Imperio et al. (2005) work on Neapolitan Italian, I consider northern varieties. Gili Fivela et al. (2013) report on all new SVO structures in Pisa Italian, which also displays a (S)(VO) pattern. Perhaps (S)(VO) vs. (SVO) is a pattern that distinguishes northern from southern Italian varieties. If so, it is possible that S is located in CP in northern varieties, while it is in the TP in southern varieties - precisely the same characteristic that distinguishes Portuguese from Spanish (according to Elordieta et al. 2005).

In summary, this study shows that intonation helps in resolving syntactic problems. Together with syntactic arguments, it is possible to create strong evidence for theoretical claims. In the present study, intonational aspects have shown that the preverbal subject in Italian SVO structures is best analyzed as being in Spec,CP.

5. Acknowledgements

I would like to thank Irene Caloi, Nicole Déhé, Elisabeth Delais-Roussarie, Martin Elsag, Izarbe García Sánchez, Georg Kaiser, Imme Kuchenbrandt, Anna Marchesini, Cecilia Polletto, and Emanuela Sanfelici for their helpful comments and discussion. My gratitude also goes to Audrey MacDougall for her assistance with editing. This work was supported by the French Investissements d’Avenir - Labex EFL program (ANR-10-LABX-0083).

6. References


Prosody, Focus and Word Order in Catalan and Spanish: An Optimality Theoretic Approach

Ingo Feldhausen1 & Maria del Mar Vanrell2

1Goethe-Universität Frankfurt, Germany & UMR 7018-LPP, Université Paris 3, France
2Freie Universität Berlin, Germany

ingofeldhausen@gmx.de, mariadelmar.vanrell@fu-berlin.de

Abstract

This paper sheds new light on the question of how neutral and contrastive focus is realized in Spanish and Catalan. By considering data collected through a production experiment based on semi-spontaneous speech, this study reveals two main findings: (a) the two languages use different strategies (most importantly for neutral focus), and (b) previous research can only be partly supported. In terms of neutral focus, the most common strategies are clefting and p-movement in Spanish and dislocating (the given material) and fronting in Catalan. As for contrastive focus, both languages use cleft constructions. In a second step, we propose a stochastic optimality theoretic approach to account for the syntactic and prosodic focus realizations in the two languages. By modifying previous proposals, we account for the variation attested in the realizations as well as for the differences between the two languages.

Keywords: Word order, Narrow neutral and contrastive focus, Prosody, Stochastic Optimality Theory, Spanish, Catalan.

1. Introduction

It has previously been assumed for Catalan and Spanish that neutrally focused elements must always be located in sentence final position, where they receive main stress by means of the Nuclear Stress Rule (NSR, Chomsky & Halle 1968, Zubizarreta 1998). Unlike English, Catalan and Spanish are said to not be able to resort to prominence shift as a strategy for varying the location of the main prominence in the sentence. Zubizarreta (1998, 1999) argues that Spanish uses a syntactic mechanism called (prosodically motivated)-movement: non-focal material is moved to a non-final position as the result of prosodic demands, with the neutrally focused element being located in sentence-final position (see (1)). For Catalan, Vallduví (1991, 1992) claims that non-focal material has to be either left- or right-dislocated; only focal material remains in the core clause (IP) (see (2)). However, a more recent study by Domínguez (2002) argues that Catalan can use syntactic operations other than dislocation (such as p-movement) to mark neutral focus, as demonstrated in (3).

a. Los alumnos se enfrentaron [ con la policia].
   ‘The students confronted the police.’
   (1)

b. *[Los alumnos] se enfrentaron con la policia.
   ‘*The students confronted with the police.’

c. Se enfrentaron con la policia [ los alumnos].
   ‘They confronted the police with the students.’

a. Ficarem el ganivet [ al calaix].
   ‘We will put the knife in the drawer.’
   (2)

b. *Ficarem [ el ganivet] al calaix.
   ‘*We will put the knife in the drawer.’

c. Hi ficarem [ el ganivet], al calaix.
   ‘We will put the knife in the drawer.’

Ficarem el ganivet [ el ganivet].
   (3)

Other strategies such as focus fronting (4), clefting (5) or focus marked prosodically in situ (2b) seem to be restricted to contrastive meaning (Solà 1990, 2002, Quer 2002, Vallduví 1991, 1992, 2002; Zubizarreta 1998, 1999, Moreno Cabrera 1999).

[(cf Un Mercedes) s’ha comprat el Jordi. ‘Jordi has bought a Mercedes’. (4)]

[(Sou [ vosaltres] que ho sabeu. ‘It is you who knows it’. (5)]

Interestingly, what most of these studies have in common is the fact that they make use of introspection and grammaticality judgments, and tend to concentrate on standard varieties of Catalan and Spanish. Recent empirical studies on the intonation of focus in Catalan and Spanish cast doubt on claims made by Vallduví (1991, 1992) and Zubizarreta (1998). Gabriel et al. (2009), Gabriel (2010), Muntendam (2009, 2013), Leal-Méndez & Shea (2012), Hoot (2012a,b) and Vanrell & Fernández-Soriano (2013; V&FS) have shown that neutrally focused elements can also remain in situ. Gabriel et al. (2009) and Gabriel (2010), for example, conducted a production test based on semi-spontaneous speech with speakers from Buenos Aires and Neuquén (Argentina). They show that neutrally focused subjects are realized in situ (i.e. [S]VO) in nearly 75% to 95% of the cases. Only a very small number of utterances show word order variation due to p-movement (< 14% in Gabriel et al 2009 and < 5% in Gabriel 2010). Muntendam (2009) demonstrates by means of a judgment task that monolingual speakers from the Andean region accepted SVO order with a neutrally focused subject in 100% of the cases (p. 194). They also accepted other constructions, though to a lesser extent. The p-movement driven word order VOS was accepted in approximately 52% of the cases. In Hoot (2012a,b) 22 speakers of Mexican Spanish performed a judgment task, showing that with a neutrally focused subject, the preference for SVO is significantly greater than VOS (p.188ff). V&FS (2013) analyzed semi-spontaneous data from different dialectal varieties of Catalan and Spanish obtained through a production test. They found that the main prominence can fall on clause-initial position in Eastern Catalan (Central and Balearic Catalonia) and Basque Spanish neutral focus declaratives or remain in situ in both neutral and contrastive focus declaratives (especially in Valencian Catalan or Spanish). All in all, these studies provide experimental evidence that when looking at non-standard varieties, the distinction between word order and intonation focal typology can be too rigid, since the ‘word order languages” Catalan and Spanish can use two different mechanisms, word order and intonation, to different degrees (cf. Face & D’Imperio 2005).

This paper, in a first step, experimentally investigates the manner in which prosody interacts with syntax in the expression of neutral and contrastive focus in Majorcan Catalan and Madrid Spanish, and shows that the data broadly support the conclusions made by Vallduví (1991, 1992) for Catalan and Zubizarreta (1998, 1999) for Spanish. In a second step, we propose a stochastic optimality theoretic (SOT) approach to account for the syntactic and prosodic focus realizations in the two lan-
The intended contribution of this research is both methodological and theoretical in nature. From the methodological perspective, it examines the same varieties (Eastern Catalan and Castilian Spanish) previously explored solely on the basis of linguistic intuition (Vallduví 1991, 1992 for Catalan and Zubizarreta 1998, 1999 for Spanish). Additionally, important factors such as focus type (broad, narrow neutral and narrow contrastive) or the constituent under focus (subject, direct object or indirect object/adjunct) are included in the design of the experiment. As far as we are aware, this constitutes the first time that these two typologically related languages, Catalan and Spanish, are carefully compared with respect to the syntax-prosody interface in a single study. As for the theoretical perspective, thus far no (optimality) theoretic model of focus exists for Catalan, whereas various proposals are available for Spanish (e.g. Gutiérrez-Bravo 2002 or Gabriel 2010). However, these approaches encounter difficulties in either accounting for the attested variation (Gutiérrez-Bravo 2002) or in integrating all possible patterns such as cleft-constructions (Gabriel 2010). In this paper, we circumvent these shortcomings by proposing an approach couched in the framework of SOT (Boersma & Hayes 2001), which accounts for the attested variation including cleft-sentences.

2. The experimental data

2.1. Methodology

The participants in our production experiment were seven Majorcan Catalan speakers between the ages of 25 and 40 (mdn = 33) and four Madrid Spanish speakers between the ages of 20 and 29 (mdn = 20). The data were obtained by means of a production test designed to elicit different focus constructions (broad, narrow neutral and narrow contrastive foci on different constituents) through question-answer pairs from short picture stories presented in a PowerPoint slide show (along the lines of those in Gabriel et al. 2009, Gabriel 2010). The short stories correspond to full sentences with a canonical syntactic structure (SVOO or OVO/Adjunct) and were controlled for the focused constituent (S, V, OOO and OOO).1 Participants were asked to respond to a series of wr-questions or tag questions and were explicitly asked to use all of the constituents that appeared in the short stories. (6) illustrates a short story and (7) the corresponding questions to which the participants were asked to reply spontaneously.

Blancanieves se llevó las manzanas con fatiga. (6)
Snow.White bring.PAST.3SG with tiredness

a. ¿Qué ha pasado? (7)
what have.PRES.3SG happened
b. ¿Qué trajo Blancanieves con fatiga? (8)
what bring.PAST.3SGSnow.White with tiredness
c. ¿Quién trajo las manzanas con fatiga? who bring.PAST.3SG the apples with tiredness
d. Blancanieves trajo con fatiga las manzanas, ¿verdad? Snow.White bring.PAST.3SG with tiredness the oranges right

e. ¿Cómo trajo las manzanas Blancanieves? how bring.PAST.3SG the apples Snow.White
f. Trajo las manzanas con fatiga Caperucita, ¿verdad? bring.PAST.3SG the apples with tiredness Little.Red.Riding.Hood, right

g. ¿Qué hizo Blancanieves con fatiga? what do.PAST.3SGSnow.White with tiredness
h. Blancanieves trajo las manzanas con vitalidad, ¿no? Snow.White bring.PAST.3SG the apples with vitality, no

Through this method, we elicited a total of 1680 contours for Catalan (24 short stories x 10 questions x 7 speakers) and 1056 for Spanish (24 short stories x 11 questions x 4 speakers). The task took approximately 30 minutes to complete. Speakers were recorded on a Zoom H4n digital audio recorder using and AKG C520 condenser microphone. The sentences were digitized at a sample rate of 44100 Hz with 16 bit amplitude resolution. The data were then annotated in Praat (Boersma and Weenink 2013) according to the orthographic transcription, syntactic strategy used by the speaker, syntactic order, focus type as well as the focused constituent and the prosodic transcription using the Cat_ToBI and Sp_ToBI systems (Prieto et al. in press, Hualde and Prieto in press).

2.2. Results

Table 1 shows the frequency with which the variable SYNTACTIC STRATEGY was observed for Catalan and Spanish in neutral focus declaratives. The results show that an important distinction is made between Catalan, which uses dislocation of the non-focused material or even fronting of the focused material, and Spanish, which resorts to p-movement and clefting. The clefting strategy is also used in Catalan, though it is limited to focused subjects. Table 1 only shows the main syntactic strategies. Here, the frequencies fail to add up to 100% due to further strategies which were used very infrequently.

Table 1: Frequency of the variable SYNTACTIC STRATEGY in neutral focus declaratives.

<table>
<thead>
<tr>
<th>CAT</th>
<th>SPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>[S]</td>
<td>[S]</td>
</tr>
<tr>
<td>Neutral/Fr 42%</td>
<td>Clefing 71.1%</td>
</tr>
<tr>
<td>Clefing 34.6%</td>
<td>P-movement 14.5%</td>
</tr>
<tr>
<td>Left-Disl. 16%</td>
<td>P-movement 47.9%</td>
</tr>
<tr>
<td>Clefting 34.8%</td>
<td>Clefing 23.3%</td>
</tr>
<tr>
<td>Fronting 34.8%</td>
<td>Clefting 23.3%</td>
</tr>
<tr>
<td>Right-Disl. 15.7%</td>
<td>Neutral 43.6%</td>
</tr>
<tr>
<td>Fronting 23.6%</td>
<td>Clefting 21.3%</td>
</tr>
<tr>
<td>Left-Disl. 16%</td>
<td>Right-Disl. 13.2%</td>
</tr>
</tbody>
</table>

Table 2: Frequency of the variable SYNTACTIC STRATEGY in contrastive focus declaratives.

<table>
<thead>
<tr>
<th>CAT</th>
<th>SPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>[S]</td>
<td>[S]</td>
</tr>
<tr>
<td>Clefing 46.6%</td>
<td>Clefing 61.4%</td>
</tr>
<tr>
<td>Left-Disl. 18.4%</td>
<td>Fronting 15%</td>
</tr>
<tr>
<td>Right-Disl. 18.4%</td>
<td>Clefting 61.8%</td>
</tr>
<tr>
<td>Fronting 23.6%</td>
<td>P-movement 15.7%</td>
</tr>
<tr>
<td>Left-Disl. 23.7%</td>
<td>Fronting 23.7%</td>
</tr>
<tr>
<td>Right-Disl. 22.5%</td>
<td>Neutral 14.4%</td>
</tr>
<tr>
<td>Fronting 30.7%</td>
<td>Clefting 41.2%</td>
</tr>
<tr>
<td>Right-Disl. 19.3%</td>
<td>Fronting 23.7%</td>
</tr>
<tr>
<td>Neutral WO 14.4%</td>
<td>Neutral WO 14.4%</td>
</tr>
</tbody>
</table>

Table 2 reports the observed frequency of the variable SYNTACTIC STRATEGY for Catalan and Spanish in contrastive focus declaratives. Both languages primarily use cleft constructions to express contrastive narrow focus on a given constituent. The languages differ, however, with respect to the following most common strategies: Left/right dislocation seems to be a strategy restricted to Catalan, while p-movement is restricted
to Spanish. Focus fronting is used in both languages, but it is limited to OIO/adjunct in Catalan.

### 3. Accounting for focus realizations

#### 3.1. Stochastic Optimality Theory

In order to account for our data, we applied the SOT (Boersma & Hayes 2001), a version of the traditional Optimality Theory (Prince & Smolensky 1993/2004). Traditional OT assumes grammars of individual languages to be a strictly ranked set of violable constraints \( C(C_1 \gg C_2 \gg \ldots \gg C_n) \). SOT, in contrast, ranks the constraints along a continuous scale. As such, the distance between the constraints can vary. Furthermore, SOT interprets constraints as ranges of (ranking) values and not as single points. As a consequence, the constraint ranges overlap when the distance between two constraints is short. This means that in a certain percentage of the evaluations (depending on the amount of overlap), a lower-ranked constraint is preferred over a higher-ranked constraint, and the suboptimal candidate wins. As a result of this mechanism, we are able to model the attested (frequency-dependent) differences in the realization of focus in Spanish and Catalan. For space reasons, we do not propose ranking values, i.e. we do not apply the full SOT apparatus - even though our data have frequency distributions.

#### 3.2. The input

Following Grimshaw (1997) and Kager (1999:344ff), we assume that the input consists of (a) a lexical head and its arguments, (b) an assignment of lexical heads to its arguments, (c) information on tense, and (d) semantically meaningful auxiliaries. The input also includes a specification of the information structure through the marking of the focused element(s) (Féry 2013, Destruel 2013). We assume that “each analysis of the input competes with other analyses of the same input”, meaning that no element is removed from the input (see the principle of Containment, Prince & Smolensky 1993/2004). Furthermore, we assume that competing candidates generated for a single input by GEN must be not only semantically equivalent (Grimshaw 1997, Kager 1999), but must also be equivalent with respect to focus domain and the type of focus (F vs. CT). This means that all competing candidates must have non-distinct focal representations. The input for a sentence such as *María se bebe un té ‘Maria drinks tea’* is \( \{ \text{beberse} (x, y), x = \text{María}, y = \text{un té}, \text{tense = present}, F = \text{un té} \} \). Functional elements that have no semantic content are not introduced, but are introduced by GEN (Grimshaw 1997, Kager 1999). This is the case for the complementizer que ‘that’ or the copula ser ‘be’. Thus, *María se bebe un té ‘Maria drinks tea’* and the corresponding cleft-construction (\( \text{Es un té que se bebe María ‘It is tea that Maria drinks’} \)) have the same input.

#### 3.3. A proposal for Spanish and Catalan

Our model relies on six constraints. (A) **StressFocus** (SF) requires that the focus be realized with main stress (see Truckenbrodt 1995:10). (B) **Head-IP** (H-IP) demands that the main stress occurs in the right-most position in its intonational phrase (Truckenbrodt 1995). (C) **Subject** (SBJ) requires that the highest A-specifier be filled by a subject (Grimshaw & Samek-Lodovici 1998). This constraint is violated when Spec;TP is not filled by a subject. (D) **FaithSyntactic** (FaS) requires that no syntactic material be added to the input (Destruel 2013:208). This constraint is violated in the case GEN introduces additional material (as in the case of clefts). (E) **P-Movement** (*P-M*) militates against p-movement, i.e. against realizing material below C, T, and V that does not belong to the verbal chain C-T-V (adapted from Gabriél's 2010 Stay-D). (F) **FocusCleft** (FCL) requires a focus element to be clefted (Destruel 2013:200). This constraint is not active in all-new / broad focus contexts, as more than one constituent is part of the focus domain. The underling constraint hierarchy is given in (8). The interaction between the six constraints is illustrated for Spanish narrow neutral focus on both the subject in Table 3 and the direct object in Table 4.

#### Table 3: Normal (upper table) and reverse ranking (lower table) for neutrally focused subjects in Spanish.

<table>
<thead>
<tr>
<th>Subject = neutral focus</th>
<th>SF</th>
<th>H-IP</th>
<th>SBJ</th>
<th>FaS</th>
<th>*P-M</th>
<th>FCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORMAL W0</td>
<td>*!</td>
<td></td>
<td></td>
<td>#!</td>
<td></td>
<td>*!</td>
</tr>
<tr>
<td>*P-MOV</td>
<td>#!</td>
<td></td>
<td>#!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLEFT</td>
<td></td>
<td></td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RD</td>
<td></td>
<td>#!</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Se lo da a Juan María, el libro.</td>
<td>*!</td>
<td></td>
<td></td>
<td>#!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As illustrated in Table 3, the ranking in (8) accounts for the cleft construction, the most frequent pattern. While the normal word order fatally violates **Head-IP**, the constructions with p-movement and left- and right-dislocations fatally violate **Subject**. Since cleft sentences are characterized by an IP break at the right edge of the clefted element (Féry 2013, Destruel 2013), **Head-IP** is not violated. We assume that the three constraints **Subject**, **FaithSynt**, and *P-MOV* overlap (see blue box in Table 3). The overlap occurs to such a degree that the frequency-dependent variation is accounted for. When FaithSynt outranks the other two constraints, the second most common realization, namely p-movement, wins (lower table). When *P-MOV* outranks the other two constraints, the winning candidate is again the cleft construction.

When the direct object is neutrally focused, the order of the most common patterns differs from that for the subjects (Table 1). In this case, p-movement is most common, followed by clefts. Again, the ranking in (8) and the proposal for the overlapping constraints easily account for these patterns (see Table 4). P-movement is the most common pattern, because it violates neither **Subject** nor **Head-IP** or **FaithSynt** upper panel. Actually, no candidate violates **Subject**, because their subject position is filled. Thus, the effect of **Subject** observed in the condition with narrow focused subjects does not appear to exist for objects. The ranking for clefts, the second most common candidate, is given in the lower table. Here, *P-MOV*, which is fatally violated by the construction with p-movement, outranks FaithSynt. The decision between the cleft-construction and the dislocation structures is made by the lowest ranked constraint FocC, which is fatally violated by the dislocations.

For space reasons, we only briefly comment on the analysis of Catalan and the contrastive focus condition in the two languages. For Catalan, the results suggest that the constraint *P-MOV* is ranked high, due to the relatively few instances of this
4. Discussion and conclusion

Our data for Spanish partially support Zubizarreta’s(1998) claims about p-movement to mark neutral focus in Spanish and focus fronting being restricted to a contrastive interpretation. However, our results also diverge from those of Zubizarreta (1998) in that clefting was the most common choice made by our Spanish participants in the contrastive focus condition. Furthermore, our results contradict the findings of Gabriel et al. (2009), Gabriel (2010) and Hoot (2012a,b), as a prosodic marking of focus in situ was realized only very rarely in Spanish. This contradiction comes as a surprise, as we used the same methodology as in Gabriel et al. (2009) and Gabriel (2010). One possible explanation could be that the other experimental studies concentrated on Latin American Spanish and not on Peninsular Spanish, as we did. Thus, there may be an important difference between the geographical varieties, confirming the important role of dialectal variation in intonational phonology (Prieto & Roseano 2010, Feldhausen et al. 2010, V&FS 2013). As for Catalan, our results strongly support the claim made by Vallduví (1992) that dislocating non-focal material is a very important strategy in this language. P-movement, in contrast, is almost non-existent. This finding contradicts claims made by Dominguez (2002), who argues that p-movement is also productive in Catalan. One must bear in mind, however, that Dominguez (2002) investigated Valencian Catalan, among other varieties, and not Majorcan Catalan. This also suggests that dialectal variation must be taken into account as a decisive factor involved in the variation of focus realization strategies (cf. V&FS 2013). The comparison of Catalan and Spanish reveals that important differences exist between the focus realizations of these languages. Most importantly, dislocations are much more common in Catalan than in Spanish, with p-movement only representing a common strategy in Spanish. The proposed optimality theoretic approach is able to account for the variation attested in the data as well as for cleft constructions. It thus offers clear advantages over the approaches proposed by Gutiérrez-Bravo (2002) and Gabriel (2010). Furthermore, the present approach also attempts to account for language-specific differences. Here, its scope is limited to the differences between Spanish and Catalan: the constraints DISLOCATEGIVEN and *P-MOV are ranked high in Catalan, while they are low or inactive in Spanish. This reflects the previously described differences in strategies used to realize focus in the two languages.

5. Acknowledgements

Our gratitude goes to Shinichiro Ishihara for his crucial help with the analysis. Special thanks are also due to Francesc Torres-Tamarit and Silke Hamann as well as to the audience of the 2014 OCP Conference in Leiden/Amsterdam, Netherlands, for their helpful comments and questions. Thanks also to Audrey MacDougall for her assistance with editing.

6. Selected References


On the use of accelerometer sensors to study nasalization in speech and singing voice

T. Fux¹, A. Amelot¹, L. Crevier-Buchman¹,², C. Pillot-Loiseau¹, M. Adda-Decker¹

¹Laboratoire de Phonétique et de Phonologie, CNRS, UMR7018, Paris, France
²Hôpital Européen Georges Pompidou, Univ. Paris-Descartes, Paris, France

thibaut.fux@hotmail.fr

Abstract

This paper presents first results of a study aiming to explore data coming from nose mounted accelerometer during speech and singing tasks. One objective was to study the variations in the piezoelectric signal under variable speech and singing voice productions. Thus, only high-pitch and high-level singing are considered in this study. Four speakers (2 males, 2 females) produced isolated vowels, CVC and VCV non-words in nasal and non-nasal consonantal contexts. Our results suggest that the discrimination of nasal consonants remains possible in singing voice. A second part of this study investigates the correlation between acoustic and piezoelectric signals in vocalic sounds. A relative stable transfer function, with a frequency dip at low frequency around 500 Hz could be measured in our data. Results highlight a relative stable transfer function between audio and accelerometer signal for the vowels.

Keywords: accelerometer, nasalization, singing voice, transfer function

1. Introduction

The proposed work is a pilot study which aim at investigating the usefulness of a nose mounted accelerometer to study its response in different contrastive contexts: oral vs nasal phonemes and speech vs singing. Nasal accelerometers have proven useful to locate a nasal sound production in a continuous spoken speech signal (Stevens, Kalikow, and Willemain 1975; Horii 1980; Lippman 1981; Brkan, Amelot, and Pillot-Loiseau 2012). The captured signal corresponds to skin vibrations during phonation which are related to the airflow through the nasal cavities and possibly to bone conduction. For example, the signal of the nose accelerometer enables the location of a nasal consonant in an oral vowel VCV context in modal speaking voice due to an increased intensity in the nasal consonant. Stevens, Kalikow, and Willemain (1975) found intensity differences of 10-20 dB between nasal consonants intensity and oral vowels intensity in the accelerometer signal but only 10 dB with high vowels such /i/. Even if the accelerometer method has already been explored for phonetic studies in speech, it remains important to study its behavior in singing voice. Do the observations done for speech still hold for singing voice? In particular, what happens if sound is produced with high intensity and relatively high pitch?

The spectral composition of the signal of a nose mounted accelerometer was studied by Tronnier (1994) (see also (Tronnier 1998). Results show that the formant structure of oral vowels is not necessarily preserved in accelerometer data whereas the structure of nasal vowels remains almost unchanged. Trying to better understand the accelerometer signal intensity variations across different vowel types is a second part of this study aiming to explore the variation between audio spectrum and accelerometer spectrum. To this end, we propose to compute a transfer function as the ratio between the acoustic and accelerometer spectra in spoken and singing voice.

2. Database

To check the relevance of nose mounted accelerometers (also named contact microphone), which are piezoelectric sensors, for future studies of nasality and nasal vibrations in singers, we designed a set of production experiments making use of contrastive phonetic contexts. Subjects produced the designed material in both speaking (modal voice) and singing conditions. The collected data include speaking and singing of 4 non-professional singers (2 males, 2 females) recorded as follows. The subjects were placed in a soundproof room and asked to pronounce nine French oral vowels (/a/, /æ/, /ʌ/, /ɔ/, /ʊ/, /ɛ/, /ɛ/, /ʌ/, /u/) and 9 CVC and 9 VCV non-words composed of the three cardinal French vowels (/a/, /ɛ/ and /u/), 2 nasal consonants (/n/ and /l/) and one occlusive consonant (/b/). Note that in a given word the two embedding consonants/vowels are identical. Recordings were carried out using a head mounted microphone (model C520/L from AKG) and a nose mounted dual accelerometer i.e. model Twin spot from K&K sound. The accelerometer was connected to a pre-amplifier and recorded, simultaneously with the microphone at 44100 Hz. The accelerometer sensors were double taped on the nose at position 6 described in (Lippman 1981). To further secure their positions surgical tape was used to fix wires on the cheeks. Both signals were then resampled at 16 kHz.

2.1. Database overview

Table 1 describes the captured data in terms of average F0 values, for the spoken and the sung material for the four subjects (F1, F2, M1, M2). The values are determined by considering only the isolated vowel productions in order to give an idea of the difference between our spoken and sung voices. Note that the production of non-words showed similar evolutions. Intensities of isolated vowels were measured in the middle of each segment (Larson and Hamlet 1987). Table 1 shows differences in average intensities between spoken and sung vowels.

3. Measurements

For the presented experiments two kinds of measurement were made. The first one is the measurement of intensities of accelerometer signal for consonants and vowels. The second one
Table 1: Average values of F0 for all subjects as measured on all isolated vowels. Bottom lines show F0 (in semitones) and intensity differences between singing and spoken productions as measured from the audio signal.

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>M1</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_{0\text{spoken}} ) (Hz)</td>
<td>202</td>
<td>207</td>
<td>96</td>
<td>123</td>
</tr>
<tr>
<td>( F_{0\text{singing}} ) (Hz)</td>
<td>355</td>
<td>369</td>
<td>196</td>
<td>343</td>
</tr>
<tr>
<td>( \Delta F0 ) (semitones)</td>
<td>9.8</td>
<td>10.0</td>
<td>12.3</td>
<td>17.7</td>
</tr>
<tr>
<td>( \Delta I ) (dB)</td>
<td>7.8</td>
<td>15.6</td>
<td>8.4</td>
<td>5.6</td>
</tr>
</tbody>
</table>

is the microphone to accelerometer transfer function, i.e. the relation in the frequency domain between the audio spectrum and the piezoelectric spectrum.

3.1. Accelerometer signal intensity

The recorded material was manually segmented into C and V segments both for speech and singing voice. The intensity was measured using a 25 ms window length. Intensity measurements were carried out as follows: for the first (second, third) phoneme of the CVC/CVC patterns the measurements were carried out at 2/3 (1/2, 1/3 respectively) of their durations. We defined an intensity ratio (in linear scale) as the ratio between the intensities of a consonant and the neighboring oral vowel measured using the accelerometer signal. The presented intensity ratios are measured only for the first couple of consonant-vowel for each non-sense word (C vs V1 for V1VC2, and C1 vs V for C1VC2).

In fully oral CVC/CVC productions, the intensity of the accelerometer signal remains low both for vowel and consonant segments. If the CVC/CVC pattern includes a nasal consonant, the intensity tends to raise in the accelerometer signal, precisely in the region including the nasal consonant. An increase of the \( I_{VC} \) intensity ratio thus signals the presence of nasality or at least nasalization. In singing voice, vocal techniques may induce nasalization without the presence of nasal segments.

3.2. Transfer function

The transfer functions were measured using exclusively isolated vowels. The spectrum of the audio and the piezo signals were computed in the middle of the vowels using a Blackman-Hanning shaped window of 1024 points (64 ms). The principle of the transfer function measurement is to divide the output (piezo) by the input (audio) in the frequency domain with linear amplitude. However, periodic speech signals include low energy components between the harmonic peaks of the spectrum. This may generate a bias in the transfer function computation. A solution consists in splitting the spectrum into small frequency bands of harmonic and non harmonic regions. However, the size of the frequency band may be chosen in accordance to the F0 value. In this paper we preferred to use only the harmonic amplitude differences. Thus, the first step, for each vowel, is the estimation of F0 and then, using a peak picking algorithm, the location of harmonic peaks in the spectrum. We chose to limit this peak picking to the region between F0 and 3 kHz. Indeed, above 3 kHz the harmonics are often hard to identify due to low harmonic-to-noise ratios in the higher frequency regions of speech signals.

This harmonic peak picking has been done in both the audio and the piezo spectrum in order to compute the amplitude difference of their respective harmonics (i.e. the audio-to-piezo transfer function). This generated discrete transfer functions of various lengths related to the initial F0 values. Thus, to average multiple occurrences, transfer function curves were interpolated in the frequency domain to obtain the same number of points. This allows us to determine a mean transfer function for each subject and voiced production types (speech and singing) using all the isolated vowel occurrences.

4. Results

This section presents various results evaluating the discrimination potential between nasal and oral productions, using an accelerometer sensor, in both the singing voice and speech. Furthermore, we also examined the frequency composition of the piezoelectric signal as compared to the audio.

4.1. Piezo Intensities: non-words, speech vs singing

Figure 1 shows average \( \frac{I_{VC}}{I_{TH}} \) intensity ratios over the whole database. The figure compares two conditions: speech (left) and singing voice (right). Each condition gives contrastive results for nasal vs oral consonants. As expected, average results show positive intensity ratios for nasal consonants and negative ones for oral consonants in both speech and singing conditions with even stronger tendencies for singing voice. Results thus confirm the potential of accelerometers to locate nasality in singing voice. However, the high standard deviations in Figure 1 signal important variabilities. Future studies should better control for different factors giving rise to variation.

Figure 2 shows the average ratio values of the three repetitions merged according to the consonantal context and vowel type. The first observation is for the vowel contexts of /l/ and /u/ where amplitudes of nasal consonants are less salient. This is a well known and frequently observed effect, due to a larger increase of the accelerometer signal for high vowels than for low vowels (/a/). It may be related to a higher oral impedance and a larger surface of palate exposed to the airflow. This may favor the propagation of vibrations (Bundy and Zajac 2006; Gildersleeve-Neumann and Dalston 2001). Observations per speaker show that the differences between nasal consonants and vowels are not necessarily consistent among speakers between spoken and sung tasks. For subject F2, a semi professional singer, it is interesting to observe that higher ratios are obtained for the /a/ context in speech, but this result is reversed in singing voice. Similar variable results were found by Chen, Ma, and Yiu (2014) who measured more intense accelerometer signals for /l/ and /u/ than for /a/ for singers trained to resonant voice techniques.

Cases where the nasal consonants are less prominent than
the neighboring vowels in piezoelectric signals could be observed. This may be explained by the fact that vowel acoustic intensities generally increase more than those of consonants when moving from spoken to singing voice. When focusing on the isolated vowel material we could confirm that an acoustic intensity increase entails an increase in the piezoelectric intensity. However, acoustic/piezo intensity ratios are not constant with globally increasing production intensities. Results suggest that the piezoelectric signal is not only influenced by the overall production intensity, but that other factors may come into play. Among these are various articulation and voice production strategies adopted by the speaker or singer. It may also be related to the measurement method.

4.2. Audio-to-piezo transfer functions: toward a piezoelectric signal prediction for oral production.

In this section, we address the question of how piezoelectric signal intensities relate to the acoustic ones in different frequency bands. By comparing the spectral compositions of the piezoelectric and audio signals as a function of vowel type, we derive audio-to-piezo transfer functions which remain comparable for all the vowels (at least until 1.5 kHz). More interestingly, these transfer functions remain almost the same for spoken and singing tasks. Figure 3 shows transfer functions averaged over all isolated vowel productions for two speakers (F2 and M2). Similar shapes are obtained with a dip around 500 Hz and they remind transfer functions of bone conduction as shown in other studies (Won and Berger 2005). Obviously, transfer functions are speaker dependent and also depend on experimental parameters such as positioning and fixing of the sensors (accelerometer and microphone). Thus, piezoelectric intensity not only depends on the acoustic intensity but also on the spectral composition of the produced sound.

To further investigate this question, figure 4 shows three glissandi of /a/, /i/ and /u/ vowels produced by a semi-professional singer. The top panel corresponds to audio data, the bottom panel to piezo recordings. For each sensor, F0 and intensity (I) curves are provided. It is interesting to compare the relative shapes of F0 and I curves in piezo and audio data.

For /i/ and /u/ vowels, the I level at the beginning and the end of glissandi are relatively constant (not increasing) while audio intensity increases. In the middle of the /i/ and /u/ glissandi, the piezo I shape tends to follow the audio I shape. Different hypotheses may be proposed to explain this observation. A first hypothesis relates to the influence of harmonics and first formant positions around the “valley” zone of the transfer function, which in turn, influences the accelerometer intensity signal. Another hypothesis is related to the passage from production mechanism 1 to mechanism 2.

5. Discussion

Accelerometer sensors are known to be useful for nasal production studies. If the objective is the localization of a nasal (or nasalized) sound in a continuous accelerometer signal, the main difficulty is the determination of a threshold to discriminate oral from nasal productions. Moreover, it appears to be difficult to predict the intensity of the piezoelectric signal for purely oral productions. High vowels are often considered to have higher piezoelectric intensity but our first results show no significant relation between vowel type and piezoelectric intensity.

Many points can be improved with respect to the recording procedure and must be tested in order to check their influ-
ence on the signal. First, the manner to position and fix the accelerometer may influence the frequency response of the sensor. A further bias may be due to velum and head movements, especially in singing. The use of multiple accelerometers to apply movement corrections can be envisioned. Finally, the accelerometer used in this study is a double accelerometer (one of each nose side) wired in parallel to obtain a single signal. If the positions of both the sensors are not exactly the same, the two accelerometers will not necessarily produce similar signals and will not necessarily be in phase. Furthermore, signal asymmetries may arise from asymmetric nose morphologies. Separate analyses of the two sensors should be studied to be sure there is no phase influence. A better control of all these factors should increase the reliability of the measurements and thus result in more reliable relations between accelerometer responses and oral productions. A more precise knowledge of this relation or transfer function will hopefully contribute to establish meaningful and robust thresholds for nasalization measurements and related studies.

6. Conclusion

The presented study investigated the use of accelerometer sensors in speech and singing, focusing on isolated vowels and nasal vs oral consonants in oral vowel contexts (CVC and VCV patterns). We made use of two types of measurements. First, an intensity $\frac{I_{ac}}{I_{ac}}$ ratio using the accelerometer signal was used to compare spoken and sung productions of the CVC and VCV patterns. Secondly, piezo vs audio spectrum intensity ratios were computed to produce corresponding transfer functions. Such transfer functions will contribute to give more precise interpretations of the observed variations in the piezo signals. Results concerning the $\frac{I_{ac}}{I_{ac}}$ ratio suggest that nose mounted accelerometers can be used in singing voice to locate nasal consonants since they remain more energetic than oral vowels (as nasal consonants also tend to do in speech). Secondly, the result of the piezo/acoustic spectrum ratios (or transfer functions) suggest that the spectral composition, in particular the phonetic type of the sound, affects its piezo intensity and that this should be taken in account when using a piezoelectric intensity normalization method such as the HONC method (Horii 1980). Some authors suggested a piezoelectric intensity F1 dependence, explained by physiological phenomena (Stevens, Nickerson, et al. 1976). The achieved results are consistent with this finding, and further suggest an F1 dependence with the sensor frequency response. To further explore these different hypotheses, future experiments will include a larger set of singers and more controlled productions.

7. Acknowledgments

This work was partially funded by the European FP7 i-Treasures project (Intangible Treasures - Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures FP7-ICT-2011-9-600676-i-Treasures). It was also supported by the French Investissements d’Avenir - Labex EFL program (ANR-10-LABX-0083).

8. References


Lexical and Surface Geminates in Maltese

Luke Galea, Martine Grice, Anne Hermes, Doris Mücke

Abstract

This study investigates the acoustic realization of word initial lexical and surface geminates in Maltese. Word initial gemination in Maltese occurs through a morphological process in both Semitic and non-Semitic verbs. Surface geminates arise through the assimilation of the definite article before coronal sounds (il → it in it-tama). In these surface geminates a bona fide vowel always precedes. In terms of duration, the singleton to geminate ratio is 1:1.6. Surface geminates are on average 7% shorter than lexical geminates. Lexical geminates are consistently preceded by a vocalic element, also of an /i/ quality, when the preceding word ends in a consonant. When the preceding word ends in a vowel, this vocalic element was also found (contrary to predictions from Maltese grammars), but is subject to within and across speaker variation.

Keywords: lexical and surface geminates, Maltese, epenthetic vowel, vocalic element, bona fide vowel

1. Introduction

Geminates are present in both Semitic and Romance languages. Although they are described phonetically as long consonants (Ladehoff and Maddieson 1996), they differ from singleton consonants along various acoustic and articulatory dimensions (c.f. Gili Fivela et al. 2007, Zeroual, Hoole & Gafos 2008, Ridouane 2010), inter alia the duration of constriction and burst characteristics as well as contact duration, virtual target position and amplitude (Löfqvist 2005). Furthermore, some languages, e.g. Italian, have compensatory shortening of the vowel before geminates (Esposito & Di Benedetto 1999), whereas others, e.g. Finnish and Japanese, do not (Lehtonton 1970, Port et al. 1987). The representation of geminates is also a controversial issue in phonological theory, especially in relation to subsyllabic constituents (Davis 2011). In particular, it is unclear whether lexical geminates should be analysed in the same way as surface geminates resulting from morphological processes.

Maltese consists of three language strata: a supra-stratum of Semitic which entails the basis of the phonology, morphology and lexicon, a Romance sub-stratum which adds to lexis and syntax, and an English sub-stratum which Maltese borrows heavily from. In Maltese, as in Arabic, geminates can occur in word initial position (see (i)). In Semitic words, gemination is obtained by a morphological process creating passivized and reflexive verbs, while non-Semitic words, which typically do not fit the root-and-pattern template in the language, undergo word initial gemination as well (see (ii) and (iii)):

(i) ddadar ‘to be disgusted’ (from Semitic: Arabic ‘dardar’, ‘muddy the water’)
(ii) ttimbra ‘to stamp’ (from Romance: Italian ‘timbrare’, ‘to stamp’)
(iii) ttargiţja ‘to target’ (from English ‘to target’)

In Maltese grammar books (c.f. Borg and Azzopardi-Alexander 1997; Mifsud 1995), word initial geminates are described as being preceded by an epenthetic vowel (technically a prosthetic vowel, as it is initial) with the quality of /i/, and with an unclear phonological status. Hoberman and Aronoff (2003:73) argue that this vocalic element is always present “unless the preceding word in the same phonological phrase ends in a vowel”. However, this is not based on acoustic data.

Moreover, Maltese has surface geminates that are derived through assimilation, in the case of the definite article, that is il- (iv), except before coronals, in which case it is assimilated (see (v) and (vi)). Here there is a bona fide vowel /i/.

(iv) il-belt ‘the city’
(v) id-dar ‘the house’
(vi) it-tama ‘the faith’

This study compares acoustic segment durations for word initial singletons and geminates (lexical geminates of Semitic and non-Semitic origin, and surface geminates). It also investigates the preceding vocalic element, comparing the bona fide vowel /i/ in surface geminates with the vocalic element /i/ in lexical geminates.

2. Speech Material

A corpus was collected including singletons (S) and three geminate types: lexical non-Semitic (LGnS) and Semitic geminates (LGS) and surface geminates (SG), in word initial position. All words contained a coronal in the target position (/d, t, s, f, z/). See Table 1.

Table 1: Speech material.

<table>
<thead>
<tr>
<th>Singleton</th>
<th>Lexical Non-Semitic</th>
<th>Lexical Semitic</th>
<th>Surface</th>
</tr>
</thead>
<tbody>
<tr>
<td>ddhabhal</td>
<td>‘to insert’</td>
<td>‘to dial’</td>
<td>‘to be entered’</td>
</tr>
<tr>
<td>jama</td>
<td>‘to hope’</td>
<td>‘to type’</td>
<td>‘to have hope’</td>
</tr>
<tr>
<td>szabbar</td>
<td>‘to comfort’</td>
<td>‘to sign’</td>
<td>‘to be consoled’</td>
</tr>
<tr>
<td>szabham</td>
<td>‘to fatten’</td>
<td>‘to be relaxed’</td>
<td>‘to be fattened’</td>
</tr>
<tr>
<td>szarar</td>
<td>‘to fray’</td>
<td>‘to zoom’</td>
<td>‘to be frayed’</td>
</tr>
</tbody>
</table>

The target words, were presented in carrier phrases in two conditions, following a consonant /m/ and following a vowel /w/:

1. *Qalilhom ____ erba’ darbiet* ‘He told them ____ four times’
II. Qalilha ____ erba’ darbiet ‘He told her ____ four times’

By looking at the production of geminates across the two contexts, we can test whether word initial geminate types exhibit different acoustic durations depending on the preceding word (whether it ends in a consonant or a vowel). Following the literature, a vocalic element is expected between a word final consonant and a word initial geminate. We also investigate whether a vocalic element is inserted after a word final vowel (which is not predicted by the literature) and compare its duration across the different contexts.

In the study there were 20 target words, which were repeated 5 times in 2 preceding contexts (n = 200 per speaker). The order of the carrier phrases was randomized. The data from seven (3 males, 4 females) native speakers of Maltese is presented here. The speakers are all speakers of standard Maltese and are all students at the University of Malta (ages: 18-27) residing in Malta. All recordings were conducted in Malta.

3. Acoustic Data: Geminates

The durations measures were taken according to geminate type (i.e. singletons, lexical geminates of Semitic origin, lexical geminates of non-Semitic origin and surface geminates), segments (i.e. /d, t, s, z/) and context (i.e. following /m/ and /v/).

3.1. Acoustic analysis

Acoustic segmentation was carried out using Praat (Boersma & Weenink, 2014). For plosives /d/ and /t/, the whole duration (constriction, burst and aspiration) was segmented. The fricatives, /s, f, z/, were identified from the onset and offset of the friction noise in the spectrogram. The vowel preceding the geminate was measured between the onset and offset of vowel related formant measures. In the vocalic contexts, the duration from the beginning of the word final vowel to the onset of the consonantal constriction in the target word was also measured.

3.1.1. Statistics

All data were analyzed with generalized linear models, using R (R Core Team, 2012) and the package lme4 (Bates, Maechler, Bolker & Walker, 2014), to test the effect of geminate type, preceding context and segment on consonant duration. We included a term for random intercepts for speakers. We included uncorrelated random slopes for the fixed effect geminate type for speakers. P-values were generated using likelihood ratio tests.

3.1.2. Results

There was a significant difference for geminate type ($\chi^2(3)$=38.859, p<0.001), segment ($\chi^2(4)$=700.48, p<0.001) and context ($\chi^2(1)$=76.953, p<0.001). Figure 1 shows the data distribution of the durations of the geminate types post-consonantally. Table 2 shows the average means (ms) of geminate type per segment.

Post hoc tests were conducted, given the significance of this model. Tukey HSD tests, using the multcomp package (Bretz & Westfall 2008), were conducted on geminate type and context. There was a significant difference between singletons and surface geminates (p<0.001, S: $\bar{x}$=109ms, sd=30, SG: $\bar{x}$=173ms, sd=30); singletons and Semitic lexical geminates (p<0.001, LGS $\bar{x}$=182ms, sd=33) and singletons and non-Semitic lexical geminates (p<0.001, LGS: $\bar{x}$=178ms, sd=30).

We expected lexical Semitic geminates and lexical non-Semitic geminates to have different durations but the difference did not reach statistical significance in our model (p=0.421, LGS: $\bar{x}$=182ms, sd=33, LGS: $\bar{x}$=178ms, sd=30). We predicted a difference between lexical and surface geminates, which was confirmed by a statistically significant difference: surface geminates and Semitic lexical geminates (p<0.001, SG: $\bar{x}$=173ms, sd=30, LGS: $\bar{x}$=182ms, sd=33) and surface geminates and non-Semitic lexical geminates (p<0.001, SG: $\bar{x}$=173ms, sd=30, LGS: $\bar{x}$=178ms, sd=30).

Post hoc tests show that geminate type duration is variant depending on the preceding context. Geminate type durations are longer after a vocalic context than a consonantal context (c.f. Table 3 below). The preceding context yielded statistical significance: /m/ - /v/ (p<0.001).

Our data shows that singletons have a shorter acoustic duration than lexical and surface geminates. Lexical geminates of Semitic origin are 9ms longer than surface geminates, while lexical geminates of non-Semitic origin are 5ms longer. Further, our data did not yield significant results between lexical geminates of Semitic and non-Semitic origin. Singletons and geminates have longer durations when preceded by a vowel.
4. What happens before the geminates?

The literature predicts that lexical geminates are preceded by an epenthetic /ɪ/ after a consonant in the same phonological phrase and is absent after vowels (Hoberman and Aronoff 2003). Surface geminates are preceded by the bona fide vowel /ɪ/, which is part of the definite article. In this study we measure the duration of these vocalic elements (i.e. the epenthetic and bone fide vowels, both of which have the quality of /ɪ/) and calculate their frequency of occurrence in both contexts (words ending in a consonant or a vowel).

4.1. Vocalic elements in post-consonantal position

Here we discuss the results for the vocalic element preceding lexical geminates and surface geminates after the preceding context /m/. The /ɪ/ preceding surface geminates is part of the definite article in Maltese, and thus is phonologically specified. In post-consonantal position, /ɪ/ was produced in all repetitions by all speakers before surface geminates. Furthermore, after /m/, the epenthetic /ɪ/, was produced almost always before lexical geminates (in 96% of the total number of repetitions across all speakers).

4.1.1. Results

Figure 3 shows the data distribution of the durations of the vocalic element /ɪ/ before lexical geminates of Semitic origin, lexical geminates of non-Semitic origin and surface geminates in post-consonantal position.

![Figure 2: Duration (ms) of vocalic element /ɪ/ preceding geminate types (singleton (S), lexical geminates of Semitic origin (LGS), lexical geminates of non-Semitic origin (LGnS) and surface geminates (SG)).](image)

Using a generalized linear mixed model as in Section 3, the duration of the vocalic element before surface or lexical geminates did not result in any statistically significant differences ($\chi^2(2)=4.2574$, $p=0.119$).

4.2. Preceding context /ʊ/

The literature (c.f. Hoberman and Aronoff 2003, Misfud 1997) claims that geminates are not epenthized (i.e. there is no epenthetic vowel preceding them) after a vocalic context. Furthermore, some languages, such as Italian, show compensatory shortening before geminates, but others, like Finnish and Japanese do not. In Maltese, the vowel /ʊ/ in /quito/ ‘qata’ (cut in small pieces), is shorter than in /qetto/ ‘qata’ (frighten). It is important to point out, however, that our data involves adjustments to duration across word boundaries, which is not the case in most of the literature cited for compensatory shortening.

Vowel lengthening is observed in most cases before lexical and surface geminates (see Table 4) if we take the whole vocalic portion spanning the word boundary into account. This vocalic portion is measured from the beginning of the vowel /ʊ/ of ‘Qalilha’ to the beginning of the segment for the geminate.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>S</th>
<th>LGS</th>
<th>LGnS</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>102</td>
<td>110</td>
<td>156</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>112</td>
<td>91</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>101</td>
<td>238</td>
<td>197</td>
<td>244</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>76</td>
<td>81</td>
<td>82</td>
</tr>
<tr>
<td>5</td>
<td>81</td>
<td>95</td>
<td>94</td>
<td>101</td>
</tr>
<tr>
<td>6</td>
<td>94</td>
<td>77</td>
<td>78</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>115</td>
<td>112</td>
<td>115</td>
<td>150</td>
</tr>
</tbody>
</table>

This vocalic portion is longer before surface geminates than lexical geminates. However, it is still longer before the lexical geminates than before the singletons.

Speakers may decide to break the hiatus between the word final vowel and the initial vowel of the definite article in surface geminates by inserting a glottal stop in between. Speaker 3 does this consistently. This could be interpreted as a resyllabification of the word, where the glottal stop is an onset, the bona fide vowel /ʊ/ is the nucleus and the geminate is part of the coda of this syllable, as well as the onset to the following syllable.

However, it is interesting to note speaker 3’s high occurrences of the glottal stop and the vocalic element in lexical geminates, which might suggest that, for this speaker, it is phonologically specified as an epenthetic vowel. Some of the longer values e.g. in speaker 3’s productions are due to the glottal insertion included in the proportion measurement. In a vocalic context, speakers can opt to lengthen the word final vowel, epenthesize by either inserting a vocalic element or by inserting a glottal stop and a vocalic element. These phenomena are prone to considerable variation. That is, they are not produced across all speakers and across all repetitions (e.g. 4 out of 7 speakers use glottal stops some of the time in all three types of geminates).

5. General Discussion

Our study shows that singletons have comparatively shorter acoustic segment durations than geminates in Maltese, the ratio being 1:1.6. There were no differences between lexical Semitic and non-Semitic geminates. Surface geminates are slightly (but significantly) shorter than lexical geminates.

Furthermore, in post-consonantal position, the bona fide vowel /ʊ/ is present at ceiling, while the vocalic element is present at 96%. The difference in duration of /ʊ/ before surface and lexical geminates was statistically significant, showing that the bona fide vowel is relatively longer.

An epenthetic vowel is required, by the language’s morphology, between imperfective prefixes and verbal stem consonant clusters (Misfud 1995). In a case like daħal ‘to
enter’, where the verbal stem for the imperfect, is –dħol (sing.)/-dħlu (pl.), results in nidħol ‘I enter’, /ɪ/ is required before the imperfective prefix -n/-, which indicates 1st person singular. Therefore, morphological prefixes and consonant clusters in verbal stems are broken up by an epenthetic vowel. For lexical geminates in the imperfective, the same process is carried out, e.g. the verbal stem –dilħal or –dilājja results in niddahal or niddajja. It is possible that /ɪ/ is somewhat more established in the phonology due to its requirement in these imperfect forms. Furthermore, following Simonović (ms), we can argue that this vocalic element is needed between consonantal (morphological) prefixes in the verbal paradigm and word initial geminates, ensuring they are not the same syllable.

It seems that Maltese shows vowel lengthening before geminates, something which was not found in other languages having geminates. This vowel lengthening might suggest that word initial geminates in Maltese are preceded by a vocalic slot that needs to be filled. This could be filled by a lengthening of the previous vowel or by insertion of an epenthetic /ɪ/. Evidence in favour of this comes from other aspects of verbal morphology, given the presence of an epenthetic vowel between a consonant and a geminate. Thus, in phonological terms, word initial and word medial geminates in Maltese may be structurally identical.

6. References


Simonović M. (ms) Kepp.ing it together and apart from others – Consonant gemination in loanword integration. Utrecht University.

Focus planning during sentence production: An eye-tracking study
L.Y. Ganushchak\textsuperscript{1,2,4}, A. E. Konopka\textsuperscript{3}, & Y. Chen\textsuperscript{1,4}
\textsuperscript{1}Leiden University Centre for Linguistics (LUCL), the Netherlands
\textsuperscript{2}Education and Child Studies, Faculty of Social and Behavioral Sciences, Leiden University, the Netherlands
\textsuperscript{3}Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
\textsuperscript{4}Leiden Institute for Brain and Cognition (LIBC), the Netherlands
a.y.ganushchak@hum.leidenuniv.nl, agnieszka.konopka@mpi.nl, yiya.chen@hum.leidenuniv.nl

Abstract
To produce a sentence, speakers must prepare and linguistically encode a preverbal message. The aim of the study is to investigate sentence planning with a discourse focus. In an eye-tracking experiment, participants described pictures of transitive events. Focus was manipulated by presenting questions before each picture. In the Neutral condition, participants were asked ‘What is happening here?’ In the Subject and Object Focus conditions, the question was Who is catching the bull? and What is the cowboy catching?, respectively. The response was the same in all conditions (The cowboy catches the bull). In the Neutral condition, consistent with earlier work, participants looked first at the subject character (cowboy) and then at the object character (bull) of the sentence. In contrast, in the Subject Focus condition, participants looked longer at the subject character and shifted their gaze to the object character only about 1500 ms after picture onset. In the Object Focus condition, it took only about 500 ms for participants to shift their gaze to the object character after picture onset. The results show that information structure affects the timecourse of linguistic formulation in simple sentences.

Keywords: focus planning, sentence formulation, incrementality, eye-tracking

1. Introduction
To produce a sentence, speakers must first prepare a preverbal message and then encode it linguistically. The planning of messages and sentences normally proceeds incrementally (e.g., Kempen and Hoenkamp 1987), but the amount of linguistic information that speakers prepare in advance of speaking can be highly variable (e.g., planning a single word vs. planning a noun phrase; Konopka, 2012). While much work has been done on formulation of individual sentences produced out of context, a largely neglected area of research is how speech is planned as a function of the discourse context in which an utterance is produced.

The aim of the present project is to investigate the online preparation of utterances within one particular discourse context, i.e. utterances produced in response to questions such as “What is the cowboy doing?” (Figure 1). Questions can affect formulation of responses in two ways. First, they explicitly mark one of characters as old information. Second, they provide a discourse context where a set of possible answers are evoked and the specific new information elicited in the answer is known to be focused (Rooth 1996).

It is well-known that interlocutors often mark focused new information differently in their utterances from given information. For example, in answer to the question of what the cowboy catches, the typical answer that the cowboy catches the bull includes cowboy as given information and bull as new, and therefore, focused information. In contrast, if the question is who catches the bull, cowboy would become focused, indicating that it is the cowboy, rather than, e.g., a person of another profession, who catches the bull. The specific question we address here is to what extent focus may affect the way utterances are planned on-line.

Figure 1: Example of a target event.

To compare formulation of sentences with and without focus, eye-tracked participants were asked to describe pictures shown on a computer screen (e.g., Figure 1). Focus was manipulated by means of questions that preceded each picture. In the Subject Focus condition, participants were asked about the subject (e.g., Who is catching the bull?). In the Object Focus condition, participants were asked about the object (e.g., What is the cowboy catching?). In the Neutral condition, participants were asked ‘What is happening here?’ The target response was expected to have the same structure and content in all conditions (The cowboy catches the bull). Differences in planning of the target responses were evaluated by comparing speakers’ eye movements to the two characters in the picture while preparing to speak.

When describing such events, speakers normally look at characters in the display in the order of mention (Griffin & Bock, 2000; also see Bock, Irwin, Davidson, & Levelt, 2003). On Griffin and Bock’s (2000) account, formulation begins with a short apprehension phase (0-400 ms) during which speakers encode the gist of the event and during which fixations to characters do not differ. Event apprehension is
then followed by a longer phase of linguistic encoding. Viewing times on a character and thus gaze shifts from one character to another during linguistic encoding are expected to vary with the ease of encoding each character: easy-to-name characters are fixated for less time than harder-to-name characters (Griffin, 2004; Konopka, 2014; Meyer & Lethaus, 2004).

Therefore, we predicted the following. If focus does not influence the timecourse of sentence formulation, then viewing times for the subject and object characters should not differ across conditions in the current experiment. However, if focus influences planning from the earliest stages of formulation, then viewing patterns in the Subject/Object focus conditions should differ from the Neutral focus condition. Specifically, speakers should preferentially fixate the character needed to answer the question and direct fewer fixations to the character that was mentioned in the question. Thus, in the Subject Focus condition, speakers should primarily direct their gaze at the subject character in the picture rather than at the object character. In contrast, in the Object Focus condition, speakers need to encode information about the object character, and therefore we should observe more fixations to the object than subject character in the picture.

Importantly, we tested when such differences in fixation patterns across conditions emerge. Differences occurring immediately after picture onset (0-400 ms) would indicate that focus information has an early effect on formulation of the target utterance. Differences across conditions emerging after 400 ms would indicate that focus information influences exclusively the timing of linguistic encoding.

2. Methods

2.1. Basic layout features

Thirty native speakers of Dutch participated in the experiment (24 women; age ranged from 17 to 23 years). All participants were students of Leiden University. Participants gave written informed consent prior to participating in the study and received a small financial reward after the experiment.

2.2. Materials

We used 178 colored pictures displaying simple actions (Figure 1). There were 58 target pictures, 116 fillers, and 4 practice pictures. Focus was manipulated by means of questions that preceded each picture. For the Object and Subject Focus conditions, we asked about the object and subject respectively (see below).

Expected target sentence: *De cowboy vangt de stier* (The cowboy catches the bull)

(1) Neutral question:
*Wat gebeurt hier?* (What is happening here?)

(2) Object Focus question:
*Wat vangt de cowboyl?* (What is the cowboy catching?)

(3) Subject Focus question:
*Wie vangt de stier?* (Who catches the bull?)

All questions were recorded by a native Dutch male speaker and were presented auditorily prior to picture onset.

Three lists of stimuli were created to counterbalance question types across target pictures. Each target picture occurred in each type of the questions on different lists, so that each participant saw each picture only once.

2.3. Design and procedure

Participants were seated in a sound-proof room. They first heard a question and then saw the picture about which they were asked the question. All participants were instructed to describe the picture as briefly as possible using a Subject-Verb-Object structure (e.g., *the cowboy catches the bull*). The task started with four practice trials.

Each target sentence was followed by two filler trials. The fillers consisted of similar pictures. However, the questions preceding filler pictures varied: for instance, the questions asked participants to name a color of an object, or how many of given items were depicted on the picture.

2.4. Data analysis

The timecourse of formulation for sentences in the three conditions was compared with by-participant ($\beta_1$) and by-item ($\beta_2$) quasi-logistic regressions performed on agent-directed fixations (Barr, 2008). We selected three time windows (0-400 ms, 400-1000 ms, and 1000-1800 ms) for analysis. Fixations were aggregated into a series of time bins of 50 ms each for the first analysis and 200 ms for the second and third analysis for each participant and each item in each condition. The dependent variable in each time bin was an empirical logit indexing the likelihood of speakers fixating the agent out of the total number of fixations observed in that time bin. Time and Focus Condition were entered as fixed effects into all models, and all models included random slopes for the Time variable.

Fixations in the three experimental conditions were compared with two contrasts. The first contrast compared the Neutral condition against the Object Focus condition. The second contrast compared the Neutral condition against the Subject Focus condition. Both contrasts thus assess how planning a sentence in response to a question changes the overall distribution of attention to the two characters relative to the focus-neutral condition. Finally, separate analyses were run with new contrasts to compare agent-directed fixations in the Subject and Object Focus conditions against one another.

3. Results

Figure 2 plots the proportions of fixations to the subject and object characters in target event pictures across conditions.

3.1. First analysis (0-400 ms)

In all conditions, speakers rapidly directed their gaze to the agent in the picture within 400 ms after picture onset (main effect of Time: $\beta_1 = 7.18$, $\beta_2 = 6.08$, both $t > 16$). All main effects and interactions comparing the Subject and Object focus conditions against the Neutral condition were not significant (all $t < 1.5$). Thus the distribution of attention to the two characters at the outset of formulation showed little influence of question type.

Comparing the Subject Focus and Object Focus conditions against one another in a separate analysis showed that the two focus conditions also differed significantly (no main effects and interactions: all $t < 1$).
3.2. Second analysis (400-1000 ms)

After 400 ms, speakers rapidly directed their gaze to the agent in the picture (main effect of Time: \(\beta_1 = .49, \beta_2 = .51\), both \(ts > 2.5\)). The first contrast in the interaction between Time and Focus Condition showed that speakers were more likely to fixate agents in the Neutral condition than in the Object focus condition (\(\beta_1 = -1.38, \beta_2 = -1.48\), both \(ts < -3\)). There were no differences between the Neutral condition and the Subject Focus condition (all main effects and interactions: \(ts < 2\)).

Comparing the Subject Focus and Object Focus conditions against one another in a separate analysis showed a significant interaction of Focus Condition with Time (\(\beta_1 = 1.38, \beta_2 = 1.47\), both \(ts > 2\)): thus, as time progressed, fixations to agents within this analysis window increased in the Subject Focus condition but not in the Object Focus condition.

3.3. Third analysis (1000-1800ms)

Speakers began shifting their gaze away from the agent between 1000 ms and speech onset (main effect of Time: \(\beta_1 = -.98, \beta_2 = -.94\), both \(ts < -7\)). They were more likely to fixate agents in the Neutral condition than in the Object Focus condition (main effect of Focus Condition: \(\beta_1 = -2.02, \beta_2 = -1.93\), both \(ts < -3\)) and in the Subject Focus condition (\(\beta_1 = 1.12, \beta_2 = 1.18\), both \(ts > 2\)). The first contrast in the interaction between Time and Focus Conditions was significant (\(\beta_1 = .92, \beta_2 = .89\), both \(ts > 2\)): fixations to agents decreased at a steeper rate in the Neutral condition than in the Object Focus condition. The second contrast in the interaction between Time and Focus Condition was not significant (all \(ts < 2\)), as the decline in agent-directed fixations was comparable in the Neutral and Subject Focus conditions.

Finally, the Subject Focus and Object Focus conditions were compared against one another. As expected, the analysis showed that speakers fixed agents more often in the Subject Focus condition than in the Object Focus condition (main effect of Focus Condition: \(\beta_1 = -2.02, \beta_2 = -1.93\), both \(ts < -3\)). The interaction with Time was also significant (\(\beta_1 = -0.91, \beta_2 = -1.93\), both \(ts > 2\)): fixations to agents decreased at a steeper rate in the Subject Focus condition than in the Object Focus condition.

4. Discussion

Participants’ gaze patterns showed large differences in attention allocation between the focus conditions. Results from the Neutral condition replicate earlier findings showing that participants look at character in the order of mention: first the subject character (cowboy) and then the object character (bull; Griffin & Bock, 2000). Gaze durations for both characters were relatively similar, and gaze shifts to the object character occurred well before speech onset.

The Subject Focus and Object Focus conditions, however, showed that fixations to the two characters were strongly influenced by the preceding discourse context. In the Object Focus condition, participants looked briefly at the subject character and shifted their gaze to the object character about 300 ms after the picture onset (see Figure 2B). The Subject Focus condition, however, showed the opposite pattern: participants looked longer at the subject character (cowboy) and shifted their gaze to the object character (bull) only about 1500 ms after picture onset (and immediately before speech onset; see Figure 2C). Thus, our results clearly show that focus exerts a strong influence on the temporal coupling of gaze and
speech onset, and, consequently, on the incremental preparation of simple utterances.

It is important to note that fixation patterns were rather comparable across the three conditions before 400 ms, although, between 300-400 ms after picture onset, there appears to be an emerging trend of subtle differences in the gaze patterns in the Subject Focus condition, compared to that in the Neutral and Object Focus conditions. The overall pattern nevertheless suggested little difference in the way speakers scanned the pictures immediately after picture onset.

Fixation patterns observed after 400 ms showed that speakers directed their attention preferentially only to the part of the display that they needed to encode to felicitously answer the question: speakers primarily fixated the new character and directed fewer fixations to the character that was mentioned in the question (i.e., the unfocused or given character). Large differences in fixation patterns to the two characters after 400 ms indicate that fixations were primarily goal-driven and thus, that gaze patterns reflected generation of a different information structure for the target utterance across conditions. In other words, speakers quickly formulated pre-verbal messages that incorporated focus information and this information exerted a strong influence on the timecourse of linguistic encoding.

Importantly, differences between the Subject Focus and Object Focus conditions show that formulation of a sentence in response to a question is less linear than formulation of the same sentence in a neutral discourse context. While speakers fixated characters in the order of mention in neutral contexts (Griffin & Bock, 2000), there was a large reduction in fixations to objects in the Subject Focus condition and in fixations to subjects in the Object Focus condition. Thus speakers did not simply devote less attention to "old" characters than to "newer" characters (Konopka, 2014). Instead, the form of linear encoding observed in the Neutral condition was adjusted in the two focus conditions which resulted in a qualitatively different and much less incremental approach. In short, our results show that sentence formulation in a discourse context is strongly driven by focus structure and thus is not as tightly coupled with eye movements as in neutral discourse new contexts (Griffin & Bock, 2000).

5. Acknowledgements

We thank Margaret den Besten for the help with data collection and the financial support via the VIDI Grant (NWO-061084338) to YC.

6. References


Konopka, A. E. (2012). “Planning ahead: how recent experience with structures and words changes the scope of linguistic planning”.

In: Journal of Memory and Language, 66, pp. 143-162.


Effort and coordination in the production of bilabial consonants

Maëva Garnier¹, Sarah Bouhake¹, Christophe Jeannin¹,²

¹ GIPSA-lab, Speech and Cognition Department, CNRS and Grenoble Universities, Grenoble, France
² Dental School, Claude Bernard Lyon I University, Lyon, France

maeva.garnier@gipsa-lab.grenoble-inp.fr

Abstract

The first goal of this study was to compare two methods of measurement of lip articulation force. Bilabial consonants (/p/, /b/, /m/) were produced by a single French female speaker in modal and whispered speech, with varying levels of vocal effort. Lip compression force measured with a pressure sensor glued on the lower lip, correlates significantly, though only partially, with lip muscle activity, measured with surface EMG.

The second goal was to better understand the variations of lip articulation force in the production of stop consonant. Variations in lip compression correlate significantly, but not with a high level of correlation, with the variations in intraoral pressure. This does not support the idea that the primary goal of lip compression is to oppose to the pressure built up in the oral cavity. Lip compression force and Pio were found to correlate significantly, though moderately, with the intensity of the burst created at occlusion release, but not with its spectrum mean.

Keywords: stop consonants, articulation force, EMG

1. Introduction

Speech production requires the precise coordination of breathing, laryngeal and articulatory gestures. Learnt by children during their speech development, this coordination is then automatized. However, it still remains a complex motor action that can become destabilized in non-standard speaking styles (whisper, shouted or hyper-articulated speech) and that can dysfunction in several voice or speech disorders (vocal straining, stuttering, dyspraxia,...).

Stop consonants are of great interest for a better understanding of this coordination as they require a coordination not only in the timing of breathing, laryngeal and articulatory gestures, but also in their force. They are the first consonants to be produced at the babbling stage, and the most problematic segments for stutterers. They are also of particular interest for the measurement and modeling of production efforts, as the complete occlusion of the vocal tract requires a contact force between the lips, or between the tongue and the palate. It also implies vertical displacements of the larynx and adduction forces during the occlusion phase, perturbed vibration of the vocal folds (for voiced consonants) and a more or less abrupt glottal onset at the occlusion release (for unvoiced consonants).

The acoustical result of this coordination has been much studied and described in the phonetic literature: The spectrum of the bursts created at the occlusion release, the voicing intensity, the duration between the occlusion release and the voicing onset (Voice Onset Time (VOT)), or the direction of the first three formants transitions after the occlusion release, are as many audible cues to (un)voicing and place of articulation features that distinguish the different stop consonants (Forrest et al. 1988; Blumstein and Stevens 1979). The aerodynamics (intra-oral pressure, oral air flow) of the stop consonants production has also been extensively documented (Koenig et al. 2011). The articulatory targets of vocal tract constriction have been characterized using ultrasound imaging, electromagnetic articulography or (electro)palatography, depending on preceding and following speech segments, and depending on the phonation mode (whisper/modal, soft/loud) (Lofqvist and Gracco 1997). So far, articulatory forces have mainly been estimated from cinematic criteria such as velocity peaks of speech articulators (Nelson 1983). No method has been consensually adopted to measure static forces yet. Some first attempts have been made to measure contact forces using pressure sensors between the lips (Hinton et al. 1992), inserted in dental prosthesis or artificial palates (Jeannin et al. 2008). Other studies have tried to estimate lip compression force using surface electromyography (Lukker and Parris 1970).

Most of the phonetic work on stop consonants aimed at describing significant differences between phonological categories (voiced vs. unvoiced, different places of articulation, comparing stops in different languages, ...). However, it is not fully understood yet (1) which parameters and production gestures control the variation of phonetic features, and (2) the physical interactions that exist between the different control parameters. The present study aims at exploring the relationship between different acoustic, articulatory and aerodynamic descriptors of the bilabial stop production, with a particular interest in the variation of the lip articulation force.

2. Material and method

One French female speaker was recorded in laboratory conditions while producing the logatoms /apa/, /aba/ and /ama/ in modal and whispered voice, with 3 levels of vocal effort in the modal mode (murmured, conversational and shouted speech) and 2 levels of vocal effort for the whispered mode (normal and shouted whisper). 60 repetitions of the logatoms were recorded in each mode and level of vocal effort. A red light flashing every second gave her the speech rate. Several signals were simultaneously recorded:

- The audio signal was recorded with a 1/4-in pressure microphone (Bruel and Kjær 4944-A) placed 30 cm away from the lips. The sound pressure level was calibrated using the 1 kHz internal reference signal of the conditioning amplifier (Bruel and Kjær Nexus).
- The intra-oral pressure (Pio) was recorded with a small capillary tube and the EVA acquisition system.
- Lip muscle activity was measured with two pairs of surface electrodes placed on the superior orbicularis (EMG1) and the depressor anguli oris (EMG2) (see Figure 1).
The compression force between the upper and lower lips was estimated using a force sensor similar to that inserted in the PRESLA device (Jeannin et al. 2008), glued on the lower lip and calibrated in cm H2O, using a water column and a latex container (see Figure 1).

The different segment phases (occlusion, burst, formant transitions) were segmented manually from the audio signal, using Praat software. The voice onset time (VOT) was measured as the time between the burst of energy, at the occlusion release, and the second formant onset point of the vowel that follows (Jovičić and Šarić 2008), which enabled us to measure a VOT for whispered speech. With this definition, the VOT of the voiced stop /b/ is consequently positive.

The 4 following descriptors were then extracted for all the logatoms: the mean intensity of the logatom (in dB SPL), the maximum Pio (in cm H20), the maximum energy of the two EMG signals, and the maximum force of lip compression (in kPa).

The 3 additional descriptors were extracted for the stop consonants /p/ and /b/ only: the VOT (in ms), the intensity of the burst in energy (in dB SPL) and the center of gravity of its spectrum (in kHz). The burst spectrum was analyzed following the method described in Forrest et al. 1988, but up to 8kHz only (instead of 10kHz). In French, the burst of the voiced stop /b/ is produced, in modal speech, with voicing. In that specific case, we first modeled the burst signal as the sum of a harmonic component and a noisy component (HNN decomposition (Stylianou 1990)). Thus we were able to subtract the estimated voicing component to the burst signal before analyzing it.

The greater lip compression force observed in modal speech compared to whispered speech can simply come from the fact that modal speech is globally more intense (by about 20dB, test LRT, df = 3, F=2039.4, p<.001 ***) and also produced with higher intra-oral pressures (Pio) (test LRT, df=3, F=294.94, p<.001 ***). A significant positive correlation is indeed observed between the lip compression force and the average intensity of the logatom (test LRT, df=4, F=25.67, p<.001 ***). So is also a significant positive correlation observed between lip compression and Pio (test LRT, df=2, F=13.34, p<.001 ***).

3. Measurement of lip articulation force

3.1. Lip compression force measured with a pressure sensor

Lip compression force, measured with the pressure sensor, was found to vary significantly with the phonation mode and the segment type (significant interaction F(2,366)=8.20, p<.001 ***): whispered stops tended to be produced with reduced compression of lips, compared to modal logatoms (see Figure 2). Test LRT, df=3, F=22.96, p<.001 ###), although that difference is significant for /b/ and /m/ only. Lip compression also depends significantly on the stop consonant (test LRT, df=16, F=19.68, p<.001 ***): always greater for /p/ than /b/.

The voice onset time (VOT) was measured as the time between the burst of energy (in dB SPL), the mean intensity of the logatom /apa/, /aba/ and /ama/ in modal or whispered speech, with varying levels of vocal effort. For /apa/ and /aba/, the voiced stop /b/ is produced, in modal speech, with voicing. In that case, we first modeled the burst signal as the sum of a harmonic component and a noisy component (HNN decomposition (Stylianou 1990)). Thus we were able to subtract the estimated voicing component to the burst signal before analyzing it.

![Figure 1: Illustration of the force sensor and its calibration system, of the position of the EMG surface electrodes, and of the signals acquired simultaneously.](image1)

![Figure 2: Lip compression force, mean intensity and maximum intra-oral pressure (Pio) measured for the logatoms /apa/, /aba/ and /ama/ in comfortable modal and whispered voices.](image2)

![Figure 3: Correlation between lip compression force and average intensity of the logatoms /apa/, /aba/ and /ama/ produced in modal or whispered speech, with varying levels of vocal effort.](image3)
in Pio and global intensity. Figure 2 shows, in particular, how the segment /m/ is produced with comparable articulation force to the two other segments, although no pressure builds up in the oral cavity for that nasal stop. Consequently, the correlation between lip compression force and global intensity or Pio, is not always underlined by the same relationship but is significantly affected by the factors MODE and SEGMENT. Thus, the slope of the linear regression between lip compression force and global intensity is always greater in whisper than in modal speech (test LRT, df=1, F=20.49, p=7.10^(-6) *** and different between the 3 segments (test LRT, df=2, F= 6.56, p=0.001 **). Likewise, the slope of the linear regression between lip compression force and Pio is greater in whisper than in modal speech (test LRT, df=1, F=21.38, p<.001 ***). However, it does not depend significantly on the segment (see Figure 4).

Consequently, we consider that these two signals are not redundant or equivalent, and that the level of muscle activity cannot entirely predict the amount of lip compression force.

Likewise, the intensity of the two EMG signals are significantly and positively correlated (test LRT, df=4, F= 58.1 p<.001 ***). See also Table 1). However, the correlation is only partial (R=0.67 at its best for /p/ in modal speech) and still depends on the mode (test LRT, df=1, F= 23.2, p=2.10^(-4) *** and the segment (test LRT, df=2, F=4.0, p=0.018, *). Consequently, we can consider that these two signals are not redundant and bring complementary information about the articulation of stop consonants.

4. Variation of phonetic cues

4.1. Voice Onset Time (VOT)

In agreement with previous studies, the VOT was found to be shorter for voiced consonant /b/ compared to the unvoiced consonants /p/ (F(1,206)= 31.59, p<.001 ***). A significant increase of the VOT was also observed from modal speech to whispered speech (F(1,206)=180.16, p<.001 ***).

A significant and positive correlation was found between the lip compression force, measured with the pressure sensor, and the level of lip muscle activity, for the superior oribicularis (EMG1: test LRT, df=4, F=44.6, p<.001 ***) and the depressor anguli oris (EMG2: test LRT, df=3, F=59.2, p<.001 ***). However, this relationship is not invariant but depends significantly on the segment: the slope of the linear regression (coefficient “a” in the tables) tends to be greater for the segment /m/, compared to /b/, and to /p/ (EMG1*Segment: test LRT, df=4, p<0.001 ***; EMG2*Segment: Test LRT, df=2, F=4.0, p=0.019, *). This correlation also depends significantly on the mode of phonation for the superior oribicularis muscle (EMG1*Mode: test LRT, df=2, p<0.001 ***), with always a greater slope of regression in whisper than in modal speech. This is not the case for depressor anguli oris muscle (EMG2*Mode, test LRT, df=1, F=1.2 p=0.26 NS).

These results indicate that the measures of lip compression force and lip muscle activity are not redundant or equivalent, and that the level of muscle activity cannot entirely predict the amount of lip compression force.

4.2. Articulation force estimated from lip muscle activity

<table>
<thead>
<tr>
<th>Force /EMG1</th>
<th>apa</th>
<th>aba</th>
<th>ama</th>
</tr>
</thead>
</table>
| Modal       | R=0.23, **  
|             | a=2.29 | R=0.25, ***  
|             | a=3.76 | R=0.44, ***  
|             | a=4.66 | Whisper     | R=0.42, ***  
|             | a=3.88 | R=0.39, ***  
|             | a=4.75 | R=0.47, ***  
|             | a=6.73 |

<table>
<thead>
<tr>
<th>Force /EMG2</th>
<th>apa</th>
<th>aba</th>
<th>ama</th>
</tr>
</thead>
</table>
| Modal       | R=0.41, ***  
|             | a=4.55 | R=0.30, ***  
|             | a=8.29 | R=0.50, ***  
|             | a=8.75 | Whisper     | R=0.32, ***  
|             | a=4.53 | R=0.20, *   
|             | a=4.71 | R=0.62, ***  
|             | a=6.50 |

<table>
<thead>
<tr>
<th>EMG1/EMG2</th>
<th>apa</th>
<th>aba</th>
<th>ama</th>
</tr>
</thead>
</table>
| Modal      | R=0.23, **  
|            | a=0.25 | R=0.41, ***  
|            | a=0.75 | R=0.38, ***  
|            | a=0.63 | Whisper     | R=0.67, ***  
|            | a=1.02 | R=0.51, ***  
|            | a=0.97 | R=0.23, **   
|            | a=0.17 |

| Modal      | R=0.45, ***  
|            | a=0.68 | Whisper     | R=0.08, NS   
|            | a=0.69 |

Table 1: Correlation between the lip compression forces measured with a pressure sensor, and the activity of two lip muscles (EMG1: superior oribicularis; EMG2: depressor anguli oris), for the three logatoms /apa/, /aba/ and /ama/.

Figure 5: Correlation between the voice onset time and the maximum Pio, for the two bilabial /p/ and /b/ produced in modal or whispered speech, with varying levels of vocal effort. The VOT was measured as the time difference between the burst and the onset of the second formant.

The variations of Pio partly explain these variations of VOT (Correlation VOT ~ Pio: test LRT, df=3, F=12.44, p<.001 ***), although such a correlation was not observed for /p/ in whispered speech (see Figure 5). Again, this relationship is not always the same and depends significantly on the segment (test LRT, df=1, F= 4.89, p=0.027 *) and the mode (test LRT, df=2, F=18.34, p<.001 ***): with positive correlations in modal speech and negative ones in whisper.
4.2. Burst intensity and spectrum
The burst intensity varies significantly as a function of both the mode of phonation and the segment (mode*segment: F(1,246)= 50.59, p<.001 ***). Although no general tendency can be drawn, the variations of the burst intensity can partly be explained by the variations in intra-oral pressure (Correlation: Test LRT, df=3, F=46.45, p<.001 ***) and lip compression force (Test LRT, df=3, F=16.86, p<.001 ***), excepted for the segment /b/ produced in modal voice. This relationship depends significantly on the mode of phonation, with positive correlations in whisper and negative ones in modal speech.

[Diagram showing burst intensity and spectrum]

Figure 6: The top figures represent the intensity and the spectrum mean (CoG) of the burst produced at the occlusion release for the logatons /apa/ and /aba/ in conversational modal and whispered voices. The bottom tables summarize the results of the correlation analysis between the burst intensity or CoG, and two control parameters: the intra-oral pressure (Pio) and the lip compression force, measured with the pressure sensor.

<table>
<thead>
<tr>
<th></th>
<th>$I_{Burst} \sim \text{Pio}$</th>
<th>$I_{Burst} \sim \text{Lip force}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>apa</td>
<td>aba</td>
</tr>
<tr>
<td><strong>Modal</strong></td>
<td>R=0.39, ***</td>
<td>R=0.02, NS</td>
</tr>
<tr>
<td></td>
<td>a=0.63</td>
<td></td>
</tr>
<tr>
<td><strong>Whisper</strong></td>
<td>R=0.46, ***</td>
<td>R=0.22, **</td>
</tr>
<tr>
<td></td>
<td>a=0.71</td>
<td>R=0.02, NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\text{CoG}_{Burst} \sim \text{Pio}$</td>
<td>$\text{CoG}_{Burst} \sim \text{Lip force}$</td>
</tr>
<tr>
<td></td>
<td>apa</td>
<td>aba</td>
</tr>
<tr>
<td><strong>Modal</strong></td>
<td>R=0.14, NS</td>
<td>R=0.01, NS</td>
</tr>
<tr>
<td></td>
<td>a=0.23, **</td>
<td>R=0.867</td>
</tr>
<tr>
<td><strong>Whisper</strong></td>
<td>R=0.13, NS</td>
<td>R=0.20, NS</td>
</tr>
<tr>
<td></td>
<td>a=0.05, NS</td>
<td>R=0.07, NS</td>
</tr>
</tbody>
</table>

The center of gravity of the burst spectrum (CoG) was found to be slightly higher in whispered speech, compared to modal speech (test LRT df=2, F=26.83, p<.001 ***). On the other hand, no significant correlation was found between the variation of CoG and the Pio or the lip compression force, except for the segment /p/ produced in modal voice.

5. Conclusion
Two methods of measurement or evaluation of lip articulation force have been compared here on a single female French speaker, for a large variety of productions (modal and whispered speech, from murmured to shouted levels). Lip compression forces measured with a pressure sensor glued on the lower lip were observed with comparable values to the literature (Hinton et al. 1996) and were found to increase with the global level of vocal effort. The activity of two lip muscles controlling the movement of the upper and the lower lips correlates significantly with this lip compression force, as well as with each other, without providing redundant information. The lip compression force correlates significantly, though not with a high level of correlation, with the maximum Pio of the bilabial plosives /p/ and /b/. Furthermore, the nasal stop /m/ demonstrates comparable articulation force to the two other segments, although it is produced with weak or null Pio. These results do not support the idea that the primary aim of lip compression is to maintain occlusion of the vocal tract by opposing to variations of Pio.

Lip compression force was also found to correlate significantly, though not with a high level of correlation, with the burst intensity. No such relationship was found with the spectrum CoG of the burst. This results support the idea that lip articulation force may partly control the variations of burst intensity.

Further investigations will explore the combined contribution of different control parameters to the variation of the burst characteristics, but also to the transient characteristics that follow the occlusion release and that are crucial to the perception of stop categories (Blumstein and Stevens 1979).

6. Acknowledgements
This work was supported by the pole CSVSB of the University Joseph Fourier (Grenoble). We also thank Thomas Hueber for his help with the HNN decomposition.

7. References
Domain initial strengthening and height contrast in French: acoustic and ultrasound data
Laurentianne Georgeton, Tanja Kocjančič Antolik & Cécile Fougeron
Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Univ. Paris 3-Sorbonne Nouvelle, France
{laurianne.georgeton; cecile.fougeron}@univ-paris3.fr; tkocjancic@gmail.com

Abstract
This paper investigates how prosodic boundary strength (IPi vs. IPm) affects the production of the vowels /i, e, æ/ which contrast on a four degrees of height scale in French. Acoustic and tongue configuration data are examined for four speakers. Results show an expansion of the vowel space in IP-initial position that is achieved by a raising of F2 for /i, e, æ/ and of F1 for /æ/. Differences in tongue configuration also contribute to an expansion of the articulatory space in IP-initial position are also observed. Measurements on the highest point of the tongue show a narrowing of the constriction for /i, e, æ/, accompanied by fronting for /e, æ/, and a widening of the constriction for /æ/ for most speakers. These variations in IP-initial position lead to a maximization of phonetic contrasts in terms of height and frontness within the pairs /æ-e/ and /æ-æ/ for most speakers, but not within the pair /i, æ/, probably due to articulatory/acoustic constraints.

Keywords: vowels, prosodic boundary, French, height contrast

1. Introduction
Several studies have reported an effect of prosodic position on vowels, with a modification of their acoustic or articulatory properties when accented or close to a prosodic boundary (see the review by Cho 2011; and recently Kim & Cho 2011, 2012; Georgeton & Fougeron, 2014). Except for the last reference, most studies have been limited to the investigation of only a few types of vowels, usually peripheral vowels. Consequently, it is not clear whether all vowel types are modified in the same way, nor whether prosodically driven segmental variations may be modulated by the density of the phonological inventory.

In order to address these questions, the purpose of this study is to investigate prosodic effect on vowels in a dense dimension of contrasts, namely the four levels of height on which the four front vowels /i, e, æ/ are contrastive in French. Variations in the lingual and acoustic properties of these vowels are tested according to the strength of the prosodic boundary, i.e., according to whether they are initial or medial in an Intonational Phrase (IPi or IPm). Few studies have investigated variations in the lingual articulation of vowels in absolute domain-initial positions (#VC sequences) and their results show that vowels are influenced by prosodic strengthening (as consonants do) with a global increase in gestural magnitude in higher prosodic constituents, which interacts with vowel in different directions. In a study investigating lingual variation of the two vowels /æ/ and /e/ in English, Lehnhrt-Lehouillier and colleagues reported a greater articulatory magnitude in IP initial position, without more description on the direction of the changes in lingual configuration (Lehnhrt-Lehouillier, McDonough McAleavey, 2010). Kim and Cho (2011, 2012) observed that for all the three front English vowels /æ/ , /e/, /i/, boundary induced variation resulted in a featural enhancement of [+/high] properties in such a way that, in IP-initial position, the high front vowels /i, æ/ were higher while the low front /æ/ was lower. For the vowel /æ/, this variation in height was accompanied by more anterior tongue position.

In French, prosodically driven lingual variations on vowels have been mostly investigated under focal accentuation or in domain-final position. Loewenbruck (1999, 2000) observed a similar expansion in height contrast between /i/ and /a/ under focal accentuation, with a higher tongue body for /i/ and a lower tongue body for /a/. Taban & Perrier investigated domain final /i/ (2005), /a/ (2003) and /æ/ (2007) in different prosodic boundaries. A lower tongue body before stronger prosodic boundaries was also found for the low vowel /a/, but for /i/ the effect was lesser and speaker-dependent: one of their three speakers showed a backing of the tongue but the other two tended to raise and front their tongue body. For the vowel /æ/, they observed tongue dorsum backing coupled with raising or lowering depending on the speaker. The authors concluded that these different strategies concurred to a common acoustic goal: the raising of F3 for vowel /i/, and the lowering of F2 for vowel /æ/ in order to prevent a perceptual confusion with /y/.

These results suggest that articulatory variations induced by prosodic boundaries may depend on the language’s phoneme inventory and the preservation of vowel contrasts (see also Cho & Jun 2000 for consonantal contrasts). In a recent study (Georgeton & Fougeron, 2014), we also observed contrast-dependent variation in domain-initial position, by looking at the labial articulation and acoustic properties of the 10 oral vowels of the French system. While all vowels showed an increase in lip area in IP-initial position, this effect was found to be larger for the unrounded vowels. Consequently, the contrast between front rounded and unrounded vowels was found to be maximized in IP-initial position.

In the present paper, we address further this question by investigating changes in tongue configuration and acoustic property for the front (unrounded) French vowels /i, æ/.

Within these four levels of height, density-dependent limitations on phonetic variation may be at play (as suggested by Manuel, 1990 for example). According to the literature cited above, an enlargement of the oral constriction can be expected for the lower vowel /æ/, while predictions are not clear for the vowel /i/, and absent for the non-peripheral vowels, the mid-closed /æ/ and mid-open /e/. We will therefore examine how prosodic effect modifies the articulation of these domain-initial vowels and whether it interacts with vowel contrasts in this dense system.

2. Material and method
The lingual configuration and acoustic properties of the four front oral vowels /i, æ, e, æ/ have been investigated in two prosodic conditions: in an Intonational phrase-initial (IPi) vs. an IP-medial (IPm) position. For the IPm position, vowels were initial in the second word of a fake compound first name. The
four vowels were produced in controlled sentences in a [ip/VC] context. C is /p/ for /i, e, a/ and /v/ for mid-open /i/, in order to prevent its pronunciation as mid-closed (see Georges & Fougéron 2014 for the description of a similar corpus). Four female speakers were recorded with a Midray DP600 ultrasound (60 i/sec) with head stabilization (Articulate Instruments Ltd, 2008).

Each sentence was produced 10 times in a random order but repetitions with un-exploitable tongue contours have been discarded from the analysis. Table 1 summarizes the number of renditions analyzed per vowels and prosodic positions (IPi, IPm) for the 4 speakers (SA, SC, SL, SZ).

Table 1: Number of renditions analyzed by speaker, prosodic position and vowels.

<table>
<thead>
<tr>
<th></th>
<th>SC</th>
<th>SL</th>
<th>SZ</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPi</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>IPm</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>IPi</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>IPm</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>IPi</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>IPm</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>IPi</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>IPm</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>IPi</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>IPm</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Target vowels were segmented and labeled in Praat in order to extract acoustic duration and formant values. F1 and F2 were taken at three successive points in the middle of target vowels and then averaged. For lingual configuration, one to three (depending on vowel duration) successive tongue contours were traced manually in the middle of the vowel with Articulate Assistant Advanced (Articulate Instruments Ltd, 2012) and then averaged for each vowel/position condition. An estimation of vowel height and place of articulation was done by quantifying the height (y-axis) and front-back location (x-axis) of the highest point of the tongue, which was automatically extracted from individual contours.

In order to test whether F1, F2, duration and coordinates of the highest point of the tongue vary according to prosodic position, analyses by speaker are done with sample t-tests for each vowel. In order to test the interaction of boundary effects with height contrast between each pair /i-/e/, /e-/a/ and /a-/e/ two-factor ANOVAs (position, vowel) were conducted for each speaker. (Note that due to space limitations, statistical details are not given here and only significant differences are reported; for speaker SA, given his small number of renditions, only tendencies are reported).

3. Results

3.1. Boundary effect (IPi vs IPm) on the lingual and acoustic properties per vowel types

Differences in tongue contours and in spectral properties according to prosodic positions are illustrated in Figure 1. For all speakers, an effect of prosodic position is observed, with speaker- and vowel-dependent patterns. Looking at F1 and F2 of vowels in IP-initial position compared to IP-medial position, the following variations are observed:

• /i/ has a higher F2 for all speakers except SZ, and no variation is found on F1;
• /e/ has a higher F2 for all speakers and a higher F1 for one speaker (SL)
• /e/ has a higher F2 and a higher F1 for all speakers;
• /a/ has a higher F1 for all speakers. An effect on F2 is found for speakers SL and SZ but in an opposite direction: higher F2 for SZ and lower F2 for SL.

These spectral variations are not accompanied by systematic differences in acoustic vowel duration. Speakers SL, SC and SA have longer /i/ in IPi, but for the other vowels, few differences in vowel duration appear, and these differences are speaker- and vowel-dependent. A lengthening in IPi is found for /e/ for SC, a shortening is found for /e/ for SL, for /a/ for SL and SZ, and for the nine other comparisons there is no change in vowel duration.

Considering articulatory variations, the differences in tongue contours between the two positions presented in Figure 1 appear to be larger for speakers SL, SZ and SA than for speaker SC, and are clearly vowel dependent.

As explained in the method section, a quantification of the differences in lingual configuration is made to estimate the degree and location of the constriction at the highest point of the tongue. In IP-initial position, the following variations are observed:

• /i/ has a narrower constriction with a rising of the highest point of the tongue for all speakers except SZ, which rather shows a baking of the constriction.
• /e/ has a narrower constriction for all speakers and a front constriction for speaker SL and SA;
• /e/ has a narrower constriction for all speakers and a front constriction for all except SC;
• /a/, on the other hand, has a wider constriction for all speakers, and a backer constriction for all except SC.

Overall, from a systemic perspective, the variations observed in IP-initial compared to IP-medial position contribute to an expansion of the acoustic and articulatory spaces. This expansion is achieved by an enlargement of the spaces both in the vertical dimension (constriction height, F1) and horizontal dimension (constriction location, F2).

3.2. Boundary effects on the contrast between vowel pairs

The effect of prosodic position is further tested here on the acoustic and articulatory contrast between adjacent vowels along the vowel height dimension of contrast. For this, we test whether the effect of boundary depends on the vowel identity within the three vowel pairs (/i-/e/, /e-/a/ and /a-/e/), and therefore whether the contrast between the members of the pairs is affected by prosodic position.

For the /i-/e/ pair, a significant interaction is found on constriction height for all speakers: both /i/ and /e/ have a narrower constriction in IPi, but the amplitude of the tongue rising for /e/ is larger than that for /i/. Consequently, /e/ gets closer to /i/ and the contrast in height dimension between these two vowels is not maximized in IP-initial position. An interaction is also found on dimensions linked to place of articulation: front-back position of the constriction for SL and SC, and F2 for SZ and SA. For all speakers except SC, both vowel are modified in the same way (higher F2 for SZ and SA and front constriction for SL) but with a larger change for /e/. Consequently /e/ gets closer to /i/. Speaker SC is the only one showing a larger contrast in constriction location between /i/ and /e/ in IPi, with a fronting of the constriction for /i/ and a backing for /e/.

For the pair /e-/a/, an interaction is found on the dimensions linked to vowel height: F1 for all speakers except SL, and constriction height for SL. This interaction reflects a maximization of the distinction between the mid-closed and mid-open vowels in IPi for all speakers, with a greater narrowing of the constriction for /e/ than for /a/ (SL), and a large rising of F1 for /i/ (SC, SZ, SA). Interactions are also found on the dimensions linked to place of articulation for all speakers except SC: on F2 for SL and SA, on the front-back
position of the constriction for SL and SZ. Again the contrast between the two vowels is maximized in IPi for SL and SA with a larger F2 rise for /e/ than /ɛ/ (SL and SA) and more fronting of the constriction for SL. For SZ however, the two vowels get closer in IPi with a slight backing of the constriction for /e/ and a slight fronting for /ɛ/.

For the pair /ɛ-æ/, interactions are found on constriction height for all speakers, on F1 for all speakers except SA, on constriction location for all except SC and on F2 for all except SZ. The two vowels are more distinct in IPi position with a wider constriction and higher F1 for /æ/ than /ɛ/ for all speakers except SA, and with a backer constriction for SL, SA, SZ, and a lower F2 for SL, SC, SZ.

4. Discussion and conclusion

In this study, we investigated the effect of prosodic position on the lingual and acoustic properties of the front unrounded oral French vowels /i, e, æ, a/ in IP-initial vs. IP-medial positions. As found in the other studies presented in the introduction, we show that prosodic position influences the articulation and acoustic properties of the vowels. In IP-initial position an expansion of the articulatory and acoustic spaces is observed in height and front-back dimensions of the lingual constriction (estimated from the highest point of the tongue) as well as in F1 and F2 dimensions.

The observation of more than two degrees of vowel height provides interesting results on the direction of this boundary effect. Different sets of vowels can be distinguished. On the dimension related to tongue height (vertical position of the highest point of the tongue), the non-low vowels /i, e, æ/ pattern together with a narrower constriction degree in IP-initial position for most of the speakers (three for /i/, four for /ɛ, ɛ/). On the other hand, the low vowel /æ/ shows a widening of the oral constriction for all speakers, as observed in other studies. Regarding the location of the constriction in the front-back dimension, a fronting of the constriction is observed in IPi for /ɛ, æ/ (two speakers for /ɛ/, and three for /æ/), while /æ/ has a backer constriction (three speakers). These measurements made on the highest point of the tongue appear to be well suited to quantify some of the differences in tongue contours illustrated in Figure 1. However, they do not capture the overall modifications in tongue shape and location. Modifications in the oral cavity resonators are therefore better accounted for by the variations in F1 and F2. In IP-initial position, the vowels /i, e, æ/ also pattern together with a rising of F2 for at least three speakers. In terms of F1, the non-high /i, e, æ/ are both characterized by a rising of F1 for all speakers (also for /ɛ/ for one speaker).

Our data further show that in a dense system of contrast, where vowels have little room for phonetic variations, prosodic boundary effects may contribute to maximize the contrast between adjacent vowels. An increase of phonetic distinctiveness is observed on our articulatory and/or acoustic measurements with differences according to speakers and vowel pairs. Overall, for all pairs but /i-ɛ/, the distinction between the vowels is maximized in IP-initial position in at least one dimension for all speakers. The contrast between the mid-open and mid-closed vowels is mainly increased in the height dimension (F1 and constriction height), and the contrast between the mid-open and open vowels is mainly increased in height and/or frontness of the construction. In most cases, the direction of the articulatory and/or acoustic variations in IP-initial position is similar for the two vowels of the pair, but is larger for one of them, leading to an increase of contrast between the two. A similar tendency was observed in

Georgeton & Fougeron (2014) for the labial configuration resulting in an increase of contrast between front rounded and unrounded vowels (see also Cole et al. 2007 for consonants). For the pair /i-ɛ/, however, no increase of contrast is found, except for the location of the constriction for SC. Acoustic and physiological limitation can explain that the tongue rising in IPi is more constrained for /i/ and is thus smaller than that for /ɛ/. Similarly, boundary induced variation at the lips were found to be larger for unrounded vowels than for the more constrained rounded vowels in Georgeton & Fougeron (2014). Taken together these results suggest that phonetic contrasts between vowels tend to be maximized in IP-initial position in dimensions that are less constrained by articulatory and/or acoustic limitations. Cross-linguistic comparisons between languages with different vowel inventories are now necessary to determine how these limitations comply with the language system of contrasts.

5. Acknowledgements

This work has been supported by the French Investissements d’Avenir - Labex EFL program (ANR-10-LABX-0083).

6. References


Figure 1: Mean tongue contours and F1/F2 values of the four vowels /i/ (orange), /e/ (green), /ɛ/ (blue) and /a/ (red) in IPi (solid line) and IPm (dotted line) for the four speakers (SZ, SL, SC, SA).
Italian Vowel and Consonant (co)articulation in Parkinson’s Disease: extreme or reduced articulatory variability?

Barbara Gili Fivela1, Massimiliano Mario Iraci1, Vincenzo Sallustio2, Mirko Grimaldi1, Claudio Zmarich3, Danilo Patrocinio2

1CRIL (Centro Ricerca Interdisciplinare sul Linguaggio), DREAM – University of Salento – Italy, 2Centre of Phoniatry and Rehabilitation of Communication Disorders – ASL Lecce – Italy, 3CNR-ISTC – Padova – Italy

Abstract

The present study investigates primarily on cross-speaker variability in the amplitude and duration of gestures, discussing the first data available on the articulation of consonants and vowels in Italian Parkinson’s Disease (PD) dysarthric speech. Observations on PD speech gesture variability reported in the literature are not consistent, ranging from a decreased to an increased gestures’ amplitude. In this study, we discuss acoustic and kinematic data (3D electromagnetic articulograph AG501) on Italian bilabial stops and two vowels ([i], [a]) produced by two mild-to-severe PD speakers and two peer healthy controls. Our results show that both increased and decreased amplitude and duration can be reported even in one subject’s productions, if considering different gestures (opening vs. closing) or different planes (vertical vs. horizontal).

Keywords: Parkinson’s Disease, dysarthria, EMA.

1. Introduction

Hypokinesia is a crucial characteristic of Parkinson’s Disease (PD) and PD is in fact the major cause of hypokinetis dysarthria (Duffy 2005). Hypokinesia is defined by Ackermann and Ziegler (1991) as reduced range of simple limb movements with consequent target undershooting. However, experimental results show evidences of both reduced and increased amplitude in speech gestures (for the former, e.g. Skodda 2011, 2012 on acoustics and Forrest et al. 1989 on kinematics; for the latter, e.g., EMA recent studies by Wong et al. 2010, 2011). As for Italian, there are currently no articulatory data on PD speech productions. Nevertheless, preliminary acoustic results on this type of pathological speech (Gili Fivela et al. 2013) show, for instance, difference in intensity between bilateral voiced plosives and following vowels which is lower in PD speakers than in control subjects, pointing to the lack of a complete closure in plosives and to a possible reduction in speech gestures’ amplitude.

The main goal of this study is to check by means of both acoustic and articulatory data whether hypokinetic dysarthria in Italian PD corresponds to consistent features across speakers or rather to cross-speaker variability, possibly depending on speaker-specific strategies. In particular, we aim to verify if variability in speech gestures is reduced or increased; moreover, hypothesizing that both cases (high or reduced variability) might affect the inter-gesture coordination, we aim to check how gestures’ phasing can be affected.

In the following, we focus on both vowels and plosive consonants. As for the latter, Italian is a quite interesting test language, as both singleton and geminate consonants are found: geminates show longer acoustic duration and are preceded by shorter vowels than singletons; from an articulatory point of view, the (lower lip) opening gesture shows, among other things, longer duration and greater amplitude in geminates than in singletons (Gili Fivela and Zmarich 2005). Differences in consonant gestures’ amplitude may then be operatively tested by referring to the singleton vs. geminate contrast.

2. Methods

2.1. Corpus and recordings

The corpus is composed of ‘CV’/CV disyllables, where Vs are /i/ - /a/ and initial or intervocalic (C/C) may be bilateral stops or nasals (/p, b/); /m/ in controls. Disyllables are: [ pi-pa], [ bi-ba], [ pip-pa], [ bib-ba], [ pip-ma] and [ bib-ma] (the latter two used to check [p, b] in non-doubtful syllable coda position) and are inserted in a carrier sentence: La X blu ‘The blue X’.

Two mild-to-severe PD speakers from Lecce – Southern Italy (PD-3: age 66; PD-4: age 81) and two peer healthy controls (CTR-1: age 64; CTR-2: age 80) participated in the study. The same severity level was reported for PD speakers according to clinical assessment, although at the time of recording PD-3 showed a higher degree of global motor impairment than PD-4. Each speaker read the corpus 7 times at normal speech rate. Acoustic and articulatory data were collected synchronously by an EMA 3D AG501 (Carstens Med., GmbH) in a quiet room at CRIL (Lecce, Italy). The articulatory data were recorded by means of 7 sensors, glued on subjects: 2 on tongue mid-sagittal plane (dorsum and tip), 2 on lips vermilion border (upper and lower), 1 on the nose and 2 behind the ears for normalization.

2.2. Measurements

The acoustic signal was manually labelled in PRAAT (Boersma and Weenink 2009), to identify consonant and vowel boundaries and three points in the central part of the vowel for first (F1) and the second (F2) formant measurements. Consonant (C) and both preceding and following vowel duration (V0 and V1, respectively) were then automatically extracted together with formants values (used to compute an average F1 and F2 value for each vowel). Articulatory data labeling was performed by means of MAYDAY (Sigona et al., in prep.). For both tongue dorsum (TD) and lower lip (LL) track, we labelled: gesture attainment, located at the zero velocity, for the tonic vowel (V0), the following Consonant (C) and the post-tonic vowel (V1); for each tracked segment, the maximum velocity was labelled at the velocity peak of the relevant coil. The duration and amplitude of lower lip/tongue dorsum gestures were calculated. For the closing gestures, for instance, the duration was calculated as the time interval between the maximum aperture and the maximum closure of the articulators; the amplitude was calculated as the vertical component of the articulator displacement during
gesture; for tongue dorsum only, the horizontal component of displacement was also computed to acquire information on the anterior-posterior displacement. Moreover, to investigate the inter-gestural coordination, we measured the time interval from the highest tongue dorsum position for [i] to the bilabial closure for [p] or [b], and the time interval from the bilabial closure for [p], [b] or [m] to the lowest tongue dorsum position for [a].

As for statistical analysis, repeated measures ANOVAs were run, both on single speaker and overall results, with Repetition as the within-subject factor (6 levels) and Consonant status (3 levels) as well as Voicing of the consonant (2 levels) as between-subject factors; Speaker (4 levels) was used as a factor in the analysis of overall results.

3. Results

In some cases, consonant clusters were realized with an intervening schwa vowel (e.g., pip[ma] and, in the case of one PD speaker (PD-4), they were even simplified inverting the consonant sequence to obtain a more familiar cluster (e.g., pipma’pimpma). Instances of both types were not considered for statistics. For this reason, at this stage, we discarded clusters produced by the two elderly subjects, that is one control (CTR-2) and one PD speaker (PD-4), and seven extra cases. To clearly show the differences found in control and PD speakers, results are reported of repeated measure ANOVAs run with Speaker and Consonant status as factors, while statistical checks of the relevance of the Consonant status factor on single speakers are reported only for those measures that, on the basis of the literature, are known to be more strictly related to the presence of linguistic contrasts involving singleton, geminate and cluster consonants. Statistics on the influence of Speaker and Voicing always showed the latter to be non-significant and its interaction with Speaker was usually the same found in the other ANOVAs. They are not reported here for space limits.

3.1. Acoustic data

The check relative to the influence of Speaker and Consonant status on the duration of the preceding, tonic vowel (V0) showed that only Speaker is a significant factor (Consonant: [F(2,10)=3.038; p>0.05]; Speaker: [F(3,10)=22.056; p<0.05]) with no interaction between the two ([F(4,10)=1.343; p>0.05]). The Tuckey post-hoc showed that one PD speaker, PD-3, and his matching control, CTR-1, show shorter vowels than others, although they differ from each other and PD-3, who shows the shortest vowels, differs from all others, while CTR-1 differs from PD-3 and CTR-2 (who shows the longest vowels). However, a statistical check for every single speaker showed that for only one speaker the Consonant status had a significant impact on the preceding vowel duration (CTR-1: [F(2,3)=4.397; p<0.05]; CTR-2: [F(1,2)=31.273; p<0.05]; PD-3: [F(2,3)=1.15; p>0.05]; PD-4: [F(2,3)=1.978; p>0.05]). In case of significance, singletons are preceded by longer vowels than geminates.

As for the preceding vowel formant values (see Fig. 1, left), results on F1 showed that only Speaker is a significant factor (Consonant: [F(2,10)=4.108; p>0.05]; Speaker: [F(3,10)=14.42; p<0.05]) with interaction between the two ([F(4,10)=3.774; p<0.05]). The Tuckey post-hoc comparison showed that speaker CTR-2 corresponds to significantly highest values, speaker PD-3 to the lowest values and the other speakers to intermediate values, with CTR-1 showing F1 values that are more similar to PD-3 than to CTR-2 and PD-4 that show values that are closer to CTR-2’s ones. Quite differently, results on F2 values showed that both factors are significant (Consonant: [F(2,10)=17.276; p<0.05]; Speaker: [F(3,10)=184.222; p<0.05]), with no interaction between the two ([F(4,10)=2.428; p>0.05]). The Tuckey post-hoc showed that clusters correspond to lower F2 in the preceding vowel than other consonant(s) and that PD speakers show lower F2 values than other speakers; the latters also differ from each other, with CTR-1 showing higher F2. Notice that a similar pattern for F2 formant value in relation to the Speaker factor is found for the post-tonic vowel /a/ (see Fig. 1, right). Its F1 values significantly differ in relation to the Speaker factor only (for Consonant: [F(2,10)=2.600; p<0.05]; Speaker: [F(3,10)=44.049; p<0.05]), with no interaction between the two ([F(4,10)=1.302; p>0.05]). PD speakers show lower F1 values than control speakers, and the Tuckey post-hoc test shows that the /a/ vowel is produced with significantly lower F1 by PD-3 and with significantly higher F1 by CTR-1. As for F2, both factors are significant (for Consonant: [F(2,10)=15.652; p<0.05]; Speaker: [F(3,10)=49.596; p<0.05]), with interaction between the two ([F(4,10)=9.262; p<0.05]). In particular, F2 is lower in PD than in control speakers, and the post-hoc test shows that /a/ is produced with significantly lower F2 by PD-3 and with significantly higher F2 by CTR-1 (but they do not differ significantly from PD-2 and CTR-2, respectively); as for Consonant status, F2 in /a/ is lower after geminates than after singletons and clusters, who in fact show the highest values.

![Figure 1: F1 and F2 (Hz) values for V0 /æ/ and V1 /a/](image)

As for consonant(s) duration, data show that both Speaker and Consonant status factors are significant (Consonant: [F(2,10)=105.992; p<0.05]; Speaker: [F(3,10)=19.746; p<0.05]), with no interaction between the two ([F(4,10)=1.006; p>0.05]). The Tuckey post-hoc showed that singletons correspond to significantly shorter values than geminates, which are significantly shorter than clusters; moreover, control speakers show longer segments than PD speakers, with CTR-1 showing significantly higher values than any other subject. The statistical check performed on every single speaker data showed that the Consonant status has a significant impact on the duration of the consonant segment(s) only for CTR-1 and PD-3 (CTR-1: [F(2,3)=58.979; p<0.05]; CTR-2: [F(1,2)=6.502; p<0.05]; PD-3: [F(2,3)=44.464; p<0.05]; PD-4: [F(1,2)=4.513; p<0.05]), for which, according to post-hoc test, geminates and singletons are shorter than clusters.

3.2. Articulatory measurements

3.2.1. Vocalic gestures

The duration and amplitude of the closing gesture toward /i/ (from preceding /æ/ was investigated on both the vertical and the horizontal axis, with reference to the TD coil. On the vertical axis, the duration of the closing gestures for V0 is only sensitive to the Speaker factor (Consonant: [F(2,10)=3.33; p<0.05]; Speaker: [F(3,10)=57.484; p<0.05]) with no interaction between the two ([F(4,10)=3.142; p<0.05]). The Tuckey post-hoc showed that PD-3 has shorter gestures and CTR-2 has longer gestures than all other speakers, with the others showing intermediate values (with CTR-1 not different from PD-3). As for the horizontal dimension, only the Speaker factor is significant (Consonant: [F(2,10)=839; p<0.05]; Speaker: [F(3,10)=16.628; p<0.05]) with no interaction between the two ([F(3,10)=1.144;
p>0.05). Post-hoc results are similar to those for the vertical component (although, among speakers showing intermediate values, PD-4 and CTR-2 do not differ). As for the amplitude of the closing gestures for V0 on the vertical component, both factors are significant (Consonant: [F(2,10)=7.869; p<0.05]; Speaker: [F(3,10)=413.182; p<0.05]) with interaction between the two [F(4,10)=10.187; p<0.05]. The post-hoc results show that all speakers and consonant conditions differ from each other: PD-3 shows values at the lowest extreme and PD-4 at the highest, with control speakers showing intermediate values and CTR-1 showing higher amplitude than CTR-2. As for consonants, clusters show the lowest amplitude values and singletons the highest. Considering the horizontal displacement, only the Speaker factor is significant (Consonant: [F(2,10)=0.047; p>0.05]; Speaker: [F(3,10)=31.603; p<0.05]) with interaction between the two [F(6,10)=5.898; p<0.05]. The Tuckey post-hoc showed that CTR-2’s gestures are narrower than other speakers’ and PDs and show the highest average values.

Concerning the opening gesture of V0 vowel /a/ (toward /a/), results concerning the duration of the vertical displacement showed that both factors are significant (Consonant: [F(2,10)=42.802; p<0.05]; Speaker: [F(3,10)=8.282; p<0.05]) with interaction between the two [F(4,10)=7.956; p<0.05]. The post-hoc test showed that CTR-1 has the longest gesture, and on average PD speakers have shorter gestures; as for consonants, the gesture is longer in the cluster condition than in others. With regards to the horizontal component, both factors are significant (Consonant: [F(2,10)=43.419; p<0.05]; Speaker: [F(3,10)=7.235; p<0.05]) with no interaction between the two [F(4,10)=3.329; p>0.05]. Post-hoc results are similar to those reported for the vertical axis, though here CTR-1 differs only from PD-3. As for the amplitude of the opening gestures for V0 in its vertical component (see Fig. 2, left), only Speaker is a significant factor (Consonant: [F(2,10)=1.475; p>0.05]; Speaker: [F(3,10)=24.724; p<0.05]), with no interaction between the two [F(4,10)=25.0; p>0.05]. The post-hoc results show that PD-4 and CTR-1 produce significantly wider gestures than the other speakers, with PD-4 showing greater amplitude values than CTR-1, who shows greater displacement than PD-3 and CTR-2. Considering the horizontal component (see Fig. 2, right), both factors are significant (Consonant: [F(2,10)=16.016; p<0.05]; Speaker: [F(3,10)=234.716; p<0.05]) with no interaction between the two [F(4,10)=3.939; p>0.05]. According to the Tuckey post-hoc, CTR-2 shows smaller amplitudes than others, and, at the other extreme, PD-3 shows greater amplitude than others; finally, the gesture is wider in cluster conditions.

3.2.2. Consonantal gestures and phasing

As for the consonant closing gesture duration, results concerning data on the LL coil showed that both factors are significant (Consonant: [F(2,10)=7.547; p<0.05]; Speaker: [F(3,10)=38.961; p<0.05]) with no interaction between the two [F(4,10)=4.50; p<0.05]. The Tuckey post-hoc showed that PD-3 and PD-4 show shorter gestures than other speakers (who do not differ significantly from each other), with PD-3 showing shorter gestures than PD-4; as for Consonant status, geminates show the greatest duration. Concerning closing gesture amplitude, results show that both factors are significant (Consonant: [F(2,10)=14.727; p<0.05]; Speaker: [F(3,10)=40.672; p<0.05]) with no interaction [F(4,10)=7.65; p<0.05]. The Tuckey post-hoc showed that PD-3’s gestures are narrower than those by other speakers, PD-4 included, who shows values similar to CTR-2 and smaller than CTR-1; as for Consonant status, geminates show the greatest amplitude.

The opening consonantal gestures show duration values that vary significantly depending on the Speaker factor only (Consonant: [F(2,10)=1.429; p<0.05]; Speaker: [F(3,10)=6.201; p<0.05]) with no interaction between the two [F(3,10)=1.095; p>0.05]. The Tuckey post-hoc showed that speaker PD-3 shows shorter gestures than CTR-2 speaker, with the other speakers showing intermediate, not significantly different, values. However, a check for every single speaker showed that the Consonant status is a significant factor for both CTR-1 [F(2,3)=5.5470; p<0.05], with gaminates being longer than singletons and clusters, and PD-3 [F(2,3)=8.888; p<0.05], with clusters being longer than both gaminates and singletons. As for the consonant opening gesture amplitude (see Fig. 3), results showed that both factors are significant (Consonant: [F(2,10)=11.627; p<0.05]; Speaker: [F(3,10)=43.000; p<0.05]) with no interaction between the two [F(4,10)=1.512; p<0.05]. The Tuckey post-hoc showed that speaker PD-3 shows smaller amplitude than other speakers; as for Consonant status, gaminates show the greatest amplitude. Moreover, a check on every single speaker showed that the Consonant status is a significant factor for all speakers, that is CTR-1 [F(2,3)=8.066; p<0.05], CTR-2 [F(2,3)=47.383; p<0.05], PD-4 [F(2,3)=9.028; p<0.05], with gaminates (and clusters) showing greater amplitude than singletons, and PD-3 [F(2,3)=25.230; p<0.05], with gaminates being wider than singletons and clusters.

As for the phasing between the tongue dorsum gesture for /i/ and the lower lip gesture for the consonant(s), results showed that only Speaker is a significant factor (Consonant: [F(2,10)=3.747; p<0.05]; Speaker: [F(3,10)=14.287; p<0.05]) with no interaction between the two [F(4,10)=1.237; p<0.05]. The post-hoc test showed that PD-3 presents the shortest interval in comparison with other speakers and CTR-2 shows the longest, though not significantly different from PD-4. As for the phasing between lower lip gesture for the consonant(s) and tongue dorsum gesture for /a/, results showed no significant factor (Consonant: [F(2,10)=3.30; p<0.05]; Speaker: [F(3,10)=2.547; p<0.05]) with no interaction [F(4,10)=3.324; p<0.05].

Figure 2: Amplitude (mm) of tongue dorsum opening gesture: vertical (left) and horizontal (right) axes

Figure 3: Amplitude (mm) of lower lip opening gesture
4. Discussion

The main goal of the study was to check on the basis of both acoustic and articulatory data whether hypokinetic dysarthria found in Italian PD speech corresponds to consistent features across speakers or rather to cross-speaker variability. The first data available on the articulation of consonants and vowels in Italian PD dysarthric speech were presented, showing interesting features from both an acoustic and an articulatory point of view. As for acoustics, data on the tonic vowel /i/ duration and first formant values show that PD speakers significantly differ one from the other (with PD-3 showing the shortest and most close vowels and PD-4 producing longer and more open vowels). In these cases, the characteristics of PD productions are not consistent and rather match those of the peer controls. In fact, both elderly speakers, PD-4 and CTR-2, show vowels that are longer and more open than those produced by younger speakers (in line with Xue & Hao 2003).

On the other hand, PD speakers group against controls as for both the second formant values of their vowels and the consonant(s) duration. In fact, PD speakers show slightly velarized vowels, in the case of both tonic /i/ and post-tonic /a/, and shorter consonant(s) average duration. Statistical checks on single subject data show that the shortest vowel duration usually observed before geminates and clusters is found only in one control speaker, and that two out of four speakers show a significant difference in consonant duration depending on Consonant status. Importantly, this happens independently from the speakers being controls or PD patients, while the scarce differentiation of vowel and consonant duration in the context of geminates and singletons is attributed to features of the variety of Italian used by speakers, such as the general tendency to strengthen plosives, especially voiced ones (Maiden and Parry 2009). As for kinematics, PD speakers do not always show gestures that would be predicted on the basis of acoustic data (e.g., wider TD gestures along the x axis despite lower F2 values that would theoretically not require a wider anterior-posterior gesture). Moreover, PD speakers show more heterogeneous articulatory rather than acoustic results. Concerning the vocalic gestures, they show consistent results as for the amplitude of the /a/ to /l/ closing gesture on the horizontal plane (which is wider for PDs, especially for PD-4), and as for the duration of the /i/ to /a/ opening (which is shorter in PDs than in controls, in particular for PD-3 vs. CTR-1) both on the vertical and the horizontal plane. Apart from these similarities, PD speakers differ for various correlates, such as for: the duration of the closing gesture, in both vertical and horizontal dimension (with PD-3 showing the shortest gesture in comparison with any other speaker and PD-4 showing much longer duration); the amplitude of the closing on the vertical dimension (with PD-4 showing the widest gestures and PD-3 showing the narrowest in comparison to other speakers); the amplitude of the opening gesture which on the vertical axis is wide for PD-4 and narrow for PD-3, while on the horizontal axis is wider for PD-3 (with PD-4 showing values more similar to controls, especially to CTR-1). As for the consonantal gesture, PD speakers often show smaller values in comparison to controls, and this is particularly true for PD-3. Nevertheless, PD-3 appears to differentiate consonantal opening gesture duration and amplitude according to the consonant status, tending to maintain correlates to linguistic contrast. Finally, PD speakers differ also as for the phasing of gestures, in that the phasing between tongue dorsum gesture for /l/ and lower lip gesture for consonants is timed earlier only in PD-3, while PD-2 shows a later timing, similar the latest timing found in speaker CTR-2.

5. Conclusion

PD speakers may differ from each other as for both acoustic and articulatory correlates and may do that especially for the vocalic gesture amplitude, although not always (not necessarily along all planes or all gestures) and not even consistently with a possible higher degree of impairment (not necessarily smaller amplitude values are found for possibly more impaired speaker, although this is quite often the case).

This is probably due to the realization of different articulatory strategies both intra- and inter-speaker. That could be reflected in the few phasing differences we observed, though our data are too preliminary for a statement in this direction. Importantly, PD speakers may show significant differences depending on the consonant status, even when they show a general reduction in duration and amplitude correlates. This may be favored by the aforementioned possible differences in articulatory strategies.

6. Acknowledgements

We wish to thank the Carstens Med. GmbH for the chance to collect data by means of AG501. A special thank also goes to Francesco Sigona, for the post-processing of EMA data.

7. References


Phonetic characteristics of filled pauses: the effects of speakers’ age

Mária Gósy¹, Judit Bóna², András Beke¹, Viktória Horváth¹,
¹Department of Phonetics, Research Institute for Linguistics, HAS, Budapest, Hungary
²Department of Phonetics, Eötvös Loránd University, Budapest, Hungary
gosy.maria@nytud.mta.hu, bona.judit@btk.elte.hu, beke.andras@nytud.mta.hu,
horvath.viktoria@nytud.mta.hu

Abstract

Filled pauses usually reveal speech planning or execution problems even though the speaker does not produce an overt error and may have a function of discourse marker as well. In Hungarian, the most frequent form of filled pauses is a schwa-like vowel of various durations. The purpose of this study was to analyze the occurrence, duration and formant structure of Hungarian schwa-like filled pauses in 16 nine-year-old children, in 16 young adults and in 16 elderly speakers. Results confirmed age-dependent patterns of filled pauses. Speakers’ age is one of those factors that influence the occurrences and formant values of filled pauses.

Keywords: duration of filled pauses, formant structure, children, young adults, elderly

1. Introduction

Filled pauses are known also as mazes, delay markers, hesitation (phenomena), editing pauses, unlexicalized filled pauses, fillers, interjection, delays, even noises, and they usually reveal speech planning or execution problems even though the speaker does not produce an overt error (Shriberg 2001, Watanabe et al. 2008). The speaker might use filled pauses to gain time for synchronization between thinking and speaking processes. In addition, filled pauses also reflect the speaker’s strategy to control his/her own speech production and they can have pragmatic and discourse functions as well (Clark and Fox Tree 2002, Simpson 2006). The term ‘filled pause’ will be used in this paper as the phenomenon containing a schwa-like vowel in Hungarian with the same functions like uh, umm or er in English.

The main theoretical interest that guided our research was to find out whether the durations and the formants of Hungarian filled pauses are the same or specifically different across ages. The effects of age on filled pauses are not completely understood and, in fact, they are vigorously debated in the literature (e.g., Gayraud et al. 2011). Filled pauses in large corpora were reported to be from one-third to over one-half of all disfluencies (Shriberg 1994). Most studies agree that there are no large differences in the frequency of disfluencies between young and adult speakers (e.g. Leeper and Culatta, 1995), even centenarians’ breakdowns (frequency and types of disfluencies) were similar to those of younger elderly speakers (Searl et al. 2002). On the contrary, elderly people were reported to use a larger number of filled pauses as opposed to young adults (e.g. Kemper 1992, Roggia 2012). Filled pauses are reported to occur in children’s spontaneous speech as early as in three-year-olds (e.g. Furman and Özyürek 2007, Hudson Kam and Edwards 2008). Analysis of children’s spontaneous speech authors concluded that it shows more adult-like disfluency patterns toward the later preschool years (DeJoy and Gregory 1985). In addition, there is a continuous decrease in the occurrences of disfluencies from 6-year-olds to adulthood, however, the proportions of diverse types of disfluencies show differences depending on age (Ito 1986). Nine-year-old Turkish-speaking children were shown to use filled pause (in form of şey) as a planning/hesitation marker and a narrative initiation marker in 62% of all items while they occurred as fillers in 35% of the cases (Furman and Özyürek 2007). The close interaction between the occurrence of filled pauses (and other breakdowns) and grammatical structures of speech samples in children were confirmed in several studies (e.g. Fiestas et al. 2005, Farantouri et al. 2008). In children, the frequency of filled pauses increases with linguistic complexity on the one hand, being greater in longer utterances (Yaruss et al. 1999, Thordardottir and Weismer 2002). Filled pauses are reported to have wide durational range from less than 100 ms to about 750 ms (e.g. Shriberg 2001, Duez 2001, Eklund 2004) but they may frequently have even longer durations. The duration of filled pauses shows a great variety also in children. The average duration was found to be about 800 ms in 4-year-olds’ spontaneous speech samples (MacWhinney–Osser 1977) while nine-year-old Spanish-speaking children had 28% filled pauses of their all analyzed disfluencies (Esposito 2005). There was no significant difference among 3-year-old, 5-year-old and 9-year-old Turkish-speaking children’s speech samples in the occurrence of filled pause şey (Furman and Özyürek 2007). Spanish-speaking six-year-olds’ narratives contained filled pauses in 2.3% (on average) of all produced words while English-speaking children had them in 3.9% of all produced words (Fiestas et al. 2005). Our hypotheses were that (i) filled pauses would show age-dependent occurrences and (ii) age would influence both the durations and formant structures of filled pauses.

2. Methodology

Spontaneous speech samples were analyzed in three groups of: (i) 16 nine-year-old children, (ii) 16 young adults (ages 22 to 28) and (iii) 16 elderly people (ages 75 to 90). All speakers were native speakers of Hungarian from Budapest. Half of them were females while the other halves were males in all age groups. Children were randomly selected from three elementary schools. No hearing or speaking disability was reported among them. Adults were randomly selected with the criteria of their ages and gender from BEA, a large Hungarian spontaneous speech database (Gósy 2012). Their hearing and speech production was appropriate for the ages. The topics of the adults’ narratives included the speakers’ families, work or past work, and hobbies while children spoke about family, school, holiday and hobbies. An average of 8.5 minutes was selected from the middle of each narrative. The samples collected from children were carried out at school in a small quiet room in the morning. Speech samples of all subjects were manually annotated by two of the authors using Praat (Boersma and Weenink 2010). Each filled pause was identified and coded by two of the
authors while the other two authors controlled the coding. No disagreement was found in identification of the filled pauses in speech samples. Functions of filled pauses were neither identified, nor classified. 71.5% of all filled pauses were schwa-like (neutral) vowels. Altogether 1054 schwa-like filled pauses were analyzed in the three groups (249 with children, 523 with young and 282 with elderly participants). The occurrences, durations, and the first two formants of the filled pauses were analyzed. Their duration was measured from the first glottal pulse to the last glottal pause of the vowel. The first two formants were measured at the midpoint of the total filled pause duration. Both duration and formant frequency measurements were carried out using an automatic Praat script followed by their manual verification using auditory feedback if necessary. The script was written for the purpose of this study by one of the authors.

MANOVA and non-parametric tests (Kruskal–Wallis and Mann–Whitney test) (SPSS, version 19.0) as well as three-dimensional Euclidian distance measures were used to test the hypotheses.

3. Results

This study concentrates on measured occurrences, durations and formants of schwa-like filled pauses in Hungarian, where they typically appear as Ő ([u] or [ɔ]) with various durations. 71.63% of all filled pauses in the present material were produced by this vowel while the rest of them consisted of a combination of a similar vowel and a [b]-like and/or a nasat-like consonant or a murmur, for example mm, Öh, Öm, Öhm. For example (filled pause is marked by bold Ő letters): nekem mondjuk Ő in mondjuk fel akarok használni ‘for me let say Ő let say I want to use’. The ratio of schwa-like filled pauses was 78.5% in young adults, 67.6% in elderly and 68.8% in children.

3.1. Occurrences of schwa-like filled pauses

The occurrence of schwa-like filled pauses showed no difference between children (2.7 indices/minute, on average) and elderly participants (2.6 indices/minute, on average) but it was more frequent in young speakers (2.9 indices/minute, on average). Females used filled pauses more frequently (435 indices) than males (487 indices); however, there is almost no difference between nine-year-old girls (122 indices) and boys (125 indices). Both young and old female adults used more filled pauses (2.99 indices/minute and 3.5 indices/minute) than young and old male adults (2.49 indices/minute and 1.6 indices/minute).

3.2. Durations of schwa-like filled pauses

Young adults produced the longest filled pauses while children produced the shortest ones while the duration of elderly’s filled pauses fell within those of the two former age groups (Table 1).

The differences depending on age is significant (Kruskal–Wallis test: Chi-Square (2) = 10.813, p = 0.004). There was no significant difference between nine-year-old girls and boys (according to Mann–Whitney test, p > 0.05). Young adults’ and elderly’s data, however, were proved to be significantly different depending on gender (Mann–Whitney test for young adults: Z = -4.947, p < 0.001 and for elderly: Z = -2.109; p = 0.035), see figure 1.

![Figure 1: The interrelations of duration and relative frequency of filled pauses depending on age.](image)

Nine-year-old boys’ filled pauses were significantly shorter than those of young and old male adults (Kruskal–Wallis test: Chi-Square (2) = 24.5453, p < 0.001), and both adults’ durations differed significantly from those of the nine-year-old boys (Mann–Whitney test between boys and young adults: Z = -4.982; p < 0.001; and Z = -2.941; p = 0.003 between boys and elderly males). No statistical difference was found between the data of the two male adult groups. Although elderly female speakers produced some longer filled pauses than young females and nine-year-old girls, the differences did not turn out to be significant.

3.3. Formants of schwa-like filled pauses

Formant frequency data were analyzed considering the factors of ‘age’ and ‘gender’ (Figure 2). Both first and second formants show differences depending on age. F1-values decrease in parallel with age with both females and males. The F2-values decrease between children and young adults both in females and males but increase between young and elderly speakers. The decrease of the F1s is assumed to reflect the slight rise of the tongue height across ages. The increase of the F2s in elderly’s filled pauses might be assumed to show the consequences of either the higher tongue position in the oral cavity or the less lip-rounding (or both). This modification seems to be characteristic particularly of elderly males.

![Figure 2: Formants of filled pauses depending on age and gender.](image)
The effect of 'age' and 'gender' was statistically analyzed using MANOVA and showed significant effects on the formants (factor of 'age' for F1: F(2, 1044) = 277.985, p < 0.001, η² partial = 0.35; for F2: F(2,1044) = 141.686, p < 0.001, η² partial = 0.21) and factor of 'gender' (for the first formant: F(1, 245) = 11.009 at p < 0.001, η² partial = 0.094; for the second formant: F(1, 245) = 29.269, p < 0.001, η² partial = 0.018). The interaction of 'age' and 'gender' was statistically significant but only for the second formant (F(2, 1044) = 9.571, p < 0.001). The first two formants were modeled by Gauss distribution in order to demonstrate the differences among the three age groups. Children’s formants differed largely from both adult groups’ while the two adult groups’ formant values showed larger overlaps. Tukey post hoc tests revealed that the formant frequency values of the two formants differed from each other in all age groups (p = 0.0001 in all cases).

As expected, the factor ‘gender’ showed significant differences in all adult groups for both F1s and F2s (one-tail ANOVA in the case of young adults for first formants: F(1, 519) = 99.606, p < 0.001; and for second formants: F(1,519) = 347.479, p < 0.001) and in the case of elderly for the first formants: F(1, 274) = 35.126, p < 0.001; and also for the second formants: F(1,274) = 40.331, p < 0.001). Although the girls’ first and second formants showed less differences from those of the boys’ than young adults had, the data turned out to be significantly different even in their cases as well (for the first formants: F(1, 245) = 11.009, p < 0.001; and for second formants: F(1,245) = 29.269, p < 0.001).

The first two formants were modeled by Gauss curve in order to demonstrate the differences among the three age groups. The distance between the groups was characterized in the model by means of the Euclidian distance that was calculated based on the mean value of the distributions of the first two formants. Results confirmed that children’s formants differ largely from both adults groups’ while the two adult groups’ formant values show larger overlaps.

![Figure 3: The Euclidian distance of the formants of filled pauses from those of the (classical) neutral vowels.](image)

We wanted to define the closest filled pause formants to those of the neutral vowel among the age groups. The F1 value of the neutral vowel is 500 Hz while that of the F2 is 1500 Hz (Pickett 1980). The distance of our data from those of the neutral vowel was calculated by means of Euclidian distance from the mean values of the distributions. Results indicate that the formants of the young speakers’ filled pauses are the most similar to those of the neutral vowel, elderly’s formants are less similar while those of children’s show large differences (Figure 3).

4. Discussion and conclusion

Our study aimed to investigate the occurrence, durations and formants of O-type filled pauses in Hungarian. The finding that filled pauses occurred most frequently in young adults as opposed to children and elderly. There was practically no difference in occurrences of filled pauses between our 9-year-old girls and boys. Hungarian-speaking elderly subjects’ narratives contained the least amount of filled pauses which contradicts some findings reported in the literature (e.g. Roggia 2012).

Elderly speakers are frequently reported being as fluent as, or sometimes more fluent than, young speakers, and filled pauses, were shown to increase with age (e.g. Lee and Orlikoff 1999, 2000). The frequency of filled pauses in our material seems to contradict those claims that occurrence of filled pauses increases in elderly (see Roggia 2012).

The hypothesis that filled pauses would show similar durations across ages was partly confirmed. Analysis of the durations disclosed that the difference across ages exists between nine-year-old boys and adult males where children had the shortest filled pauses and young males the longest ones. No such difference exists among females, the durations of the filled pauses are very similar irrespective of age. The durational difference of the filled pauses seems to be rather a factor of gender than a factor of age.

Our study confirmed that ‘age’ and ‘gender’ are decisive factors for the first two formants of the Hungarian schwa-like filled pauses. The elderly participants’ formant values overlap with those of young adults’ as a succeeding that the formants of the schwa-like filled pauses show the acoustic consequences of the articulation in the mid part of the oral cavity. Our findings seem to partly contradict the respective facts reported in the literature. Children’s vowel formants are reported to be higher than those of adults’ (Busby et al. 1995) because of the different shapes and sizes of their speech organs including the developing vocal cords (Baken–Orlikoff 2000). Elderly’s physiological configurations, somewhat centralized articulation, and motor control of their articulation may explain the acoustic differences in their vowel formants (Hooper–Cralidis 2009).

The effect of gender on formants yielded significant differences in all our age groups. There are contradictory findings in the related literature concerning the gender differences appearing in the formant values in children. No differences depending on gender were found in children’s speech until the age of 15 in a study by Lee et al. (1999) while the opposite was confirmed in the study by Busby and Plant (1995) claiming that formant values were higher in females than in males (children’s ages varied between 5 and 11). Vorperian and Kent (2007) found gender differences in formant frequency values by age 4 with more apparent differences by age 8. Since children’s speech organs, including vocal cords, can be considered to be identical at the age of 9 (Simpson 2009) therefore no large gender differences were assumed for the formants of the filled pauses in our children. Although supraglottal anatomy and the vocal tract configurations (particularly vocal tract length and mouth opening size) are relevant for formant frequencies; however, other factors of learned behaviors and the vowel type should also be considered in gender differences of speech beside the age (Busby and Plant 1995, Simpson 2009). We can conclude that by the age of 9 children are expected to show gender differences in schwa-type Hungarian filled pauses. As expected, female formants were in general higher than those of male formants, and the vowel space defined by F1 and F2 is shown to be larger in females than in males irrespective of age (see also Simpson 2009).

Our findings partly supported the claims about the occurrence and durations of filled pauses reported in the literature. The
differences found can be explained primarily by the specific articulation of the Hungarian filled pauses. The vowel-like sound constituting filled pause in Hungarian is the realization of two vowels (the front, labial, mid vowel and the neutral vowel) that are both members of the Hungarian vowel inventory but they have different phonemic status.

5. Acknowledgements

This study was supported by OTKA project No. 108761, and the Bolyai János Research Scholarship.

6. References


Clark, Herbert H. and Jean E. Fox Tree (2002). “Using uh and um in spontaneous speaking”. In: Cognition 84, pp. 73–111.


Phonetic Properties of Hungarian Hiatus Fillers versus Intervocalic Palatal Approximants: Physical Realities and Theoretical Implications

Máría Gósy

Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
gosy.maria@nytud.mta.hu

Abstract

Hiatus rules of Hungarian include various ways of resolving vowel sequences, such as adding the consonant [j]. Our research question is whether there are differences, either in durational or spectral domains, between the two [j] consonants depending on their function (hiatus filler vs. palatal approximant). Thirteen young female native speakers read 80 sentences that contained the target words. The duration and the first two formants of [j] were measured. The mean duration of [j] was 58.5 ms in VJV sequences while those of hiatus fillers was 52 ms between the same vowels. The F1 values of [j] were similar irrespective of the function; however, the F2s differed significantly depending on function. We can conclude on the basis of acoustic-phonetic data that our speakers reading various sentences did not encounter any functional or contextual pressure to produce hiatus fillers largely differently from the corresponding [j] consonants. Thus, the hiatus filling in the analyzed vowel sequences seems to be a specific type of coarticulation.

Keywords: hiatus filling, durations, formant frequency values

1. Introduction

In connected speech, neighboring speech sounds may influence the pronunciation of one another to a variable extent: this array of phenomena is known as coarticulation. Coarticulation phenomena are determined by a number of physiological, phonetic, and phonological factors, as well as specific mechanisms that are characteristic of the language at hand (e.g., Beddor et al. 2002). Research over the past few decades has confirmed that the mutual influence of speech sounds may also apply across syllable boundaries and word boundaries, and even across intervening segments (e.g., Farnetani and Recasens 1999, Manuel 1999).

The occurrence of two adjacent vowels, or rather the gap between them, is referred to as hiatus both in phonetics and in phonology, in cases where the two vowels constitute two separate syllable nuclei rather than a diphthong (Siptár 2012). The occurrence of vowels in two neighboring syllables can be found in a number of languages, although some other languages exclude that possibility (see Kabak 2007, Odden 2006). The general issue of hiatus and hiatus resolution can be analyzed both in a phonetic and a phonological perspective; typically, the literature includes more papers on its phonological aspects (e.g., Kabak 2007, Padgett 2008, Siptár 2012) than on phonetic ones (e.g., Aguilar 1999, Markó 2012). Hiatus resolution strategies are language specific, meaning that individual languages may have different ways of realizing vowel clusters, with transitions or with various strategies of resolution or elimination (Casali 1998, Hardcastle 2006, etc.). A given language may have several mechanisms or strategies of hiatus resolution; the issue may also concern the grammar (morphology) of a language in addition to its phonetics and phonology.

The realization of pairs of adjacent vowels, among other language-specific factors, depends on the quality of the vowels concerned in Hungarian, too. In principle, three possibilities of hiatus resolution are available: the deletion of one of the vowels; the insertion of a segment between the two, and the transformation of one of the vowels into a semivowel (Siptár 2012). Siptár and Türkény (2000) offered a detailed theoretical account of the criteria of hiatus resolution by glide insertion and stated, among other things, that whenever one of the vowels was either [i] or [iː], speakers obligatorily resolved the hiatus by inserting a [j]-like transitional segment. They added that this was independent of whether the VV cluster occurred intramorphemically or across a morpheme or word boundary (pp. 283). In addition, they claimed that the quality of the ‘hiatus filler’ may differ from that of a realization of the phoneme /j/ in that the former may be weaker or more transitional than the latter. This distinction, they claimed, was mainly observable in what is called ‘careful speech’; they allowed for the possibility that it was hardly noticeable in fast or casual speech styles. In their analysis, they attributed the phonetic differences to the phonological claim that the former was a glide whereas the latter was a liquid. Markó (2012) presented measurements concerning the specific pronunciations of various types of vowel clusters in read speech, with respect to positions of occurrence, and she was the first to report on the occurrence and assumed function of irregular phonation in cases of hiatus.

The present paper addresses the issue of whether hiatus fillers and realizations of the phoneme /j/ only differ in their function or whether that functional difference is also mapped onto a distinction in pronunciation. If it is (that is, if the articulation gestures involved differ sufficiently to result in detectable acoustic differences), the acoustic differences between hiatus fillers and phoneme realizations will demonstrate that they do differ in terms of articulation, too. The aim of this paper, then, is to present an acoustic analysis of the assumed difference between hiatus fillers and phoneme realizations in vowel clusters involving [i] and either [ɔ] or [ɛ]. In view of the fact that such pairs of vowels may also be separated by (underlying) palatal approximants in this language, the two types of surface representations can be analyzed in identical contexts, e.g. in idek ’yours-pl’ vs. idek ‘frightening’. We tried to find out whether realizations of /j/ do indeed differ from hiatus fillers in their physical durations and/or in the values of their first and second formants. Our hypothesis followed the phonologists’ claim, that is, we expected to find (i) significant differences in all relevant parameters between the two types of [j] across functions, but (ii) no differences depending on whether the consonant at hand occurred word internally or across a word boundary (in either of the two functions).

2. Participants, material, method

Thirteen young female native speakers of Hungarian (between 22 and 32 years of age) participated in the experiment. All of
them had normal hearing and no speech defect; all came from Budapest and were university graduates. The articulation rate of their performance in sentence reading was between 9.8 sounds/s and 11.2 sounds/s; the mean rate was 10.4 sounds/s. We collected bi- and tri-syllabic words, compounds, and phrases (total of 80 items). The items either contained a hiatus ([V][j)V] or a pair of vowels separated by the palatal approximant ([V][j)V].

The vowel clusters consisted of [i] or [e] or [i] and [3]; the [V][j)V sequences also contained these vowels. The vowel clusters (represented orthographically), then, were as follows: ia, ai, ie, ei. The [V][j)V sequences, accordingly, were ija, aij, ije, eij. Word internal VV and [V][j)V sequences were collected into one category, including items like laikas ‘lay’, tied ‘yours’, kijavit ‘correct’, fejik ’theyl milk’. Another category was made up by items involving a word/compound boundary, e.g. almaital ‘apple drink’, hajillat ‘smell of hair’. We made sure that the words and phrases (i) should contain the target sequences either at the boundary of the first and second syllable or at that of the second and third syllable (whether this was a word internal position or one involving a word boundary), and (ii) should be familiar words/phrases often occurring in everyday situations. We had five items in each type, 10 in each category. The whole list ran into 80 items; read out by the 13 subjects, this resulted in 1040 tokens to be analyzed. We embedded the items into carrier sentences of three or four words and of similar length in terms of syllables (6 to 8). The participants’ task was to read each sentence as a single breath group. Before recording began, the subjects had time to read the sentences for themselves, and even to practice reading them aloud. The sentences were recorded in a sound-insulated booth, digitally, direct to the computer, with the GoldWave sound editing software, with 44.1 kHz sampling, and saved in a 16 bit format. 86 kbytes/s, mono. The recording microphone was AT4040.

The recorded material was analyzed (and checked) by the author and another trained phonetician. Of the phonetic properties of the clusters/sequences, the individual durations of their three segments and the first and second formants of the hiatus fillers or /j/ realizations were measured in Praat (Boersma and Weenink 2010). In order to measure the durations, we had to demarcate the hiatus filler or the [j] from both the preceding and the following vowel. We used hearing-based segmentation and also the Praat settings in defining the boundaries of hiatus fillers and /j/ realizations.

The main criterion of segmentation was defined as the changing point of the second formant (both in terms of the automatic results of Praat and of the analyt’s own visual impression); if necessary, oscillographic information and changes of the first formant were also taken into consideration. The determination of the beginnings and ends of vowels was also assisted by the acoustic representation of the preceding/following consonants; here, as conventional, we considered the first (last) oscillations of the second formants to be decisive. In cases where more accurate delimitation was called for, we also considered changes in intensity and/or fundamental frequency. In all cases, we also checked the sound boundaries perceptually. Formants were measured at midpoints of the durations of hiatus fillers and /j/ realizations, respectively, and this was complemented by FFT analyses. For statistical analyses, we used repeated measures ANOVA and correlation analysis in SPSS (19.0), at a confidence level of 95%.

3. Results

In the statistical analyses that follow, we refer to the difference between hiatus fillers and phoneme realizations as the factor of ‘function’. Both in analyses of duration and in those of first and second formants we also studied the factors of ‘cluster’ and ‘boundary’. The former refers to whether the quality of the front vs. back low vowels or the order of those vowels and [i] had any effect on the values of the parameters studied. The factor ‘boundary’ refers to whether the position of the sequences under study (whether they occur word internally or across a word boundary) influenced the values that those parameters exhibited.

3.1. Durations

The initial question concerned the values of physical duration of the hiatus filler and the palatal approximant. Our data revealed that (disregarding any other factors) physical durations significantly differed in terms of ‘function’ (F(1, 104) = 69.417; p < 0.001; partial $\eta^2 = 0.42$). This factor relatively strongly correlated with durational differences. The mean duration of hiatus fillers was 52 ms, while that of /j/ realizations was 58.5 ms. The minimum value was 32.6 ms for hiatus fillers and 39 ms for /j/ realizations, the maximum value was 76.2 ms for the former and 89.2 ms for the latter (Fig. 1: left side).

The factor ‘cluster’ had no significant effect on durations. The values measured were rather similar across all types. Both hiatus fillers and /j/ realizations were somewhat longer when [i] was the first vowel or the cluster or preceded (rather than followed) the /j/. However, the differences were between 1 and 4 ms only. Depending on the position of the [i], the minimum values of [j] durations were higher in aij and eij than in ija and ije; but no such difference was found in ai vs. ia, and the hiatus filler of ei was shorter than that of ie. Differences were very small again, in phoneme realizations they did not reach 4 ms, and in the case of ie they averaged on 4.6 ms. The tendency of maximum values was similar; the difference in ajija was somewhat larger (7.4 ms) than in eijije (2.8 ms), and that of aija/aia was 6 ms. The maximum value of ei (just like its minimum value) was lower than that of ie, the difference was 1.6 ms.

The factor ‘boundary’, on the other hand, did have a mathematically significant effect on the duration of hiatus fillers and /j/ realizations (F(1, 104) = 4.653; p = 0.033; partial $\eta^2 = 0.046$). But this factor accounted for the differences very slightly (below 5%). The duration of word internal hiatus fillers was 53.5 ms, that of ones across a word boundary was 50.6 ms; the duration of word internal /j/ realizations was 58.3 ms, and across a word boundary it was 58.6 ms. Irrespective of the quality and order of vowels, /j/ realizations were longer across a word boundary than word internally, while hiatus fillers were shorter across a word boundary than within words. The difference of averages was 3 ms for /j/ realizations, that is, there was practically no difference between the durations within and across words. The distance between hiatus fillers’ averages was also very small, 2.9 ms; speakers produced hiatus fillers so much longer within words on average (Fig. 1: right side).
The duration of hiatus fillers and realizations of the phoneme /j/ was determined by both the type of cluster and its position ($F(3, 104) = 4.734$; $p = 0.004$; partial $\eta^2 = 0.129$). However, with respect to the two functions of [j], neither ‘cluster’ nor ‘boundary’ yielded significant differences, meaning that irrespective of the type of cluster and of the position, there was no mathematically supported temporal difference between hiatus fillers and /j/ realizations. Both hiatus fillers and /j/ realizations were slightly longer within words than across words. But this was only true for cases of ai, aij and ia, ija. In eji and ije cases, phoneme realizations exhibited just the opposite trend: word internal consonants were longer here. Hiatus fillers also differed depending on cluster type: in ei items the consonant was longer within words, whereas in ie clusters it was longer across word boundaries. The differences were very small everywhere: just a few milliseconds.

The temporal relationships between hiatus fillers and phoneme realizations may also be characteristic in how their durations relate to those of the full VCV sequences. Our data showed that there was no statistically relevant difference between the two functions in relative durations. The mean proportion of /j/ realizations within the full V[j]V sequences containing them was 30.2% (SD 3.98%), whereas that of hiatus fillers was 30.6% (SD 3.99%). Relative minimum and maximum values were also very similar in the two functions: in /j/ realizations they were 22.1% and 41.4%, while in hiatus fillers they were 21.6% and 38.9%. The paired factors ‘function’ and ‘cluster’ or ‘function’ and ‘boundary’ did not have any mathematically confirmed effects on relative durations, either. We also analyzed the combined effects of ‘function’, ‘cluster’, and ‘boundary’, and we failed to receive significant differences here, too. Thus, the relative durations of the consonants within those of the whole VCV sequences were the same (or at least statistically non-distinct), irrespective of their function or of any other factor.

### 3.2. Formants

The factor of ‘function’ had no effect on the values of first formants (there were no significant differences there). The mean of first formants of all /j/ realizations was 525 Hz (std. dev. 48.39 Hz), and that of all hiatus fillers was 529 Hz (std. dev. 45.47 Hz). For the former, the minimum value was 386 Hz, and the maximum value was 621 Hz; for the latter, the minimum of first formants was 407 Hz, and their maximum was 626 Hz.

The factor ‘cluster’ had a significant effect on F1 values ($F(1, 104) = 5.841$; $p = 0.001$; partial $\eta^2 = 0.154$). This means that both the location of the [i] and the quality of the other vowels affected the first formant values of /j/ realizations and hiatus fillers. In the case of phoneme realizations, first formants were almost identical in aij/ia cases, the latter being 10 Hz lower on average. In pronunciations of eij/ie, the differences in F1 values were somewhat larger, but here ije exhibited the higher value, and the differences were again small, around 20 Hz. With hiatus fillers, the tendency was just the opposite for ai/aia the value of the first formant was higher for ai, the difference being 22.2 Hz. Pronunciations of ei/ie were similar to those of eij/ije; in the cases of hiatus fillers, too, the latter type of items had a higher F1 value, the difference was 34 Hz. First formants of the consonants at hand were higher where the [i] came first in the sequence. In hiatus fillers, first formant values exhibited a more consistent tendency than in phoneme realizations, irrespective of the vowel qualities or of the location of the [i].

The factors ‘function’ and ‘boundary’ were both crucial for the pronunciation of the consonants under study. With respect to values of first formants, we found significant differences ($F(1, 104) = 10.937$; $p = 0.001$; partial $\eta^2 = 0.102$). The distance between average values of phoneme realizations and hiatus fillers was 6.2 Hz within words; across word boundaries, that value was somewhat larger, 13.6 Hz. In /j/ realizations, the mean difference of first formants between positions was 13.7 Hz, in hiatus fillers it was 6.1 Hz. We can conclude that function affects these values but slightly; the first formants of phoneme realizations were somewhat higher word internally, whereas those of hiatus fillers reached higher frequencies across word boundaries – but the differences were very small in both functions.

Statistical analyses confirmed significant differences in the values of first formants when all three factors – ‘function’, ‘cluster’, and ‘boundary’ – were taken into consideration ($F(3, 104) = 9.290$; $p = 0.001$, partial $\eta^2 = 0.225$). However, within the function-based groups ( /j/ realization vs. hiatus filler) neither cluster type nor word boundary had a significant effect on first formant values.

A statistical analysis of second formants showed significant differences with respect to ‘function’, that is, the difference in F2 values between phoneme realizations and hiatus fillers can be statistically confirmed ($F(1, 104) = 10.841$; $p = 0.001$, partial $\eta^2 = 0.101$). The mean of second formants in all occurrences of /j/ realizations was 2218 Hz, and in those of hiatus fillers, it was 2189 Hz. Minimum and maximum values of the former were 1947 Hz and 2609 Hz, those of the latter were 1819 Hz and 2584 Hz. The averages differed by 39 Hz; second formants of the phoneme realizations were slightly higher (Fig. 2: left side).

![Figure 2: F2 values of phoneme realizations (][j]) and hiatus fillers (hf) (left side) and F2 values within words and across a word boundary (right side), means and standard deviations.](image)

The combination of the factors ‘function’ and ‘cluster’ did not influence our data; differences in formant values were vanishingly small. The mean values of second formants of /j/ realizations in aij/ia differed by 43 Hz, and in eij/ie they differed by 36 Hz. Second formants of hiatus fillers differed across cluster types by a mere 4 Hz. Minimum and maximum values differed across sequence types in a particular manner. The F2 maximum of aij was higher by 90 Hz than that of the consonant of ija, while the highest second formant of eij was 90 Hz lower than the F2 of ije. The minimum value of the second formants of hiatus fillers was almost 90 Hz higher in ai than in iia, whereas the maximum value of the latter was 83 Hz higher than that of the former. The minimum value of hiatus fillers in ei was higher by 112 Hz than in ie, while maximum values exhibited a difference of 143 Hz (with the higher value in ie). The frequency ranges of F2 tended to differ slightly across functions. In the case of phoneme realizations, minimum values were practically identical; the difference between maximum values came close to (but did not reach) 100 Hz. In hiatus fillers, both minima and maxima differed more spectacularly across cluster types. Minimum F2s were higher for ai and ei, while maximum values were higher in iia and iie; a kind of leveling can be observed here. If we compare maximum values across functions, we see that the tendency is identical in eij/ie and ei/e but a contrary tendency can be observed in aijai and iaja.
If we disregard the order of vowels, the difference of /j/ realizations in the contexts of front vs. back low vowels was 211 Hz on average, and that of hiatus fillers was 237 Hz, but the differences did not turn out to be statistically significant.

The factor ‘boundary’ confirmed significant differences for F2 frequency values (F(1, 104) = 14.989; p = 0.001; partial η² = 0.135), meaning that the values of second formants in word internal vs. word external cases differed statistically. The mean difference of F2 of /j/ realizations was a mere 22 Hz, the marginally higher value was found across word boundaries. On the other hand, the mean value of second formants of hiatus fillers was slightly lower at word boundaries than word internally; the difference was 49 Hz. That is, the tendency was contrary between functions. The mean values of second formants of phoneme realizations and hiatus fillers were almost identical (the difference was 6 Hz); across word boundaries, the mean F2 of phoneme realizations was higher by 65 Hz than that of hiatus fillers (Fig. 2: right side).

A combination of the factors ‘function’, ‘cluster’, and ‘boundary’ did not have a significant effect on the second formants of consonants in the sequences. Within the groups separated by function, neither ‘cluster’ nor ‘boundary’ turned out to be significant. Mean values of the second formants of both phoneme realizations and hiatus fillers differed between consonants occurring within and across words by less than 100 Hz in all cluster types. The largest difference (78 Hz) was found in the F2 of the hiatus filler of ia between word internal and word external cases, with a higher value in the word internal consonants.

Neither the durations nor the second formant values were affected by the type of vocalic context in either function. The first formants, however, were significantly affected by the position of [j] (F(1, 104) = 11.134; p = 0.001; partial η² = 0.10).

4. Discussion and conclusion

The results basically confirmed our hypotheses in that durations and formant values did statistically differ in terms of the factor of function. However, the results revealed minute differences only; in durations, the differences were below 10 ms, in first formants, they were below 50 Hz, and in second formants, they were below 100 Hz (but were nevertheless statistically significant). Disregarding concrete numerical results, it was only between functionally different [j] durations that remarkable differences were found. Cluster types did not affect durations, and although the factor ‘boundary’ accounted for significant differences, it only explained those differences in 4.6%. Taking all three factors into consideration, durational differences were significant (F(3, 104) = 4.734; p = 0.004), and accounted for this result in 13%. First formants showed no significant differences in terms of function, but they did in terms of the factors ‘cluster’ and ‘boundary’. On the other hand, the properties of the cluster or sequence had no effect on F2 values. Our conclusion must be a kind of paradox, given that acoustic data differ but very slightly, hence the qualitative difference of the consonants at hand with respect to their function is not convincing. It is less probable that the temporal and formant value differences reach the threshold of perceptibility although this may also depend on the speech sound and on the listeners’ task (Fourakis and Port 1986, Warner et al. 2004). The data do not unambiguously support the claim that hiatus filler differs from that of other palatal approximants; we have to consider all instances of [j] to be palatal approximants whose duration and formant structure vary in a broad range in both functions.

Further research might give us some clues concerning the effect of word frequency or whether larger context does or does not influence the pronunciation of yods; or what kinds of individual pronunciation habits can be forecast (Cole et al. 2010). An especially important factor to be studied later on would be the effect of speech style.

Although a number of assumptions have been offered in the literature as to why speakers should produce an extra segment between two adjacent vowels (in various languages), the most probable explanation, after all, may be that they strive for ease of articulation. It is also quite plausible that speakers would employ as hiatus filler a segment that is part of the sound inventory of the given language anyway. This is likely to have a positive effect on the improvement of articulatory gestures. The fact that we have failed to find large important differences between classes of items pertaining to the two functions of [j] suggests that speakers do not try to signal definitely a functional distinction by pronouncing their [j]’s differently.

5. Acknowledgements

This research was supported by OTKA project No. 108762.

6. References


Multiple bursts in Hungarian voiceless plosives and VOT measurements
Tekla Etelka Gráczi 1, Anna Kohári 2

1 Research Institute for Linguistics, HAS, Hungary
2 Eötvös Loránd University, Department of Phonetics, Hungary
graczi.tekla.etelka@nytud-mta-hu, koharianna@gmail.com

Abstract

Stops can occur with no, one or more bursts. The main determining factor of this phenomenon is the place of articulation, however further factors were also found to influence the release of the stops. Some studies measure VOT-values from the most intense bursts, while most of them do not indicate the start point of the measurement. Intervocalic voiceless unaspirated stops were analyzed in ten speakers’ reading in three vowel contexts. Besides differences across the places of articulation, interspeaker variability was also found in the release types. The interaction of closure and release duration, and the factor speaker with the appearance of multiple bursts showed different tendencies in the three stops. Three VOT-measurements were analyzed (from the first, the most intense and the last burst). Their difference was significant at all places of articulation and among the speakers, though all the analyzed factors of VOT-values showed similar interactions in all three methods.

Keywords: multiple bursts, stops, VOT, Hungarian

1. Introduction

The release of stops may be various. It can be realized without detectible burst, or with one or multiple bursts (henceforth: MB) (Olive et al. 1993). Their realization depends on several factors. Velar stops more often appear with MBs (e.g. Keating et al. 1980, Olive et al. 1993), than anterior ones, however, in the case of bilabials either the lack of detectible bursts (Gráczi and Kohári 2012) or also the MBs (Savitrhi 1989) were found to be more frequent than for alveolars. The speakers’ characteristics also play a role in the realizations. Children with 1;6 show similar burst-patterns to adults, but MBs are more frequent in their speech. Sönmez et al. (2000) found that the posterior the PoA the more frequent MBs in English and Polish, while Savatrhi (1989) claims that in Kannada bilabials also tend to realize more often with MBs than alveolars.

VOT is measured from the release (Lisker and Abramson 1964). However, MB-realization may rise questions. As Fant (1969) notes “Ambiguity often arises as what is the true release transient of palatal and velar stops.” Some studies say that they measure from the most intense burst (e.g. Fuchs 2005, Lousada et al. 2010), while most studies do not mention where they measure the VOT from. The question may arise what differences the chosen measurement can cause in the results. To answer this question we analyzed three voiceless (unaspirated) Hungarian plosives.

2. Methods

Ten native speakers of Hungarian (6 females and 4 males; between 20 and 27 ys) participated in our experiment. None of them reported any hearing problem or showed any speech disorder. They read aloud nonsense words of IVCVl structure in a carrier sentence (A képernyőn a XXX alak látható. ‘The word XXX can be seen on the screen.’) The intervocalic consonant in the target word was /p, t, k/, the two neighboring vowels were identical /i, u, a/. The sentences were randomly displayed by SpeechRecorder (Draxler and Jänsch 2004), and read aloud four times each. This gives altogether 360 sentences. One repetition had to be eliminated. In this case /litil/ sequence was either realized as [litiil]. Though /i/ exists in Hungarian, the authors could not decide based on perceptual and visual inspection of this item whether it was a misreading or an alternative realization.

The recordings were made in a sound treated room with AT4040 microphone at a fixed appr. 40 cm distance from the speakers’ mouth, at 44.1 kHz and 16 bit. The speech material was manually labeled (Figure 1) in Praat 5.3 (Boersma and Weenink 2013). The start of the closure, the start of the following vowel, the first, the most intense and the last burst, and the restart of voicing was annotated. The number of bursts and in case of no bursts the possible alternative realization (e.g. fricative-like pronunciation) were also documented.

The following parameters were calculated: i) duration of the consonant (from closure start to the start of the following vowel); ii) duration of the closure (to the first burst), iii) duration of the entire release (from the first burst), iii) VOT from the three labeled bursts. The relation of the number of bursts and these duration data, and the relation of the three VOT-measurements were attested with SPSS 19.0. Crosstab analysis; Mann-Whintey U-test and Kruskal-Wallis-test were run with Monte-Carlo simulation; and Repeated measures with Bonferroni correction.
3. Results

3.1. Number of bursts

No burst was detected in 5.0%, one burst in 47.1%, two in 33.7%, three in 12.8%, and four in 1.4% of the data. When no burst was detected, /k/ realized once as a fricative, and weak frication appeared as release 11 times (/p/, /f/). In those cases (4 /l/ and 2 /p/) the stop realized with a closure phase and no detectible aperiodic part.

As Table 1 shows /p/ and /k/ was realized the most frequently without any burst, that in the former mostly meant closure and no burst, while in the latter friction was the most often alternative type of realization. /p/ appeared rarely with more than two bursts, while /k/ was the most prone to this effect. These results agree with the most of earlier studies. The number of data required to run the statistics on a two-group model where the distribution of one-burst and the MBs categories could be attested. The Crosstab analysis proved a significant difference among the PoAs (\(\chi^2(2, 341)=39.823, p=0.001\), Cramer’s V=0.342). /p/ had one burst 79 times, and MBs 33 times, while /k/ had one 31 times and MBs 79 times. In the /l/ realizations it was equal (59, 60, respectively).

Table 1: The frequency of occurrence (PoA) of MBs and no burst at the analyzed PoAs.

<table>
<thead>
<tr>
<th>PoA</th>
<th>Number of bursts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>/p/</td>
<td>6.7</td>
</tr>
<tr>
<td>/t/</td>
<td>0.8</td>
</tr>
<tr>
<td>/k/</td>
<td>7.6</td>
</tr>
</tbody>
</table>

The vowel context showed some effect on the appearance of MBs in the case of /k/. In the context of /i/ 77.5% of the realizations had at least two bursts, and four bursts occurred only this context. In the case of the other PoAs no similar tendencies were found. In general, the factor vowel did not show a significant effect on the appearance of one or MBs according to the Crosstab analysis (\(\chi^2(2, 341)=1.104, p=0.586\), Cramer’s V=0.057), neither when attesting the results for the consonants separated (/p/, /f/, /t/, /k/) (\(\chi^2(2, 112)=6.55, p=0.447\), Cramer’s V=0.122; /t/: \(\chi^2(2, 119)=2.008, p=0.380\), Cramer’s V=0.130; /k/: \(\chi^2(2, 110)=2.525, p=0.273\), Cramer’s V=0.152). Six speakers had 1-6 stop realizations without any burst. The most often non-burst opening was found in two speakers’ speech: a male participant’s releases were frications in 11.1% and a female speaker in 8.3%, and she also had non-released stops in 8.3%. MBs appeared in 36-61% of the stop realizations of each speaker (Figure 2). Four bursts were rare (0-5.5% per speaker), while the frequency of appearance of two- or three-bursts were different among them. While the stops of m4 and f4 had MBs in less than 40% of their speech, in the case of m1 and f2 MBs appeared in over 60% of their stops. When also considering the PoA, we can find great differences among the speakers (Figure 3). In seven speakers the highest ratio of /k/ was realized with MBs, in the case of other three participants’ /h/ realizations showed higher MBs than their /k/ ones. In four speakers’ pronunciation /p/ was not less frequently realized with MBs than /h/ or /k/. The speaker itself did not show a significant effect on the appearance of one or MBs according to the Crosstab analysis (\(\chi^2(9, 341)=7.687, p=0.382\), Cramer’s V=0.169). However, when separating the results for the PoAs the factor speaker proved to play a role on the two-group categorization in the anterior PoAs, while not for /k/ (\(\chi^2(9, 112)=19.175, p=0.021\), Cramer’s V=0.414; /i/: \(\chi^2(9, 119)=26.781, p=0.001\), Cramer’s V=0.474; /k/: \(\chi^2(9, 110)=11.218, p=0.272\), Cramer’s V=0.319).

Table 2: Tendencies of IBs being identical with the first or last burst

<table>
<thead>
<tr>
<th>PoA</th>
<th>2 bursts</th>
<th>3 bursts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tendency</td>
<td>tendency</td>
</tr>
<tr>
<td>/p/</td>
<td>31</td>
<td>100%</td>
</tr>
<tr>
<td>/t/</td>
<td>43</td>
<td>15%</td>
</tr>
<tr>
<td>/k/</td>
<td>47</td>
<td>29%</td>
</tr>
</tbody>
</table>

Table 3: Frequency of MBs being identical with the first or last burst across vowel contexts in the 2-burst realizations

<table>
<thead>
<tr>
<th>PoA</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>16</td>
<td>78.6</td>
</tr>
<tr>
<td>/t/</td>
<td>12</td>
<td>56.3</td>
</tr>
<tr>
<td>/k/</td>
<td>16</td>
<td>76.9</td>
</tr>
</tbody>
</table>

In the case of the MB-releases, the most intense burst (henceforth: IB) can be identical with the first, or the last burst, or it can appear as a separate one (in case of at least three bursts). In the case of the two-burst realizations the anterior PoAs showed a preference of the IB being identical with the first burst, while the velar did not (Table 2; Crosstab analysis: \(\chi^2(2, 121)=13.481, p=0.001\), Cramer’s V=0.334). IB did not tend to be identical with the last burst either in the few four-burst cases (/p/: 0, /t/: 2, /k/: 3). The three-burst realizations, however, showed the tendency that the IB was never the first burst. Some effect of the neighboring vowel was found at the posterior PoAs, however, its effect is not significant in the two-burst cases (Crosstab analysis: \(\chi^2(2, 121)=0.145, p=0.939\), Cramer’s V=0.035). The analysis of IBs by speakers shows diffusse picture. Some tended not to show
“preferences”, while others’ IBs followed the results of the PoAs in Table 2 and 3.

3.2. Relation of duration patterns and the characteristics of release

Our question was whether the duration patterns of the stops (duration of closure, release and the consonant) had any interrelation with the type of the release. All three measured values were different across the PoAs and the vowel context according to the Kruskal-Wallis test (χ² was between 13.296 and 115.937, p<0.001). Attesting the effect of the vowel context at the different PoAs, all three duration parameters showed a significant difference at each PoA (χ²(2, 120)=2.250, p=0.325). The interaction of three durations and the one-burst or MB categories were analyzed (Table 4 and 5).

The results of the three measurements show a significant effect of the PoA (F(2, 339)=9.541, p=0.049, η²=0.234). The results of the three measurements showed a significant difference at each PoA (F(2, 120)=2.250, p=0.013). Attesting the effect of the vowel context with the type of the release. All three measured values were different across the PoAs and the vowel context according to the Kruskal-Wallis test (χ² was between 13.296 and 115.937, p<0.001). The interaction of three durations and the one-burst or MB categories were analyzed (Table 4 and 5).

The results show that the duration of the closure was longer, and that of the release was shorter in the one-burst realizations of /t/ than in its MB-realizations. The Mann-Whitney U-test was run on the data separated by the PoA and the vowel contexts, as well. In the case of /k/ no differences were found of the two groups in any contexts, while in /p/ in the /u/ context the duration of the closure was longer (Z=-2.302, p=0.020), and in the /i/ context these the former was shorter, the later was longer in the MB-realizations than in the one-burst ones (Z=-2.391 and -2.698, p<0.001).

3.3. Relation of VOT and multiple bursts

In the VOT analysis the no-burst realizations were eliminated, and also one realization of /lutul/ where the following vowel remained voiceless during its entire duration. The number of analyzed tokens is shown in Table 6.

The VOT-values measured from the three analyzed bursts can be more or less different due to the distance and the number of the MBs, and the IB-tendencies. Figure 4 shows the difference of the three measurements methods independently of the number of bursts of the realizations. (These values are of course equal with the distance of the analyzed bursts.) In the case of MBs, /p/ showed a difference of the first and the IB based VOT-values between 0-7.9 ms, that of the IB and the last burst measurement was between 0-10.0ms, and that of the first and the last burst based VOT ranged between 0-10.0ms. In the same order for /t/, these values are: 0-20.0ms, 0-12.5ms, 0-20.7ms, and for /k/: 0-32.3ms, 0-19.7ms, 2.1-37.4ms.

\[ \chi^2(2, 120)=2.250, p=0.325 \]

\[ \chi^2(2, 120)=2.250, p=0.013 \]

\[ Z=-2.302, p=0.020 \]

\[ Z=-2.391, p=0.013 \]

\[ Z=-2.698, p<0.001 \]
\[ \eta^2=0.055 \], the speaker (F(2, 339)=1.627, p<0.001, \eta^2 =0.071),
and their interaction (F(2, 339)=1.545, p=0.025, \eta^2 =0.071)
was not proven to play a significant role in this analysis.

Though the test proved the results of the three measurements
to significantly differ from each other, the main effects that
were found in earlier studies to affect the VOT-values (e.g.
Keating et al. 1980, Barry and Moyle 2011, Cheung and Wee
2009) have the same tendency of effect on each measurement
according to the Between Subjects Effects (Table 7).

4. Discussion and conclusion

The posterior PoAs showed higher number of MBs, and also
more bursts. These results agree with earlier studies. The
present analysis showed interspeaker differences in the
appearance of MBs at the anterior PoAs, while not in the case
of the velar voiceless stop. The variability may be explained
by the anatomical-aerodynamic and the temporal thus
durational differences among the speakers mentioned in the
Introduction (see Imbrie, 2005, Song et al. 2012). The
frequency of occurrence of the MBs did not show any
interaction with the vowel context of the analyzed stops.
Possibly because the differences among the articulatory
distances of the vowels do not affect the pressure increase
during the consonant, thus its own characteristics and
the speaker’s specific articulation determine the results. The lack
of the interspeaker variability in the occurrence of MBs in /k/
suggest that the main factor of this phenomenon is the PoA.

The IB of the stops with two and with three bursts showed
different tendencies. In the former ones usually the first burst
was the most intense one with a difference among the PoAs,
while in the latter ones the first burst was never the most
intense one. The differences across PoAs in the two-burst
cases might have articulatory-aerodynamic reasons. Maybe
the first burst in the velars is more frequently caused by the
intraoral pressure (see the suggestion by Keating et al. 1980),
while at the other PoAs the first burst might be also resulted
by muscular activity. This hypothesis, however, needs
articulatory investigation.

The duration patterns of the stops showed a diverse picture. In
the case of the velar stop no interaction was found, while in
the alveolar the duration of the closure and the release varied
between the one-burst and the MB-realizations. In the case of the
latter, the longer closure phase might result in higher
intraoral pressure causing MBs, and the release might be
shorter due to some compensation of the duration. In the case
of velar, the larger surface, the larger inertia, and the higher
intraoral pressure (due to the smaller volume) might result that
the duration of the closure does not influence the realization
like at the alveolar PoA. The greater intraoral cavity and thus
the slower pressure build-up in bilabials might also support
this hypothesis.

The differences of the results of the three measurement
methods are influenced by the frequency of the appearance
of MBs, and the typical appearance of IB. The significant effect
of the PoA might appear due to the release differences across
PoAs. The speakers’ effect also may arise from the interspeaker differences of the frequency of occurrence of
MBs. Although the results of the three VOT measurement methods
differed among the PoAs and the speakers, the main effects
(PoA, vowel context, speaker) had similar interaction in all
three methods with the data. The reason may be that the more
posterior the PoA the more frequently and more bursts appear
that does not influence the pressure tendencies of the PoAs i.e.
that the more anterior the PoA the lower intraoral pressure
build-up is during the closure, and the quicker the decrease can
be after the start of the release.

5. References

Barry, Jeff and Maura Moyle (2011). Covariation among vowel height
effects on acoustic measures. In: Journal of the Acoustical
Society of America, 130, pp. 365–371.

Boersma, Paul and David Weenink (2012). Praat: doing phonetics by
computer (Version 5.3.13). http://www.fon.hum.uva.nl/praat/
download_win.html

parameter for speaker identification: Evidence from Hong
Kong. In: Current Issues in the Unity and Diversity of
Languages: Collection of the papers selected from the CIL 18,
pp. 3921–3934.

Draxler, Christopher and Klaus Jansch (2004). SpeechRecorder, a
universal platform independent multi-channel audio recording
software. In: Proceedings of The International Conference

Fant, Gunnar (1969). Stops in CV-syllables. Quarterly Progress and

Fuchs, Suzanne (2005). Articulatory correlates of the voicing contrast
in alveolar obstruent production in German. In: ZAS papers in
cations_zaspil

Gráczi, Tekla Etelka and Anna Kohári (2012). Zöngek ezdési idő egy
metodológiai kérdés függvényében. In: Markó A. (ed. )
Tudomány. Az anyanyelv-elsajátítástól a zöngekézdés i idög, pp.
228–248.

Imbrie, Annika Karin Karlsson (2005). Acoustical study of the develop-
Harvard-MIT Division of Health Sciences and Technology,
Massachusetts Institute of Technology.

stop-consonant release for different places of articulation. In:

Song Jae Yung and Katherine Demuth (2008). Compensatory vowel
lengthening for omitted coda consonants: A phonetic
investigation of children’s early representations of prosodic

Consonant discrimination in elicited and spontaneous speech: a
case for signal-adaptive front ends in ASR. In: Proceedings
ICSLP, Beijing, China. pp. 325–328.

Wang, Yu-tsai, D. R. Kent, J. R. Duffy, J. E. Thomas and G.
Weismer (2004). Alternating motion rate as an index of speech
and Phonetics 18.1, pp. 57–84.
Does prosodic similarity matter in L2 production?
The case of Korean learners of French
Bénédicte Grandon, Hiyon Yoo
Université Paris Diderot, UMR 7110 Laboratoire de Linguistique Formelle, LaBex – EFL, France
b_grandon@yahoo.fr, yoo@linguist.univ-paris-diderot.fr

Abstract
French and Korean share similar prosodic characteristics as far as rhythm and intonation are concerned. In order to determine how this prosodic similarity affects the second language production, we propose in this paper to examine these prosodic parameters in read productions by Korean learners of French as a second language compared to French native speakers. We show that the productions of Korean learners and French native speakers present minor differences: concerning rhythm, Korean learners are less systematic in lengthening the last vowel at a phrase-final position while the overall pitch contour is similar for both groups of speakers, especially for the subject and object phrases. We argue that these minor differences are not sufficient enough for the detection of a “foreign accent” only with prosodic cues.

Keywords: Second Language, prosody, rhythm, intonation, French, Korean, Speech production

1. Introduction
In the last years, several studies have shown that the types of errors made by second language (L2) speakers depended on whether phonological categories were similar or different in both First (L1) and Second (L2) languages (see Best 1995, Flege 1995 among others for the segmental component). As far as the prosodic level is concerned, studies on intonation (Jilka 2007, Mennen 2007) and rhythm (Barry 2007) have shown that when the intonational contour or rhythm pattern exists in both languages, if the distinct features are minor, speakers are more prone to making mistakes. The present study focuses on the prosodic components of production in French by Korean native speakers. French and Korean are both described as “syllable-timed” languages (Di Cristo 1999, Jun 1993), with common prosodic features: (1) Primary stress, realized through syllabic lengthening, is located on the last syllable of the last lexical word of a phrase (among others Martin 2009, Di Cristo 1999 and Jun 1993), (2) non-stressed syllables have a constant duration (Di Cristo 1999, Lee 1993), and (3) declarative sentences have a descending pitch contour beginning on the first accented syllable in French (Di Cristo 1999) and on the sentence’s second syllable in Korean (Lee 1990) and continuing through the end of the sentence. Furthermore, the intonation of modality is seen as the result of F0 realization at IP-final level both in French (Delattre 1966, Martin 2009) and Korean (Jun 1993).
This quick comparison of the prosodic structure of the two languages reveals their prosodic similarity. In this paper, we present the results of a production experiment which aims at understanding how rhythm and intonation of native Korean speakers translate into French, and examining if there are enough cues in L2 realizations to consider the existence of a Korean foreign accent in French based on rhythm and intonation only.

2. Production experiment
In this section, we present the experimental protocol of our production experiment, followed by the results we obtained.

2.1. Method

2.1.1. Participants
All speakers were students living in Seoul at the time of the recordings: Two female native speakers of Standard Parisian French (19 and 20 years old) and three female (23, 23 and 26 years old) and one male (23 years old) native speakers of Standard Seoul Korean, with variable proficiency levels in French.

2.1.2. Corpus
Since our aim was to compare the prosodic characteristics of two groups of speakers, we chose to work on a read corpus, which allowed keeping the same segmental information for both groups while avoiding differences due to hesitations or lexicon and syntax.
Declarative sentences were designed in French following the pattern “NP subj-VP-NP obj”. All sentences are balanced: in each sentence, the three phrases (NP subj, VP and NP obj) have an identical number of syllables, varying from 1 to 10 syllables: there are two sentences of 3 times 1 syllables, two sentences of 3 times 2 syllables, two sentences of 3 times 3 syllables, etc. up to 3 times 10 syllables. Our corpus contains a total of 20 sentences. (1), (2) and (3) and sample examples of the corpus, in French, with English translation:

(1) 3x2 (6 syllables) sentence
(le chat)SUBJ (a bu)VP (le lait)OBJ
(the cat) subje (drank) verb (the milk) obj

(2) 3x4 (12 syllables) sentence
(Les deux garçons)SUBJ (ont rencontré)OBJ (un vieil ami)OBJ
(The two boys) subj (met) verb (an old friend) obj

(3) 3x9 (27 syllables) sentence
(Le voyageur perdu en Coré)SUBJ (a eu l’occasion d’utiliser) verb
(The lost traveler in Korea) subj (was able to use) verb

2.1.3. Recordings
Recordings were made with the Audacity software (Audacity Version 2.0), in mono, using a sampling frequency of 22050 Hertz and 32bits. All recordings were done on a laptop, using an external microphone, in a quiet room.
The twenty sentences of the corpus were read five times in a random order. The order was different for each series of sentences. All speakers read the five series of sentences in the same order, with pauses between each series. The recordings
were made without any time constraint. Speakers read the sentences before the recording session, and could ask for explanation if they had trouble understanding the meaning of the sentences. They were allowed to take breaks whenever they wished and to re-read sentences if they faced difficulties.

2.1.4. Data analyses

The corpus was segmented and annotated using Praat (Boersma & Weenink, 2013). The annotation was first made automatically using the Easyalign software (Goldman 2011) and verified manually. Segment durations and F0 values of vowels in phrase-final syllables were automatically extracted using a Praat script. For rhythm, we chose the vowel over the syllable as the unit of analysis. Vocalic durations were normalized in order to put aside “inter-speaker” and “intra-speaker” variations: instead of comparing raw durations, we used a ratio of the duration of each occurrence produced by the speaker divided by the mean duration of this type of vowel in this speaker’s corpus.

F0 values were measured at three points per vowel: at the beginning, middle and end. In order to avoid “inter-speaker” variation, the data were normalized in semi-tones calculated using each speaker’s mean F0, using the following formula (1) from (Martin 2009):

\[ F0(ST) = 12^*(\log(F0/speaker’s\, mean F0)/\log(2.00)) \] (1).

We ran ANOVA tests in order to compare phrase-final vocalic durations and phrase-final F0 modulations of the two groups (L1 and L2). We used regression tests to study the difference in realization of the declination line over sentences by the two groups. Statistics were run on the R software (R Development Core Team 2012).

3. Results

3.1. Rhythm

The study of rhythm consisted in measuring the presence or absence of vocalic lengthening in a given position. For the final vowel of each phrase (subject, verb and object), we set that a vowel can be considered as lengthened when its normalized duration is above the threshold of 1.2 (mean + 20%). We considered that choosing the mean duration value was not sufficient enough to determine a lengthening compared to the threshold of 1.2, above which lengthening can clearly be perceived. ANOVA tests were conducted for every sentence, but because of limits of space, we show the results for three sentences only, illustrating our purpose. Figure 1 shows the variation of mean vocalic durations for the two groups of speakers (French L1 and Korean L2) for three sentences (3x2=6, 3x4=12 and 3x9=27 syllables). In most cases, both French L1 speakers and Korean L2 learners lengthen the last vowel of the subject phrase. French speakers almost never lengthen the end of a verb, which shows that they tend to group the verb with its object and to place lengthening only at the end of the sentence. Korean speakers present more diverse results, with vocalic lengthening found in six cases out of ten (sentences with 3x4, 3x5, 3x6, 3x7, 3x8, 3x10 syllables), which might correspond to a more frequent segmentation of the sentence for learners than for native speakers. Vowels at the end of object phrases (which represents also the end of sentences) are systematically longer for French speakers while Korean learners do not produce this expected lengthening (vocalic lengthening of 1,2 can be seen only for sentences with 3x2, 3x4 and 3x10 syllables).

We ran ANOVA tests to compare the realizations of vowels at the end of phrases for the two groups of speakers. Thus, for 3x2 syllable sentences, durations are similar for both groups of speakers, with a lengthening of the 2nd vowel (Subject-final syllable), of the 4th vowel (verb-final syllable) and of the 6th vowel (Object and sentence-final syllable). The difference of lengthening is significant for both groups for each position (for the 2nd vowel, \( F(1,57) = 4,380 \) \( p=.0408 \), for the 4th syllable \( F(1,55) = 1,113 \) \( p=.0001 \) and the 6th vowel \( F(1,55) = 17,378 \) \( p=.0001 \)).

For 3x4 syllable sentences, the two groups have a different realization of lengthening: no lengthening of the 4th and 8th vowels (subject-final and verb-final syllables) but lengthening of the 12th vowel (sentence-final syllable) for the French speakers an moderate lengthening of the 4th, 8th and 12th vowels for the Korean speakers. The ANOVA tests show significant differences between the two groups for the three vowels: for the 4th vowel \( F(1,54) = 9,863 \) \( p=.0027 \), for the 8th vowel \( F(1,54) = 16,118 \) \( p=.0002 \) and for the 12th vowel \( F(1,54) = 14,353 \) \( p=.0004 \). However these results differ from the predicted results, since the group of Korean speakers and not the group of French speakers shows a systematic vocalic lengthening in phrase-final syllables for each phrase (subject, verb and object).

For the 3x9 syllable sentences, there is a lengthening of only the 9th vowel by French speakers and of the 9th and 11th vowels by the Korean speakers. The differences between the two groups for the 9th and 18th vowels are not significant, contrary to the difference between the two groups on the sentence-final vowel which is significant \( (F(1,48) = 18,292 \) \( p=.0001 \)).

![Figure 1: Mean vocalic duration of the two groups for 3x2, 3x4 and 3x9 syllable sentences.](image)

3.2. Intonational patterns at phrase-final level

For intonation, the F0 values for the last vowel of each phrase (subject, verb and object) were compared for the two groups of
speakers. Results show that the two groups (French native speakers and Korean learners of French) produce very close patterns. Figure 2 illustrates the differences and similarities of the F0 measures for the two groups of speakers (French L1 and Korean L2 speakers) for three sentences (3x2=6, 3x4=12 and 3x9=27 syllables).

Thus, at the end of sentences (i.e. at the end of object phrases), both French and Korean speakers produce massively a falling pattern (sentences 3x2, 3x3, 3x6, 3x8, 3x9, 3x10), which can be followed by a small rising (sentences 3x2, 3x6, 3x8). Korean learners have more random productions with more final risings, and even a rising pattern for sentence 3x7. However, the ANOVA test reveals a non-significant difference for F0 realizations of the two groups of speakers on the last vowel of the sentence.

The F0 realization on the last vowel of the subject is also similar for both groups. However, the type of pattern can vary, with a fall-rise pattern for sentence 3x1, a rising pattern for 3x5, 3x6 and 3x10 syllable sentences, and a falling pattern for the 3x7 syllable sentence. The realizations of the two groups of speakers differ for sentences with 3x2, 3x3, 3x8 and 3x9 syllables.

Results are less homogeneous at the end of verb groups. Within the group of French speakers, the last vowel at this position is produced as a flat pattern in short sentences (3x1 and 3x2 syllables), a rising pattern in sentences with 3x3, 3x7 and 3x9 syllables, a rise-fall pattern in sentences 3x5 and 3x10 and a falling pattern in sentences 3x4, 3x6 and 3x8. Korean learners realize also the same patterns but not for the same sentences: rising pattern is found in sentences 3x3, 3x5 and 3x7, a flat pattern in sentences 3x1 and 3x8, a rise-fall pattern in sentences 3x2, 3x9 and 3x10, a falling pattern in sentence 3x4, and a falling pattern followed by a rise in sentences 3x6.

For 3x2 syllable sentences, the realizations of the two groups are very close, with significant differences at only two points (F(1,55) = 12,056 p=.0010) at subject-onset point and vowel-midpoint at verb-final level (F(1,57) = 19,818 p=.0001).

For 3x4 syllable sentences, the two groups have a different realization of F0 patterns: flat F0 declination for French L1 speakers and peak pattern for Korean learners on the subject final vowel with significant differences on F0 values on the subject final vowel (F(1,36) = 6,306 p=.0168) at vowel-onset, (F(1,54) = 7,740 p=.0074) at mid-point and (F(1,36) = 4,514 p=.0405) at end-point) and at end-point of the verb-final vowel (F(1,12) = 4,890 p=.0472).

For 3x9 syllable sentences, both groups have similar realization of F0 patterns at verb and object level, but different at subject level: French L1 speakers have a rising pattern and Korean learners have a rising-falling pattern, with a significant difference at end-point of the subject-final vowel (F(1,22) = 9,172 p=.0062). At verb-final level, the group of French speakers have a rising F0 pattern whereas the Korean learners have a rising-falling pattern, with a significant difference between both groups at mid-point (F(1,47) = 5,895 p=0.0191) and at end-point (F(1,29) = 5,582 p=0.0251). Both groups have a falling F0 pattern at object-final level, but the Korean learners have a stronger, earlier fall, with a significant difference between groups at mid-point (F(1,42) = 4,229 p=0.0460).

3.3. Declination line

The analyses of declination line (Table 2, Figure 3) show that all slopes are negative, which means that the global F0 progressively declines for all sentences. For all sentences, slopes of Korean L2 speakers of French are systematically lower than those of the French L1 speakers, which indicates that the declination is systematically stronger for French native speakers.

Table 1: Slopes of declination lines for the two groups of speakers.

<table>
<thead>
<tr>
<th>Number of syllables per phrase</th>
<th>slope of declination for French speakers</th>
<th>slope of declination for Korean learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.201</td>
<td>-0.722</td>
</tr>
<tr>
<td>2</td>
<td>-0.393</td>
<td>-0.377</td>
</tr>
<tr>
<td>3</td>
<td>-0.228</td>
<td>-0.187</td>
</tr>
<tr>
<td>4</td>
<td>-0.116</td>
<td>-0.110</td>
</tr>
<tr>
<td>5</td>
<td>-0.112</td>
<td>-0.080</td>
</tr>
<tr>
<td>6</td>
<td>-0.090</td>
<td>-0.090</td>
</tr>
<tr>
<td>7</td>
<td>-0.040</td>
<td>-0.048</td>
</tr>
<tr>
<td>8</td>
<td>-0.055</td>
<td>-0.032</td>
</tr>
<tr>
<td>9</td>
<td>-0.054</td>
<td>-0.029</td>
</tr>
<tr>
<td>10</td>
<td>-0.035</td>
<td>-0.025</td>
</tr>
</tbody>
</table>

Furthermore, for both groups, the longer the sentence is, the more the slope in absolute value decreases. This fact points out a diminution of the declination depending on the F0 target at the end of sentences, which is approximately constant. The Korean L2 speakers however seem to have a greater F0 amplitude around the declination line than the French speakers, which could be explained by differences in group size or a greater F0 modulation by Korean speakers. Further experimentation and more speakers could help understand these particular results.
Figure 3: Declination lines of the two groups for 3x2, 3x4 and 3x9 syllable sentences.

4. Conclusion

For both French and Korean speakers, results of the present study reveal a systematic and strong lengthening of the subject-final vowels and the object-final vowels (i.e. sentence-final vowels) but not of the verb-final vowels as far the rhythmic component is concerned. In L2 productions however, systematic lengthening occurs at almost every final position. This vowel lengthening can also be observed in other positions in longer sentences, which reveals a different segmentation strategy in L2 speakers.

The analyses of phrase-final contours (subject, verb and object) show that French and Korean speakers have a similar subject and object-final realization of F0. However for L2 speakers, modulation of F0 appears to have a more random shape and height than for L1 speakers. Comparative analyses of declination lines through regression tests show a progressive decrease of F0 for both groups, with L1 speakers’ declination being steeper than L2 speakers’ declination. These differences in the production of L1 and L2 speakers in French are consistent with Jilka (2007) but are not sufficiently pronounced to precisely identify which of the four types of errors they represent. The prosodic similarity of the two languages in contact seems to be a criterion to consider in determining the types of errors. This result is confirmed by a perception study we conducted (Grandon & Yoo 2014) with native speakers of French, since these minor differences were not sufficient to allow the listeners perceiving a foreign accent in productions of French by Korean speakers.

5. Acknowledgements

This work is supported by a public grant as part of the “Investissements d’Avenir”-Labex EFL program (reference: ANR-10-LABX-0083).

6. References


Goldman J.-Ph. 2011. “EasyAlign: an automatic phonetic alignment tool under Praat” in Proceedings of InterSpeech, September 2011, Firenze, Italy

Grandon B. & Yoo H. 2014 “Do Korean L2 learners have a “foreign accent” when they speak French? Production and perception experiments on rhythm and intonation”, in Proceedings of Speech Prosody 07, Dublin


The Characteristics of Sublexical Errors in Spontaneous Speech
Dorottya Gyarmathy, Mária Gösy

1Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary

gyarmathy.dorottya@nytud.mta.hu, gosy.maria@nytud.mta.hu

Abstract

In this study we analyzed the phonetic patterns of sublexical phonological speech errors that were found in spontaneous narratives of Hungarian. A total of 210 speech errors (replacements, insertions and deletions) and their probable sources were analyzed. For a large proportion of speech errors, target and error sound differed in just one feature. The erroneous vertical movement of the tongue in vowels was responsible for the majority of speech errors as opposed to horizontal movement of the tongue. Phonetic distance in consonants appeared primarily in manner of articulation, followed by place of articulation and in a very low number of cases in voicing. Speakers showed different durations of editing phases depending on the type and the source of the error.

Keywords: occurrences, phonological features, editing phases

1. Introduction

Errors occurring in spontaneous speech have been analyzed in a number of perspectives. They have been used for explaining individual processes within speech production, for demonstrating the mental representation of words, for characterizing cognitive operations, for exploring the structural organization of language, and so on (Garrett 1980, Fromkin 1980, Shattuck-Hufnagel and Klatt 1980, Cutler 1982, Baars 1986, Levelt 1989, Dell 1990, etc.). The mechanism whereby the processes of speech production are modeled is often based on what are called ‘naturalistic speech errors’ (Perez et al. 2007). These can be classified into contextual vs. noncontextual errors. The former term refers to errors for which there is some motivation in the context, while the latter refers to the lack of such motivation (Pouplier 2007). A given error can be traced back to a planning deficiency at the phonological or the phonetic level, even though it is not always easy to tell these two sources of errors apart (Cutler and Henton 2004). Sublexical errors come about during further operations performed on already activated lemmata. Errors in which a given segment is substituted for by another are indicative of a difficulty in phonological planning, while incorrect implementation is a property of phonetic errors (Pouplier 2007). Such errors are sometimes indiscriminately referred to as “Ph errors”, that is, phonological/phonetic errors (Cutler and Henton 2004) or as phonemic errors (Postma 2000). Investigations into articulatory characteristics have confirmed that an intended and an erroneous articulatory gesture can occur either in sequence or in coproduction; that is, many errors are indeed based on incorrect articulation features (e.g. Goldstein et al. 2007, Pouplier and Goldstein 2010). The most popular speech production models (e.g., Dell 1986, Levelt 1989, Levelt et al. 1999, Howell 2007) assume that word form retrieval in speech production proceeds through independent phonological and phonetic processing modules. Sublexical errors are either thought to be independent of the phonological level (Dell et al. 1993, Levelt et al. 1999) or to be impossible to tell apart on the basis of a surface analysis (Cutler and Henton 2004, Goldstein et al. 2007). Phonological activation is never on its own during speech planning; it is invariably accompanied by semantic activation (Jescheniak et al. 2002). This seems to explain word activation errors based on both phonological and semantic similarity. On the other hand, phonological errors in which the activation of the mental lexicon is not directly involved (as they can be traced back to some contextual motivation) seem to contradict the above claim. Errors have been traditionally identified perceptually using transcription based data (Fromkin 1980, Perez et al. 2007) while acoustic, articulatory and perceptual measurements have shown different findings about the nature of errors (e.g., Mowrey and MacKay 1990, Pouplier and Goldstein 2002, Frisch and Wright 2002). Controlled, experimental data have been involved in a smaller extent, and recorded spontaneous speech has been used even less often. In the present paper, errors attested in spontaneous speech are analyzed in a variety of respects, first time in Hungarian as an agglutinating language. We restricted our attention to phenomena in which the error failed to result in an (unintended but) meaningful word. We tried to find out what phonological features were involved in Hungarian sublexical errors, and what information was provided by the duration of editing phases in repairing errors.

According to Dell et al. (1997), errors can be analyzed along two different dimensions, one being the size of the linguistic unit involved in the error, the other being the nature of the error. Our analyses were aimed at the phoneme realizations and the assumed sources for the erroneous segments that came into being. We defined three types of sublexical errors: replacements, insertions, and deletions. An example of replacement is this: az erkölc s az embereknek a becü becsületetessége ‘morale is the one- honesty of people’; an example of insertion: an enea- engineer’; and an example of deletion is bizonyos métekig mérettig ‘to some extent- extent’. As far as the sources of errors are concerned, we have discriminated four contextual possibilities, all of them involving a serial misordering of segments: anticipations, perseverations, anticipations-cum-perseverations (a mixed error type where both types of context effects may be involved simultaneously), and metatheses (i.e. an exchange of segments, see e.g., Nooteboom 2005) while the remaining errors were identified as noncontextual ones.

We formulated two initial hypotheses: (i) The phonetic distance (see Nooteboom 2005) between uttered and intended sounds is determined by the source of the error, and (ii) the duration of editing phases is characteristic of both the type of error and the probable triggering factor.

2. Subjects, method, material

We have analyzed sublexical errors in the material of 12 speakers selected from the BEA Hungarian Spoken Language Database (Gösy 2012). In the approximately 8 hours of material, we found a total of 140 occurrences, that is, errors occurred once in 3.43 minutes (410 words) on average. Our subjects committed 11.43 errors each (on average). The errors
were manually annotated, and then both authors categorized them individually. The two authors’ decisions agreed on the relevant categories in 98% of all cases (a third phonetician helped us make the final decision in the remaining few instances).

The corpus included 108 replacements, 18 insertions, and 14 deletions. We used four criteria in our analyses: (1) characteristics of the surface form of the error; (2) the presumed motivation triggering the error; (3) phonetic properties of the speech sounds involved and the phonetic distance between them; and (4) durations of the editing phases. We expressed phonetic distance in terms of phonological features of the segments. We defined editing phases as the length of time elapsing between the last sound of the erroneous sequence and the first sound of the corrected sequence (Levitt 1989). Measurements were made with the Praat software (Boersma and Weenink 2010); statistical analysis (one-way ANOVA) was made with the SPSS software.

3. Results

Most sublexical errors took the form of the speaker uttering another speech sound (both consonants and vowels) instead of the one intended (77.1%), that is, most errors were replacements. The share of deletions and insertions was far smaller. The former constituted 12.8% of the corpus, and the latter constituted 10%. The share of replacements was similar to that of the data reported for other languages (e.g., Schwartz et al. 1994). 43.6% of the errors occurred in initial syllables, 30.7% word medially, and 25.7% in word final syllables ($X^2$: (140, 2) = 7.129, p = 0.028).

In the case of replacements, speakers uttered the whole sequence in 32.4% of the cases, that is, they did not stop articulation at the point of error. In deletions, it was in 50%, and in insertions, in 88.9% of all instances that the whole (erroneous) sequence was uttered.

![Figure 1](image1.png)

Figure 1: The occurrence of sublexical error types in terms of the sources.

Overall, most sublexical errors (37.86%) were due to anticipation, 20.71% to perseveration, and in 13.57% of the cases both perseveration and anticipation may have triggered the phenomenon. 22.86% of the errors were non-contextual errors, and it was in a mere 5% of cases that metathesis was found to be responsible for the error. As far as the three error types (replacement, insertion, and deletion) were concerned, the percentages are summarized in Figure 1.

3.1. Patterns of speech errors

The question arises what types of segments were more often involved in the phenomena under investigation here. In our material, more errors concerned consonants (59.29%) than vowels (40.71%). This proportion was roughly similar both for replacements (proportion of consonants: 56.48%) and for insertions/intrusions (proportion of consonants: 61.11%), whereas deleted segments were even more dominantly consonants (78.57%). Both insertions and deletions concerned rather small numbers of cases (18 insertions and 14 deletions); thus, further analyses were worth performing for replacements only.

We wanted to find out whether phonological features can predict that some speech sounds are replaced more often while other less often. Of the consonants, those produced in the central area of the oral cavity (mainly alveolar ones) were replaced by other consonants to the largest extent, and the consonants used as substitutes were also typically of that place of articulation (e.g. [l], [n], [r]). The central place of articulation of the replacements can be explained by the fact that, due to the larger space available, it is easier to produce the necessary articulatory configuration in that area, irrespective of the manner of articulation. This is apparently contradicted by the fact that the replaced consonants were also mostly central. In our material, central consonants were invariably replaced by another central consonant. In the case of vowels, replacements were mainly lower in tongue height than the intended vowels (e.g. *utonk* for *utunk* ‘our way’). Again, it is probably due to the larger available space that it is easier to produce lower vowels. The phonetic distance between the erroneously realized sound and the intended one was a single feature in most of the cases (60.75%), irrespective of whether consonants (59.32%) or vowels (62.50%) were concerned. 25.23% of all segment pairs involved in replacements differed in two features, and 14.02% differed in three. In the latter two cases, however, vowels and consonants behaved differently. 32.20% of produced vs. intended consonants differed in two features, and a mere 8.47% of them differed in three features. In the case of vowels, the two percentages were closer to one another: 16.67% differed in two and 20.83% differed in three features from the intended vowels. The single feature difference between pronounced and intended consonants was typically that of manner of articulation (in 65.71% of the cases). In cases where the phonetic distance was several features wide, mainly replacements in terms of place and manner of articulation occurred (58.33%).

In errors due to anticipation and metathesis, it was typically the manner of articulation that was affected. In errors caused by perseveration, place of articulation and manner of articulation were equally involved. It was only in the case of perseverations and anticipations that change of voicing was found. Errors simultaneously due to anticipation and perseveration exclusively concerned manner of articulation. Among non-contextual errors, twice as many errors involved place of articulation than manner of articulation (Figure 2).

![Figure 2](image2.png)

Figure 2: Single-feature phonetic distance in consonantal errors (replacements only).

In cases where the phonetic distance involved several features, perseverative replacements involved place and manner of articulation or place and manner plus voicing. In the other categories, most occurrences involved place and manner of articulation (Figure 3).

167
The majority of erroneously produced vowels also differed from the intended quality in a single feature (62.5%). These errors typically involved the vertical position of the tongue. Errors concerning the horizontal tongue position were less numerous (10% only). In terms of the sources, tongue height errors did not differ from one another (Figure 4).

In the case of multiple-feature vocalic errors (37.5% of all vocalic errors), more than half of the instances involved three features (horizontal and vertical tongue position and rounding). In anticipations, horizontal and vertical tongue position or those two plus rounding were the most heavily involved. Rounding and horizontal tongue position are hardly ever involved jointly. Errors simultaneously due to anticipation and perseveration, as well as simple slips of the tongue, invariably involved all three features. In perseverations, too, this happened in the majority of cases. On the other hand, errors due to metathesis involved horizontal tongue position and lip rounding only (Figure 5).

In over 40% of all cases, the speakers committed errors in word initial syllables; one-third of the errors were word medial; and a fourth of them occurred in word final syllables. The relevant literature shows that word initial errors are around 80% of all attested errors in other languages (Nooteboom 1967, Cutler and Henton 2004). The discrepancy may be primarily due to structural differences across languages. In Hungarian, the fact that words tend to be multisyllabic, the rich morphological arsenal, the numerous suffixes obviously require more complex planning and implementation, hence the occurrence of errors in all possible positions within the word is expected.

The proportions of vowels and consonants involved in errors may be universal, as similar figures were found in a number of other languages (Cutler and Henton 2004). The phonetic distance between intended and pronounced segments usually concerned a single phonological feature in the case of replacements demonstrating a more economical process than that involving more features. Consonants produced in the central region of the oral cavity were mispronounced to a larger extent, and as substitutes it was also central consonants that was a vowel or a consonant that was repaired and, in some types, according to the source of the error, too. The mean duration of editing phases in consonantal error corrections was 135.6 ms (SD: 150.82 ms), and in vocalic error repairs it was 164.5 ms (SD: 162.14 ms). In terms of error types, speakers took the longest time, for repairing deletions; repairs of replacements and insertions took far less time (Figure 6, left panel). In terms of the sources of errors, the longest editing phases were required for anticipations. Somewhat shorter time was needed for repairing noncontextual errors; the fastest repairs were those of perseverations (Figure 6, right panel). However, these differences were not statistically significant in any of the cases.

### 4. Discussion and conclusion

The aim of this study was to describe the occurrence and phonetic properties of sublexical errors attested in a recorded corpus of Hungarian speech material, with characterization of the types of errors and of the assumed sources behind them. An advantage of this series of investigations was that it used real spontaneous speech material rather than a mass of transcribed data, thus excluding the participation of the listener’s (transcriber’s) speech perception mechanism, a potentially biased filter (see Perez et al. 2007, Pouplier and Goldstein 2002, Pouplier 2007). Most attested sublexical errors were replacements. The ratio of insertions and deletions was roughly equal; the two processes “counterbalanced” one another, as it were. The errors concerned consonants more often than they concerned vowels. Our analysis of the sources of errors showed that anticipations occurred the most frequently, perseverations and noncontextual errors occurred less (but nearly equally) often, even fewer errors were due to a combined effect of anticipatory and perseveratory forces, whereas metatheses occurred in vanishingly small numbers.

In over 40% of all cases, the speakers committed errors in word initial syllables; one-third of the errors were word medial; and a fourth of them occurred in word final syllables. The relevant literature shows that word initial errors are around 80% of all attested errors in other languages (Nooteboom 1967, Cutler and Henton 2004). The discrepancy may be primarily due to structural differences across languages. In Hungarian, the fact that words tend to be multisyllabic, the rich morphological arsenal, the numerous suffixes obviously require more complex planning and implementation, hence the occurrence of errors in all possible positions within the word is expected.

The proportions of vowels and consonants involved in errors may be universal, as similar figures were found in a number of other languages (Cutler and Henton 2004). The phonetic distance between intended and pronounced segments usually concerned a single phonological feature in the case of replacements demonstrating a more economical process than that involving more features. Consonants produced in the central region of the oral cavity were mispronounced to a larger extent, and as substitutes it was also central consonants that
occurred most of the time. This may be due to the fact that in this way the phonetic distance between the intended and the actually uttered sound is smaller. The differences between intended and pronounced consonants mainly concerned manner of articulation; they involved place of articulation to a smaller extent, and voicing was hardly affected at all. This contradicts the results based on transcribed error data (Cutler and Henton 2004). The explanation needs further articulatory analysis of the consonants that participate in these errors. Noncontextual errors follow the frequency hierarchy for errors since the place of articulation is less stable here than manner of articulation.

The involvement of tongue height in vocalic errors can be explained by the fact that the distance between neighboring height categories is relatively small and a minor error in realizing tongue height results in a categorically different vowel segment.

Our first hypothesis was confirmed: the phonetic distance is determined by the assumed source of error. More phonological features were involved in anticipations and perseverations than in methateses and in noncontextual errors. The latter sources due to their nature seem to restrict the spreading the erroneous features.

Sublexical errors were repaired in over 60% of the cases, suggesting that speakers find their repair essential in order to facilitate the listeners’ speech comprehension. The slightly shorter editing phases in repairing consonantal errors may be explained by the fact that they normally occur more often and hence self-monitoring is more sensitive to them. In the cases of deletions, even syllable structure may differ from what is intended, thus re-planning is an even more complex process requiring longer editing phases. Repairing insertions can be done faster since it is only the deletion of the extra sound that has to be performed. The repair of replacements requires the (revised) production of one or several articulatory gestures that needs, in general, more time. The monitor seems to be the most sensitive to cases of perseveration. The relative shortness of the editing phases of noncontextual errors may be related to the fact that in most relevant cases the speakers stopped speaking after one or two short syllables and so they only had to re-plan a relatively small part of the intended sequence. Since we can only find general tendencies with respect to the relationship of length of editing phases and type of error or source of error, we were unable to confirm our second hypothesis beyond reasonable doubt. The nature of anomalous words in spontaneous speech helps to understand the internal representations and mechanisms of the speech planning processes.

5. Acknowledgements

This research was supported by OTKA project 108762.

6. References


Characterizing Post-Glossectomy Speech Using Real-time MRI

Christina Hagedorn\textsuperscript{1}, Adam Lammert\textsuperscript{2}, Mary Bassily\textsuperscript{1}, Yihe Zu\textsuperscript{3}, Uttam Sinha\textsuperscript{3}, Louis Goldstein\textsuperscript{1}, Shrikanth S. Narayanan\textsuperscript{2}

\textsuperscript{1} Department of Linguistics, University of Southern California, USA
\textsuperscript{2} Viterbi School of Engineering, University of Southern California, USA
\textsuperscript{3} Department of Otolaryngology, Head and Neck Surgery, Keck School of Medicine, University of Southern California, USA

Abstract

We use real-time magnetic resonance imaging (rtMRI) as a tool to investigate post-glossectomy speech by examining articulatory behavior. Our data reveal that listeners perceive speech produced by postoperative partial-glossectomy patients whose surgical procedure most affected the base of tongue to be typical, while speech produced by patients whose procedure affected the oral tongue is perceived to be atypical. Mechanisms of both preservation and compensation are exhibited by post-glossectomy patients whose speech is deemed atypical by listeners. These patients employ the preservative behavior of maintaining durational differences in tense and lax vowels, as well as range of F1 (vowel height). Range of F2 (vowel backness), however, is severely reduced. Compensatory behavior is exhibited when coronal tongue movement has been impeded and is exemplified by (i) production of labial stops in place of target coronal stops and laterals and (ii) frication being produced by formation of a constriction between the tongue dorsum and palate, in place of alveolar fricative /s/.

Keywords: post glossectomy speech, oral tongue, compensatory, acoustic vowel space

1. Introduction

Real-time magnetic resonance imaging (rtMRI) is a particularly useful tool with which to study speech production, as it allows for movement of all vocal tract components to be observed over time. Other methods of articulometry, such as electro-magnetic articulography (Perkell et al. 1992) and X-ray microbeam (Westbury et al. 1994) allow for high temporal and spatial resolution, yet they provide information about only specific flesh points, generally only in the oral cavity, and cannot be used to observe the coordination patterns of all articulators within the vocal tract.

Patients with advanced tongue cancer oftentimes undergo surgical resection with/without reconstruction and radiation/chemoradiation therapy. Each treatment or the combination of treatments may result in short and long-term (often permanent) morbidity because of disfigurement, highly viscous saliva, trismus, dysphagia, and voice or resonance problems (Clark and Frei 1989). Many studies have investigated speech articulation following the partial-glossectomy procedure using videofluoroscopy (Georgian et al. 1982) and electropalatography (EPG) (Fletcher 1988; Imai and Michi 1992; Michi et al. 1989). The findings of these and other studies suggest that articulation is least affected in patients who have undergone resection of the base of tongue (Logemann et al. 1993), that stop consonant articulation is most distorted for postoperative glossectomy patients and that

mobility, rather than volume, of the residual tongue is most critical in maintaining speech intelligibility. rtMRI is an ideal tool with which to identify and further characterize these and other aspects of post-glossectomy speech as it is minimally invasive to the patient and provides a global, unobstructed view of articulator behavior in all parts of the vocal tract. Detecting and measuring movement of all vocal tract components is particularly critical when studying post-glossectomy and radiation therapy speech, because it has been observed that post cancer treatment patients sometimes form vocal tract constrictions compensatorily, with articulators other than those conventionally used by healthy subjects. Using rtMRI, an analytical method of estimating constriction kinematics based on pixel intensity and formant frequency analysis, we aim to temporally and spatially characterize various articulatory mechanisms that postoperative tongue cancer patients use in their attempts to form intelligible speech. Particularly, in this pilot study, our goal is to use rtMRI to (i) determine whether post-glossectomy patients preserve particular articulatory aspects of the vocal tract gestures they produce even after the procedure impedes their lingual mobility and (ii) illustrate the ways in which patients might compensate for their inability to form particular vocal tract constrictions.

2. Method

5 advanced tongue cancer patients (3 base of tongue, M1, M2 and M3, 1 oral tongue, M4, and one base of tongue and partial oral tongue, F1; 4 male, 1 female), ages 52 to 70, were imaged after having undergone partial glossectomy, neck dissection, free flap reconstruction and radiation therapy. MRI data were collected for subjects M1, M2, M4, and F1 more than 6 months post cancer treatment (after which point post-glossectomy speech intelligibility scores have been reported to reach a plateau (Imai et al. 1988)) and 4 months postoperatively for subject M3. None of the subjects received speech therapy between the time of post cancer treatment and the MRI scan.

The patients were imaged and had their speech recorded and subsequently denoised using a custom MRI protocol (Narayanan et al. 2004; Bresch et al. 2008), while producing read speech consisting of short phrases and single words, as they lay supine in the scanner. The subjects were prompted to read a series of short phrases and single words (presented visually by the experimenter) 2-3 times in random order. The stimuli included a subset of the phrases contained in “The Rainbow Passage” and the MOCHA-TIMIT corpus (Wrench and William 2000), as well as monosyllabic, labial stop-initial words containing the vowels /i, ɪ, ɛ, e, æ, ɔ, o, ɑ, ʌ, u, ʊ, ɝ/ as syllable nuclei.

chagedor@usc.edu
http://sail.usc.edu/span
2.1. Data acquisition
Image data were acquired on a 1.5T GE Sigma scanner, using a 13-interleaf spiral gradient echo pulse sequence (TR = 6.376 msec, FOV = 200 × 200 mm, flip angle = 15° (F1, M1, M2, M4) and 20° (M3)) and a custom 4-channel head and neck receiver coil. The scan plane (3 mm slice thickness) was located midsagittally; pixel density in the sagittal plane was 84 × 84 yielding a resolution of 2.38 × 2.38 mm. Image data were acquired at a rate of 18.52 frames per second, and reconstructed at 23.8 frames/sec. using a sliding window technique. Audio was recorded inside the scanner at 20 kHz simultaneously with the MRI acquisition, and subsequently denoised.

2.2. Articulatory and acoustic analyses
For all stimuli in the experimental corpus, audio and MRI video recordings, and MR image frame sequences of the subjects’ speech were examined. For all monosyllabic tokens, acoustic vowel duration was measured and formant frequency values at the acoustic midpoint of the vowel were extracted using Praat (Boersma and Weenink 2014). Additionally, jaw height was measured at the acoustic midpoint of the vowel by manual selection of the air-tissue boundary along a single vertical axis (defined independently for each subject) in a posterior region between the rear-chin and neck. This region was chosen, to maximize the amount of vertical elevating and lowering motion captured and to minimize the amount of rotation captured. For every token in which atypical articulatory behavior was visually observed, time series illustrating articulatory activity in regions of interest (labial, alveolar, velar; Figure 1) were automatically generated by calculating the mean intensity of pixels in each region. This method provides a robust estimate of constriction degree in noisy data, without relying on computationally intensive articulator segmentation along air-tissue boundaries (Lammert et al. 2010).

3. Results and analysis
3.1. Type of glossectomy predicts speech intelligibility
A brief listening task was completed by 5 native speakers of American English (3 male, 2 female) with no history of hearing or speech deficits, between the ages of 22 and 26. The task involved listening to select sentences read by each of the 5 subjects and judging the speech in each sample as “typical” or “atypical”. Results of this perception task reveal that all listeners perceived the speech of patients who underwent partial glossectomy and radiation therapy of the base of tongue (M1, M2 and M3) to be typical, while the speech of patients who underwent partial-glossectomy and radiation therapy for cancer of the oral tongue (M4 and F1) was perceived to be atypical.

3.2. Articulatory preservation
We observe that the speech of subjects F1 and M4 is perceived to be “atypical”, and that patients are likely aware that the acoustic patterns they produce do not match their acoustic targets. Nonetheless, they continue to produce systematic differences in their articulation and resulting acoustics that are available to them. Both subject M4 and subject F1 preserve expected differences in F1 across vowels. Further, both subjects maintain durational differences between tense and lax vowels.

3.2.1. Preservation of F1 values in postoperative speech
F1 and F2 values were averaged across repetitions of the same token for subjects M4 and F1. Upon comparing the acoustic vowel spaces of subjects M4 and F1 with those of healthy male and female individuals (as reported in Hillenbrand et al. 1995), respectively, both striking differences and peculiar similarities are observed (Fig. 2-3).

The vowel spaces of both subject M4 and subject F1 are visibly reduced when compared to those of their healthy counterparts. Interestingly, we find that differences in F1 values across vowel targets differing in height are generally preserved in postoperative speech, while the range of F2 values is severely reduced. Since F1 reflects vowel height while F2 reflects backness, the present findings seem to indicate that differences in tongue height, or constriction degree are maintained in postoperative speech, despite appropriate differences in tongue backness, or constriction location not being achieved. Furthermore, we observe that the F2 values of subject M4 are generally aligned with those of normal speakers for back vowels, but not for front vowels; subject M4’s vowel space is compressed rightward, rendering all ‘back’ vowels. This pattern is precisely what is expected given that subject M4’s glossectomy procedure caused damage to the anterior portion of the oral tongue that prevents lingual constrictions from moving forward. The F2 values of subject F1, on the other hand are compressed centrally, causing target front vowels to be produced with lower F2 values and target back vowels to be produced with higher F2 values than for normal speakers. This pattern, as well, is predicted given that subject F1’s glossectomy affected both the entire superior
portion of the oral tongue and the base of tongue, hence impeding horizontal movement in either direction.

3.2.2 Vowel duration varies as a function of tense/lax
Subjects M4 and F1 exhibit acoustic vowel length differences between tense and lax vowels (/i, u/ and /ɪ, ʊ/). Acoustic vowel length, measured from formant onset to offset, was longer for tense vowels than lax vowels (paired samples t-test, p<.05, p<.05). For subject M4, tense vowel length (avg. 374.3 ms.) was on average 87.4 ms longer than lax vowel length (avg. 286.9 ms.). For subject F1, tense vowel length (avg. 331.2 ms.) was on average 60.2 ms. longer than lax vowel length (avg. 271 ms.).

3.3. Compensatory mechanisms
Using videofluoroscopy, it has been observed that post-glossectomy patients sometimes use compensatory mechanisms to form intelligible speech whereby articulators other than the ones typically used to make certain constrictions in the vocal tract are used (Georgian et al. 1982). rtMRI reveals that subject M4 employs two types of compensatory mechanisms in attempt to produce (i) target alveolar stop constrictions and (ii) target alveolar fricative constrictions.

3.3.1 Compensatory production of stops and laterals
Subject M4, who is unable to execute finely controlled movements of the tongue tip, exhibits compensatory behavior by replacing the tongue tip gesture required for stops and laterals with labiodental stop constrictions sometimes accompanied by a dorsal constriction gesture. Subject M4 typically produces the target coda /t/ in isolated words by forming both a dorsal constriction and a labiodental constriction (as evidenced by triangle-shaped lower lip deformation caused by compression of the upper teeth into the lower lip, outlined in Figure 4). The labiodental constriction in place of /n/ can be compared to the bilabial constriction in onset /b/, during which extensive outward deformation of the lower lip is apparent (Fig. 4).

Figure 4, Top: Acoustic waveform and time-aligned estimated constriction functions (labial, alveolar, velar) in M4’s production of ‘sunlight’. Bottom: MRI frames displaying articulatory postures for onset /b/ (1), and labiodental and dorsal constrictions in place of coda /t/ (0).

Compensatory behavior of this kind is not limited to isolated tokens, but is exhibited in running speech as well. Subject M4 forms a labiodental nasal stop (evidenced by slight inward lower lip deformation, circled in Figure 5) in place of word-final target /n/ of “division” (Fig. 5)

Figure 5: Acoustic waveform and time-aligned estimated constriction functions (labial, alveolar, velar) in production of ‘division’ in running speech. MRI frame shows labiodental gesture in place of word-final /n/.

In running speech, subject M4 produces a labiodental stop in place of the /nl/ portion of “sunlight” (Fig. 6). The intensity patterns observable in the MRI frames corresponding with the /nl/ constriction duration suggest that the outer edge of the lower lip is compressed against the upper teeth (evidenced by complete inward deformation of the lower lip).

Figure 6: Acoustic waveform and time-aligned estimated constriction functions (labial, alveolar, velar) in M4’s production of ‘sunlight’ in running speech. MRI frame shows articulatory posture of labiodental stop (inward deformation of lower lip) in place of /nl/.

3.3.2 Compensatory production of frication
Compensatory behavior is not limited to stops, but also occurs in place of alveolar frication. Subject M4 produces frication between the palate and tongue dorsum rather than between the apical tongue and teeth to achieve target /s/ in ‘sun’. By using an algorithm that automatically detects the pixel of maximum dynamic intensity during the utterance of interest, we are able to robustly determine constriction location (Proctor et al. 2011). Using this method, we confirm striking differences in
4. Discussion

A major contribution of this work is to illustrate that there are important aspects of post-glossectomy speech that cannot easily be detected using traditional diagnostic methods relying solely on acoustic data or by using invasive imaging techniques. A brief listening task revealed that, as consistent with previous findings (Logemann et al., 1993), naïve listeners perceive speech produced by tongue base cancer patients as typical, while speech produced by oral tongue cancer patients is perceived as atypical.

The rtMRI and acoustic data suggest that the patients whose speech was perceived as atypical preserve select mechanisms used by healthy speakers to distinguish tense and lax vowels and vowels differing in height; namely, vowel duration and F1 modulation. While F1 (vowel height) modulation is maintained, difficulty modulating F2 (vowel backness), reflecting the damage caused to the oral tongue of subjects M4 and F1, is exhibited. Jaw height data collected in this pilot study reveal that subjects M4 and F1 exhibited jaw height differences between tense and lax vowels (/i/, /u/ and /ɪ, /ʊ/), with jaw height being higher in tense vowels (paired samples t-test, p<0.05, p<0.05). While these systematic differences likely contribute to F1 modulation, follow-up analyses on this data must be done to determine to what extent tongue movement, with respect to the jaw, contributes to the F1 modulation observed. Nonetheless, the presence of these preservative mechanisms where fine control of the tongue body is absent serves as evidence in support of an articulatory framework within which gestural constriction degree and gestural constriction location are controlled independently. Thus, when mobility of a particular constrictor is compromised (in this case, influencing constriction location), the remaining gestural specifications (e.g. constriction degree, activation duration) of the articulatory target remain unchanged. In the future, we aim to collect jaw height data both before and after the glossectomy procedure, to help determine whether differences in jaw height between vowels differing in height ought to be considered strictly a preservation mechanism or a compensation mechanism.

The rtMRI data and automatically generated pixel intensity functions show that subject M4 produces compensatory labiodental stop gestures (sometimes accompanying a dorsal stop gesture) in place of coronal stop gestures. Subject M4 also exhibits compensatory behavior with alveolar fricatives, for which he substitutes velar fricatives. The difference in constriction location between subject M4’s production of target /s/ and that of subject M1 is confirmed using an automatic method of identifying the pixel of maximum dynamic intensity over the segment of interest.

This study demonstrates that rtMRI is capable of capturing both aspects of typical speech that are preserved in post-glossectomy speech in addition to postoperative patients’ deviations from expected normal speech articulation. It is hoped that ultimately, these tools and the information that they provide will be used to aid in tailoring therapy programs that will effectively provide patients with the instruction necessary to help them to once again produce intelligible speech.

5. Acknowledgements

Research supported by NIH Grants R01 DC007124-01, DC008780-05 and DC009975-01.

6. References


Subject-Specific Biomechanical Modelling of the Tongue: Analysis of Muscle Activations During Speech

Negar M. Harandi 1, Jonghye Woo 2, Maureen Stone 2, Rafeef Abougharbieh 1, Sidney Fels 1

1 Department of Electrical and Computer Engineering, University of British Columbia, Canada
2 Dental School, University of Maryland, USA
negarm@ece.ubc.ca

Abstract

Generic biomechanical models of the oral, pharyngeal and laryngeal structures have been adopted into the ArtiSynth simulation framework (www.artisynth.org). Forward-dynamics tracking of FE model of the tongue was previously addressed through solving the inverse problem (Stavness, Lloyd, and Fels 2012). The estimated biomechanics were evaluated using either the average motion reported in the literature or those of a different subject. We expand the existing generic platform to allow for subject-specific simulations, in order to (1) better evaluate the simulated biomechanics, (2) investigate the inter-subject variability and (3) provide additional insight into the speech production.

Keywords: speech, muscle activations, biomechanical modelling, subject-specific simulation

1. Introduction

The tongue is deformable and possess hydrostatic properties. Physics-based deformable models of the tongue have been proposed to study speech motor control (Fang et al. 2009; Badin et al. 2002) and speech complications resulting from tongue surgery (Buchaillard, Brix, et al. 2007). However, these models are generic and do not provide any individualized information. Moreover, the workflow to create them, including segmentation, volumetric meshing, and activation, is highly manual, tedious, non-trivial, and hence, not suitable for creating subject-specific models in clinical settings.

If the biomechanical model identifies muscle groups of the tongue, dynamics of speech can be simulated using forward dynamics. Forward dynamic simulation consists of manual trial-and-error tuning of the muscle activation signals to produce the desired kinematics. However, the level and patterns of muscle activations are difficult to measure during complex motor tasks such as speech. Simultaneous Electromyography (EMG) recording of multiple active muscles is limited due to the lack of suitable technology. In addition, the moist surface and the highly deformable body of the tongue prohibited excessive use of fine EMG electrodes on tongue surface (Yoshida et al. 1982). Furthermore, the relationship between EMG and muscle forces is complicated in dynamic movements.

In addition, the tongue contains more muscles than the dimensionality of its modelled motion space and therefore is redundant (Stavness, Lloyd, and Fels 2012). This is mostly due to the computational limitations of the simulation process for differentiating all the possible kinematic DOFs of the soft tissue. This phenomenon is called motor redundancy and has been confirmed for the 3D shape of the tongue in speech movements (Stavness, Lloyd, and Fels 2012).

As an alternative, muscle activations can be predicted from the kinematic and force measurements by solving an inverse problem. The result from the inverse can be input to a forward simulation system to provide the necessary feedback to the inverse optimization. The forward-dynamics tracking simulation is a popular inverse modelling technique which has been widely used for musculoskeletal systems. Stavness, Lloyd, and Fels (2012) have extended the forward-dynamics tracking approach to include the dynamic FE models. These models are suitable for simulation of the soft-tissues, such as the tongue, which are activated without the mechanical support of a rigid skeletal structure.

In this paper, we generate the subject-specific finite element (FE) volumetric model of the tongue coupled with the related biomechanics available in the form of the generic model. The elements’ regularity and quality are preserved and the muscle groups and their attachments are identified to enable the simulation in ArtiSynth. We further simulate the subject-specific tongue dynamics during the speech-task ageese by applying the inverse solver to the tissue point trajectories extracted from tagged-MRI. The details of the methods are described in the following section.

2. Materials and Methods

Figure 1 shows the workflow we use. The subject-specific finite element model of the tongue is created after registering the generic model, available in the ArtiSynth, to the subject’s anatomy. Further, tissue points of the tongue are tracked during a speech task using tagged-MRI. Trajectories of these points are used as the target for the inverse tracking-based controller in the ArtiSynth (Stavness, Lloyd, and Fels 2012). The controller solves a local optimization problem at each time step of a forward-dynamics simulation in order to find a set of muscle activations that drive a biomechanical model through a target movement trajectory.

2.1. Model Creation

We use the generic FE tongue model available in ArtiSynth and fully described by Buchaillard, Perrier, and Payan (2009), as the source model for FE registration. The model provides 2493 DOFs and consists of 21 muscle groups including Genioglossus...
2.2. Tagged-MRI Tracking

In this study, a 26-year-old white female repeatedly speaks an utterance *ageese*. Dynamic tagged and cine MR image sequences are acquired, using consistent protocol, during the speech task. The images are acquired at multiple parallel axial slice locations covering the tongue. The resolution scheme is 1.875 mm in-plane (dense) and 6.00 mm through-plane (sparse). Super resolution MRI volumes are further reconstructed with an isotropic resolution of 1.875 mm, for each of the 26 time frames (Xing et al. 2013).

In order to track the tongue’s tissue points during the speech task, we combine the estimated motion from tagged-MRI, calculated by the harmonic phase algorithm (HARP), with surface information from cine-MRI using diffeomorphic demons. A 3D dense and incompressible deformation field from tagged-MRI is reconstructed based on divergence-free vector splines with incomplete data samples. It was observed that the tracking results are accurate at internal tissue of the tongue, but are noisy near the boundary. Interestingly, the opposite situation occurs for tracking results using 3D deformation registration methods. Thus the tracking was enhanced by adding boundary tissue tracking information derived from 3D deformable registration to the internal tissue tracking information derived previously. The details of the method is explained in Xing et al. 2013.

2.3. Inverse Simulation

The solver considers forward-dynamics simulation of the constrained mechanical system to compute the system velocities, $\mathbf{u}$, in response to muscle-activation-dependent forces. The system can be linearised if the force is considered to be linear w.r.t. set of muscle activations, $\mathbf{a}$, subject to the condition $0 < \alpha < 1$. Let $\mathbf{v}$ be the target velocities defined in a sub-space of the total system velocities. The target velocity sub-space $\mathbf{v}$ is related to the system velocities $\mathbf{u}$ via a Jacobian matrix $J_m$ so that $\mathbf{v} = J_m \mathbf{u}$. The normalized activations are calculated by solving a quadratic program:

$$\mathbf{a} = \arg \min ( \| \mathbf{v} - H \mathbf{a} \|^2 + \alpha \| \mathbf{a} \|^2 + \beta \| \hat{\mathbf{a}} \|^2 )$$

$\alpha$ and $\beta$ are the $\ell^2$-regularization and damping coefficients. We refer to Stavness, Lloyd, and Fels 2012 for more details on the system.

3. Experiments and Results

We first evaluated our subject-specific tongue model by activating the muscle groups one at a time. The observed motion was...
in the range of reported movements in the literature (Stavness, Lloyd, and Fels 2012). The recorded tagged-MRI experienced some noise in the left half of the subject’s tongue; hence we selected our target points to be in the right half of the tongue, while forcing identical muscle activations for both left and right. We used 40 uniformly distributed nodes of the FE model as targets for the inverse simulation.

In order to test the reliability and accuracy of the inverse solver, we designed an experiment in which the muscles were first manually activated one after another, to provide a set of synthetic muscle activations. The displacements of the target nodes were recorded over time and were fed back as input to the inverse solver. The results show high accuracy (Figure 2).

In the final experiment with the tagged-MRI, we calculated the average displacement in the neighbouring region of each target point, in order to reduce the noise. The estimated activations are shown in Figure 3. The average of maximum tracking error was 2.88 ± 1.2 mm.

4. Discussion and Future Work

4.1. Muscle Activations

The estimated muscle activations show a plausible explanation for the known features of speech such as co-articulation. Co-articulation means that sequential speech sounds are not actually concatenated, but are produced with a temporal overlap. The model correctly predicts co-articulatory relationships during the motion from /g/ to /i/, as displayed in Figure 3 for the activation patterns of the mylohyoid (orange) and the genioglossus posterior (red). Mylohyoid activation acts to elevate the whole tongue as a single unit. The inverse model estimates the mylohyoid as primarily responsible for the rapid elevation of the tongue body from the low position of the maxilla and jaw; Stylohyoid (SH) and Posterior Digastric (PD) between the cranium and jaw; Superficial Masseter (SM), Deep Masseter (DM), Medial Pterygoid (MP), Superior Lateral Pterygoid (SP), and Inferior Lateral Pterygoid (IP) between the maxilla and jaw; Stylohyoid (SH) and Posterior Digastric (PD) between the maxilla and hyoid; The insertion and origin points of each muscle should also be tailored to match the anatomy of our subjects. In future we will provide a more detailed motion analysis of the articulatory model, including activation of the muscles of the jaw and hyoid.

4.2. Muscle Configuration

The magnitude of the tracking error depends on various parameters related to the characteristics of the FE model, noise in the trajectories and the inverse simulation. We noticed that the maximum error happens before /g/ when the muscles of the tongue need to push forward and upward. The forward simulation of the model results in similar error. Muscle bundles of Genioglossus are shown in the figure 4. We argue that the aforementioned forward-upward motion is generated from the contraction of part of the posterior muscle bundle at the very back of the tongue. However, current arrangement of the muscles will expand this motion to the middle part of the tongue, causing additional tracking error. In future, we are planning to subdivide the GGP muscle into smaller action units to compensate for the error.

4.3. Jaw and Hyoid Model

In this paper, we assume that the jaw and hyoid have minimal displacement and, hence, their effect on the tongue is modelled as fixed attachments. The generic coupled model of tongue-jaw-hyoid has been implemented in ArtiSynth (Stavness, Lloyd, and Fels 2012); the jaw and hyoid were modelled as rigid bodies with their own set of muscles and constrained motion. In addition, the mass of the jaw and hyoid introduces inertia and damping effects to the system. Bone structures are partially visible in MRI. This makes manual segmentation of MRI inadequate for creation of the model. On the other hand, CT scanning for healthy subjects requires extra ethics due to the hazard of radiation, and hence, is opted out of. In future, we plan to register the available generic model of the jaw to the partial surface segmented from the Cine-MRI volume using a non-rigid point set registration technique, such as coherent point drift (CPD) algorithm (Myronenko and Song 2010). There are 11 pairs of point-to-point, hill-type muscles associated with the jaw and hyoid generic models. These include Anterior Digastric (AD), between the jaw and hyoid; Anterior Temporal (AT), Middle Temporal (MT) and Posterior Temporal (PT) between the cranium and jaw; Superficial Masseter (SM), Deep Masseter (DM), Medial Pterygoid (MP), Superior Lateral Pterygoid (SP), and Inferior Lateral Pterygoid (IP) between the maxilla and jaw; Stylohyoid (SH) and Posterior Digastric (PD) between the maxilla and hyoid; The insertion and origin points of each muscle should also be tailored to match the anatomy of our subjects. In future we will provide a more detailed motion analysis of the articulatory model, including activation of the muscles of the jaw and hyoid.

4.4. Muscle Material

We use linear muscle material for our FE model. This provides more stability for the simulation, however it does not fully account for hyper-elasticity of the tongue tissue. We believe
that using a non-linear hyper-elastic muscle material, such as Mooney-Rivlin (Mooney 2004), will yield more accurate results for this particular speech task, both with respect to the estimated activations and the tracking error. In future, we will adapt our FE model to use Mooney-Rivlin material by addressing the instability issue.

4.5. Inter-subject Variability
In our future work, we will expand our experiments to other normal subjects, in order to measure the inter-subject consistency and variability of the estimated muscle activations. We have access to similar tagged and cine MRI data for 20 normal subjects and 15 patients with partial glossectomies. We also plan to investigate the possible effects of head posture on the muscle activations during speech.

5. Conclusion
We, for the first time, adapt the generic FE model described by Stavness, Lloyd, and Fels (2012) to the subject domain to investigate the speech task ageois. The subject-specific FE Model of the tongue is created based on the anatomy of the subject, captured in volumes of dynamic cine-MRI. We simulate our model based on the motion of tissue points of the tongue, extracted from the tagged-MRI of the speech sequence. The preliminary results show a plausible explanation for the known features of speech, such as co-articulation. Future work will involve further adjustment to the model, in order to estimate more reliable muscle activations while minimizing the tracking error.

Although our current simulations focus on speech, we believe that our model is suitable for investigating other neuro-muscular functions of the tongue such as swallowing; Our collaborators in ArtiSynth are creating a complex model of the soft palate and pharyngeal wall. A separate project is developing an air/water tight tube that can attach to the FE model of the tongue, deforming accordingly. Our efforts in tailoring these generic models to match the anatomy of different subjects will provide additional insight into the nature of oropharyngeal functions, as well as the extent of reliability of the biomechanical simulations.

6. Acknowledgements
This work is supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC), NSERC Collaborative Health Research Project grant and Network Centre of Excellence on Graphics, Animation and New Media (GRAND).

7. References
The emergence of rhythmic strategies for clarifying speech: variation of syllable rate and pausing in adults, children and teenagers

Valerie Hazan¹, Michèle Pettinato²

¹²Speech Hearing and Phonetic Sciences, UCL, London UK

v.hazan@ucl.ac.uk, Michele.Pettinato@uantwerpen.be

Abstract

This study focuses on the development of two temporal aspects of speech production, articulation rate and pausing, in children aged 9 to 14 years. Both measures were examined in spontaneous speech produced in interaction with a friend of the same age range and gender while completing a collaborative picture task. To investigate the use of articulation rate and pausing as clear speech strategies, two conditions were used that differed in the ease of communication between the two talkers. Articulation rate was higher in adults than in child groups, in male than female speakers, and was used as a clear speech strategy by all age groups. Pause frequency was variable within groups, but there was greater evidence of an increase in pause frequency as a clear strategy in adults than child groups, with no gender effect found for this measure. Overall, this study provides evidence for ongoing developments in certain aspects of speech production until adulthood.

Keywords: speech production, speech development, connected discourse, articulation rate.

1. Introduction

There has been much attention on early stages of development in speech production, but while children are efficient communicators by the time they enter school there is also evidence that their production of speech sounds is more variable than adults until early adolescence at least (e.g., Lee et al., 1999). Given this lesser degree of control over speech production, one area of interest is the development of intrinsic articulation rate in spontaneous speech, as this is likely to reflect development in articulatory control (Lee et al., 1999). A review of studies of articulation rate, mostly involving American children, suggest a developmental trend in children with increases in articulation rate over the first ten years or so of life (Logan et al., 2011); Sturm and Seery (2007) found no difference in articulation rate for conversational speech between 9 and 11 year olds; the task involved children answering open questions by an adult researcher. Neither of these studies provides comparative articulation rate data for adults. Jacewicz et al. (2010) did include both children aged 8 to 13 and adults aged between 20 and 91 years, and one of the tasks collected conversational speech in response to open questions (e.g. speaking about family, friends, daily activities). The model fitted to the data suggested an increase in articulation with age approximately until the age of 45 years, suggesting that adolescents use slower articulation rates than young adults.

Evidence for gender effects on articulation is mixed. Some adult studies have found evidence of faster articulation rates in men than women (e.g. Jacewicz et al., 2010; Quené, 2008) but, to our knowledge, the effect of gender has not been investigated in depth in studies of articulation rate in children.

The frequency and duration of pausing is another aspect of the development of timing control in speech. Pausing is influenced by cognitive processing and planning, and developmental trends have been found in pause frequency and duration, with changes obtained for these measures between children in kindergarten (aged 5-6) and 7-8 year olds (Sabin et al., 1979).

In a recent study comparing 5-6 year olds to adults, using a story narration task to elicit spontaneous speech, Redford (2013) found that young children were less adept than adults at pausing in linguistically-appropriate locations.

Changes in articulation rate and pausing are two prominent strategies that are used when adults attempt to clarify their speech in adverse listening conditions (for a review of clear speech strategy, see Smiljanic and Bradlow, 2009). There has been much interest in the study of clear speaking styles in adult speakers, as these reflect adaptations that speakers make in order to maximize communication efficiency, either in adverse listening conditions, or when conversing with an interlocutor with reduced language abilities (e.g. L2 speaker or young child). Clear speaking styles involve enhancements of a wide range of acoustic-phonetic characteristics; in addition to changes in articulation rate and pausing, as mentioned above, changes can involve pitch characteristics, vowel space area and intensity. It is also known that there are significant individual differences in clear speech strategies (e.g. Ferguson, 2004; Maniwa et al., 2009) and that such strategies are, to an extent, adapted to the specific communication barriers they are seeking to overcome (Hazan and Baker, 2011). Being able to increase the clarity of one’s speech for the benefit of an ‘impaired’ interlocutor is therefore likely to be a skilled aspect of speech production but little is known about the range and extent of clear speech adaptations in children. A study involving 3 to 5 year olds showed that the older children in the study were making some adaptations to their speech but that these adaptations were not adult-like (Redford and Gildersleeve-Neumann, 2009). A more recent study involving 11 children of a similar age engaged in an object naming task involving interaction with a puppet did find evidence of clear speech characteristics: the clear speech condition was characterised by a larger vowel area together with changes to the duration of vowels and to their fundamental frequency (Syrett and Kawahara, in press). In that study, perception tests with adult listeners showed that the two speaking styles produced by the children were discriminable by adults. There is therefore evidence of early stages of clear speech adaptations in preschool children but with the potential for further developments.

The research objectives were as follows. First, we wished to establish whether children’s articulation rate and pause frequency had reached adult-like values within the 9 to 14 year old age range as suggested by the previous literature. Note that
our measurements are based on spontaneous speech produced with communicative intent during a problem-solving task with another speaker rather than from picture descriptions, answers to open question or story narrations. Second, we wished to examine the use of decrease in articulation rate and increase in pausing frequencies as ‘clear speech’ strategies when carrying out the task with the same interlocutor hearing in adverse conditions. Third, we wished to investigate whether there was evidence of a gender effect emerging within this age range.

Our initial results (Pettinato and Hazan, 2013) based on a group of 40 children (twenty 9-10 year olds and twenty 13-14 year olds) suggested that a reduction in mean word duration was a primary means of clarifying speech addressed to an ‘impaired’ interlocutor, but while adults increased the frequency of their pauses in the ‘communication barrier’ condition, younger children did not. Here, we have extended this work by significantly increasing the number of child participants in our study (96 children aged between 9 and 14 years) and by using a measure of articulation rate rather than mean word duration, to enable an easier comparison with previous studies.

2. Method

2.1. Speech recordings

Recordings were made of spontaneous speech produced in interaction with another talker of the same age and gender while completing a collaborative ‘spot the difference’ picture task using the diapixUK materials (Baker and Hazan, 2011). The two talkers, who knew each other, were seated in separate sound-treated rooms and could not see each other; they communicated via headsets. Articulation rate and pausing were examined in two conditions that differed in the ease of communication between the two talkers. In the ‘no barrier’ (NOB) condition, both talkers could hear each other normally and therefore the measures obtained reflect intrinsic articulation rates and pausing frequency in connected discourse. In the vocoder (VOC) condition, communication was degraded by passing the speech of one of the talkers through a three-channel noise-excited vocoder before it reached the other talker. This caused the unimpaired talker to clarify his or her speech in order to maintain efficient communication. This condition is used to investigate whether a talker changes articulation rate and pausing frequency as clarification strategies when communicating with an ‘impaired’ interlocutor. The speech was analysed for both the NOB and VOC conditions for the unimpaired talker, named ‘talker A’. The average duration of each recording was around 10 minutes yielding around 4 minutes of speech to be analysed per talker per condition. Analyses of task transaction time (time taken to find the first 8 differences in the pictures) did not show evidence of an age effect suggesting that the task itself was equally suitable for children and adults. Further details of recording procedure can be found in Hazan and Baker (2011) which reported the analysis of adult speech recording elicited using the same procedures.

2.2. Participants

Recordings were made for 96 child participants as ‘talker A’: thirty 9-10 year olds (14F, 16 M), twenty-four 11-12 year olds (16F, 8 M), forty-two 13-14 year olds (20F, 22 M). The child recordings were compared to those of 20 adults (9F, 11 M) from the published LUCID corpus (Baker and Hazan, 2011), recorded under the same conditions.

3. Results

The following measures were calculated for each talker A in each condition. Articulation rate was calculated as the number of syllables produced by talker A divided by the total duration of speech segments (excluding fillers, silences, etc) for that talker. Syllable counts were calculated from the orthographic transcriptions of the spontaneous speech using the qdap package in R (Rinker, 2013).

Pauses (SIL) had also been manually annotated in the Praat textgrids, using the criterion that a SIL was defined as a within-talker pause of at least 500 ms in duration. A normalised measure of pause rate was calculated as the number of pauses divided by the number of words (excluding fillers and interrupted words). Mean pause duration was also calculated.

Linear mixed effect models were used, with participant treated as a random effect and condition, age group and gender as fixed factors. Statistics were carried out in R using the lme4 package (Bates et al., 2014).

3.1. Articulation rate

Box plots of articulation rate as a function of age group and condition are shown in Figure 1, and a table also shows mean articulation rate as a function of gender and age group per condition (Table 1). There was a significant effect of condition [F(1,108)=333.56; p<0.0001]; articulation rate was slower in the VOC condition (M = 3.2) than NOB condition (M = 4.1). The effect of age group was also significant [F(3,108)=18.09; p<0.0001]. The age effect was examined in more detail in post-hoc analyses. Adults had a higher articulation rate than all child age groups, and the 13-14 year old group had a higher articulation rate than the 9-10 year old group. The age group by condition interaction was not significant, suggesting that all age groups used a reduction in speech rate as a clear speech strategy. The effect of gender was significant [F(1,108)=17.84; p<0.0001] due to a slower articulation rate in female (M=3.5) than male (M=3.9) talkers. A gender by condition interaction (p<0.04) seemed due to a greater gender effect in the VOC condition. The condition by age by gender interaction (p<0.006) was not clearly interpretable.

Table 1: Means and standard deviations for articulation rates (syllables/second) as a function of age group and sex.

<table>
<thead>
<tr>
<th>Age group</th>
<th>sex</th>
<th>NOB</th>
<th>S.D.</th>
<th>VOC</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-10 yr olds</td>
<td>F</td>
<td>3.8</td>
<td>0.4</td>
<td>2.6</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>3.7</td>
<td>0.4</td>
<td>3.1</td>
<td>0.5</td>
</tr>
<tr>
<td>11-12 yr olds</td>
<td>F</td>
<td>3.9</td>
<td>0.4</td>
<td>2.9</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>4.3</td>
<td>0.3</td>
<td>3.4</td>
<td>0.5</td>
</tr>
<tr>
<td>13-14 yr olds</td>
<td>F</td>
<td>4.0</td>
<td>0.4</td>
<td>3.1</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>4.2</td>
<td>0.6</td>
<td>3.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Adults</td>
<td>F</td>
<td>4.3</td>
<td>0.5</td>
<td>3.4</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>5.0</td>
<td>0.4</td>
<td>3.9</td>
<td>0.7</td>
</tr>
</tbody>
</table>

The scatterplot in Figure 2 shows individual talkers’ articulation rate in the NOB condition plotted against the VOC condition. This highlights the variation in intrinsic articulation rate (NOB) in all age groups. The fact that the great majority of data points are in the lower area of the scatterplot suggests that a majority of talkers did use a reduction in articulation rate as a strategy to clarify their speech for the benefit of their
interlocutor in the VOC condition. It is the case though that a number of talkers in all age groups showed little change in articulation rate across conditions (points near the diagonal).

Figure 1: Box-plot showing articulation rate (syllables per second) as a function of age group and condition. Adults had a significantly higher articulation rate than all child groups. All groups reduced their articulation rate in the VOC condition.

Figure 2: Scatterplot showing articulation rate (syllables per second) as a function of age group and condition. Adults are labelled as ‘0’.

3.2. Normalised pause frequency

Box plots of normalised pause frequency as a function of age group and condition are shown in Figure 3. For certain age groups at least, within-group variability was large suggesting that certain children used pausing as a strategy while others did not, both when communication was easy (NOB) and difficult (VOC). Given this individual variability, a more meaningful measure, in terms of investigating the effect of condition on pausing strategy, is one representing the relative change in normalised pause frequency across the NOB and VOC conditions (Figure 4). The effect of age group was significant \( F(3,110) =5.87; \ p<0.0001 \). Post-hoc analyses showed that this effect was due to adults showing a greater relative change in pause frequency than all of the child age groups. The effect of gender was not significant.

Table 2: Means and standard deviations for pause duration (seconds) as a function of age group

<table>
<thead>
<tr>
<th>Age group</th>
<th>NOB</th>
<th>S.D.</th>
<th>VOC</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-10 year olds</td>
<td>1.17</td>
<td>0.4</td>
<td>1.13</td>
<td>0.3</td>
</tr>
<tr>
<td>11-12 year olds</td>
<td>1.13</td>
<td>0.3</td>
<td>1.22</td>
<td>0.4</td>
</tr>
<tr>
<td>13-14 year olds</td>
<td>1.12</td>
<td>0.3</td>
<td>1.18</td>
<td>0.4</td>
</tr>
<tr>
<td>Adults</td>
<td>0.89</td>
<td>0.2</td>
<td>1.00</td>
<td>0.2</td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

In summary, intrinsic articulation rate as measured in a communicative task involving pairs of participants of the same gender and similar age had not reached adult-like levels by the age of 13-14 years. However, all child and adult groups did use a reduction of articulation rate as a strategy to clarify their speech when their interlocutor was experiencing an adverse listening condition. Although there was evidence of individual children and adolescents increasing pause...
frequency as a clear speech strategy, group effects showed that the adult group used this strategy to a significantly greater extent than all groups of children and adolescents. Gender effects were found for articulation rate, with girls and women generally using a slower speech rate than boys and men. No gender effects were obtained for pause frequency and duration though.

It should be noted that the intrinsic articulation rates (in the 'no barrier' condition) obtained in our study are rather higher than the 2.7-3.3 syllable per second range cited in the literature for this age range (Logan et al., 2011), although closer to the articulation range seen in Jacewicz et al. (2010); this could be due to the use in our study of a collaborative task carried out between friends rather than more formal picture narrations or open-question interactions with adults used in previous studies. There was some evidence of an ongoing developmental trend within this age range, with 13-14 year olds using a faster articulation rate on average than 9-10 year olds. There was also a further significant difference in articulation rate between young adolescents and adults; this concurs with the findings of Jacewicz et al. (2010) for conversational monologues as they observed a trend for increasing articulation rate until middle age. The finding of a gender effect, with slower articulation rate for female talkers, also concurs with previous findings (e.g. Jacewicz et al., 2010). In that study, sociophonetic factors were argued for this gender effect and so it is of interest to see that these gender effects are already apparent within this age range. Female talkers tend to be more listener-oriented than male talkers, and have been shown to generally be more intelligible than male talkers (e.g., Hazan and Markham, 2004). Less of a developmental trend within the 9 to 15 year old age range was in evidence for pause frequency, and adults tended to use less pauses overall. A further analysis of the location of pauses in relation to discourse structure is needed for a better understanding of these differences (Redford, 2013). The finding that children were using pauses of longer duration than adults could be related to the communicative task involving a greater cognitive load for younger talkers, possibly because of the lack of face-to-face interaction.

This study also examined clear speech strategies used by children within the temporal domain. Articulation rate reduction appeared to be a strategy used by all age groups as shown by the lack of an age by condition interaction. This comes as no surprise as this strategy is evident in the great majority of clear speech studies. It should be noted though that the kind of individual differences in clear speech strategies often seen for adults (e.g., Krause and Braida, 2004; Maniwa et al., 2009) is also present here, as a number of children and adults showed little or no reduction in articulation rate in the condition involving effortful communication between the two speakers. Increasing the frequency of pauses as a clear strategy was however much more variable, as shown by the sizeable within group variance. It was mostly in evidence in adult speakers, as a significant age effect was obtained, with adults differing from all groups of children, but in all groups, as shown in Figure 4, there was evidence of some individuals using this strategy. Overall, this study provides evidence for ongoing developments in certain aspects of speech production until adulthood.

5. Acknowledgements

This work was supported by Economic and Social Research Council Grant No. RES-062- 23-3106. Data can be accessed by contacting the corresponding author. We acknowledge Dr Outi Tounaimen’s contribution to the analysis of the data.

6. References


Acoustic Correlates of Height in the Production of Whispered Fricatives

Willemijn F. L. Heeren
Leiden University Centre for Linguistics, Leiden Institute for Brain and Cognition
Leiden University, The Netherlands
w.f.l.heeren@hum.leidenuniv.nl

Abstract
As part of an investigation into the coding of intonation in whispered speech, this study examined acoustic characteristics of the whispered fricatives /s/ and /f/ produced at different heights. In addition, the acoustic correlates were interpreted as secondary versus compensatory cues to pitch, through a comparison with fricatives produced in normal speech. In whisper, as in normal speech, fricative consonants /s/ and /f/ systematically changed with the intended height at which a speaker produced an utterance. Most of the correlates were secondary; only in the case of intensity development in time it may be argued that compensatory information was provided by whispering speakers.

Keywords: intonation, whispered speech, fricatives

1. Introduction
Prosody, and more specifically lexical tone and sentence melody (intonation), can – to an extent – be perceived in whispered speech (e.g. Fónagy 1969; Kong and Zeng 2006; Liu and Samuel 2004). Whisper, however, is produced without the quasi-periodic vibration of the vocal folds that generates the speaker’s fundamental frequency (F0), which is assumed to hold the main cue to intonation. In the few studies to date, acoustic correlates of pitch in whisper have been sought in vowel content. This work has found that the lower formants, F1 and F2, could carry prosodic information in the absence of F0 (Higashikawa and Minifie 1999; Heeren and Van Heuven 2009), and that other contributing cues seem to be intensity (Meyer-Eppl 1957; Denes 1959), higher formants (Meyer-Eppl 1957; Fónagy 1969), and duration (Liu and Samuel 2004). In recent years, studies of normal (i.e. phonated) speech have demonstrated acoustic correlates of intonation in voiceless consonants as well (Cho and McQueen 2005; Niebuhr 2008; 2012); moreover, listeners may be sensitive to such correlates (Kohler 2011; Niebuhr 2008). In the present study we examined how consonants, rather than vowels, may contribute to the coding of intonation in whispered speech.

The phonetic implementation of prosody affects the characteristics of the segments, both vowels and consonants, over which the prosody is produced. In normal, phonated speech the acoustic content of both stops and fricatives is influenced by prosodic factors. As for stop consonants, Niebuhr (2008) found that aspiration of utterance-final /s/ is systematically different between two accent contours in its duration, intensity maximum and in the location of the spectral peak above the lower spectral energy boundary. In domain-initial position, the stops /t/ and /d/ showed longer closure duration with stronger prosodic boundaries (Cho and McQueen 2005). Both stops also showed a difference in centre of gravity (CoG) with prosodic boundary depth, but the differences were not consistently found for the same boundary depths and were not in the same direction. In that same study, the fricatives /s/ and /z/ showed longer durations with stronger boundaries. Intensity was found to be lower with stronger boundaries, which was most clearly found for /s/, but CoG did not vary with prosodic boundary strength in domain-initial position. Recently, Niebuhr (2012) investigated the interaction of utterance-final changes in F0 and the realization of voiceless fricatives /θ/, /s/, /ʃ/ and /h/ in utterance-final position, placed directly after an F0 fall or F0 rise. In rising contexts, intended as questions, the voiceless fricatives in general had a higher CoG, and higher intensity. When looking at /s/ and /ʃ/ specifically, the fricatives that are also studied here, only CoG systematically differed between high-F0 and low-F0 contexts; intensity did not differ between contexts in either fricative, and the CoG range did not differ between the contexts in /θ/.

It can be assumed that if consonant realizations are influenced by intonation in normal speech, they will also be in whisper. The goal of this investigation was to study the acoustic characteristics of whispered fricatives produced at different heights. In addition, the acoustic correlates to height in whisper were compared to characteristics of fricatives produced in normal speech, and interpreted as secondary or compensatory cues to pitch. On the one hand, under the assumption that speech is redundant, consonants may contain acoustic cues secondary to F0 in normal speech, and these cues might be present in comparable ways in whisper. On the other hand, given that speakers’ options for coding intonation in whisper are restricted, it may be the case that the segments, including the consonants, are altered more or in different ways by whisperers to express intonation. The latter would be predicted under the hyper- and hypo-speech theory (Lindblom 1990), which states that the speaker adapts the speech signal to the needs of the listener. Speakers adapt their speech when addressing e.g. hearing impaired listeners or listeners in a noisy surrounding (e.g. Krause and Braida 2004; Picheny et al. 1986). In the case of whispered speech, the listener may have higher needs as to how speakers convey intonation, prompting speakers to use more hyper-speech.

2. Method

2.1. Materials
Nonsense vowel-consonant-vowel (ViCVi) structures were used in which both vowels were one of the three corner vowels (/i, a, u/) and the same within an item, and the consonant was a voiceless fricative, either /s/ or /ʃ/. This resulted in the stimuli isi, ili, asa, afa, usu and ufu. To collect both whispered and normally spoken VCVs at different heights, speakers produced four subsequent, identical VCVs in either a rising or a falling series. The average range of speakers’ productions was 7.5 semitones. Measurements were taken from the first three realizations only. The fourth was spoken at the same height as the third and had been added to prevent boundary effects (e.g. final lengthening) on the third target item.

Twelve native speakers of Dutch took part in the recordings (6 males, 6 females, with self-reported normal hearing), and each was paired up with a listener (3 male-male,
5 male-female, 4 female-female pairs). The speaker-listener setup was intended to record listener-directed rather than read speech. The speaker was seated in a sound-treated booth and spoke into a Rode NTG-2 condenser microphone with ‘deadcat’ windscreen, connected to an Edirol R-44 portable solid state recorder (44.1 kHz, 24 bits, mono). Written production targets were presented in pseudorandom order on a computer screen, and the height change in the course of the series was indicated by shifting the subsequent production targets visually in either upward or downward directions on screen. The listener was seated outside the recording booth wearing Sennheiser HD 414 SL headphones listening to the speaker’s productions on both ears. Speaker and listener could not see each other. After each series, the listener indicated if the speaker had just produced a rising or a falling series, by means of hitting a key on a keyboard. The correctness of this judgment was directly presented on the speaker’s computer screen as listener feedback. After feedback presentation, the next production target appeared on screen. The listener was not informed about the correctness of his/her responses.

Participants received written instructions, filled out consent forms, and were paid a small amount for their efforts. Recording time was 3-5 minutes per speech mode, and the entire session lasted about 30 minutes. Listeners labeled 98% of phonated series and 91% of whispered series correctly (note that each speaker-listener pair was different). Using a 6-point Likert scale (1 = very difficult, 6 = very easy) speakers rated the difficulty of their task, directly after the recording of either speech mode. A Wilcoxon signed ranks test for paired samples showed that the task was judged more difficult in whisper (median = 4) than in normal speech (median = 6), Z = -2.9, p < .01. Both scores were in the upper half of the scale, suggesting that the task was perceived as double in both speech modes.

In normal speech, two repetitions of each of the six series were recorded (2 repetitions × 2 height directions × 6 stimuli = 24 items), and in whisper three repetitions of each were collected (36 items). The order in which speech modes were recorded was counterbalanced over subjects, and four different pseudorandom lists were used per speech mode. Data were transferred from the portable recorder onto a computer and each series was saved as a separate wave file (44.1 kHz, 24 bit). Each wave file was semi-automatically annotated at the phoneme, word and series levels. Per item, one instance was annotated manually, which was used to automatically annotate all other instances of the same item using a dynamic time warping procedure in PRAAT (Boersma and Weenink 2013). These were then manually checked, and corrected if necessary. For analysis, the second repetition of each utterance was selected. If that was not available, due to voicing in a fricative or a vowel (the latter in whisper only) or to non-speech sounds (tongue smack, etc.), another repetition was chosen.

### 2.2. Acoustic measurements and analysis

In the present study, higher productions are expected to show a higher centre of gravity (e.g. Niebuhr 2012), and, with increasing height, speakers may need to put in more effort, as in clear speech. This has been associated with longer durations (e.g. Picheny et al. 1986), and an increase in intensity in the 1-3 kHz range (Krause and Braida 2004). A less negative or even a positive spectral slope can also be expected for higher productions due to higher effort involved (e.g. Glave and Rietveld 1975; Gauffin and Sundberg 1989). The acoustic measures taken per fricative were (1) relative duration (that is the proportion of C within the VCV), (2) intensity development in dB by taking measurements at 20%, 40%, 60%, and 80% into the segment, (3) CoG in Hertz over the 0.5-8 kHz range (weighted by the power spectrum), and (4) spectral balance in decibels over the 0.5-2 and 2-8 kHz frequency regions in the fricative’s long term average spectrum. For acoustic analysis, 864 instances were used (12 speakers × 2 speech modes × 6 stimuli × 2 height directions × 3 heights). Measurements were taken from all fricatives, using PRAAT. Per acoustic cue, a repeated measures (RM) ANOVA was run with within-subject factors Speech Mode (normal, whisper), Fricative (/s/, /f/), Vowel Context (/a/, /i/, /u/), Height (low, medium, high), and Direction (rising versus falling). Time (20%, 40%, 60%, 80% into the segment) was also included as a factor for analyzing intensity development.

If sphericity was not met, Huynh-Feldt corrections were applied.

Correlations between the different measurements were computed, which showed a high correlation between CoG and spectral balance in both speech modes (whisper: r = .87, p < .001; normal speech r = .86, p < .001). For that reason, only CoG results are given below. The analysis was expected to show different types of results. In addition to effects of height, intrinsic differences between the speech modes and between the fricatives were expected (e.g. as speech mode or fricative main effects). Relevant to the research questions were main effects of height and interactions of height by speech mode. A main effect of height would indicate a secondary cue to height, as the cue would be available similarly in normal speech and in whisper. A speech mode by height interaction could indicate a compensatory cue to height, namely in the case that the acoustic difference is present only or larger in whispered than in normal speech. In the results, effects directly relevant to the research questions are presented first, followed by a discussion of other effects present in the data.

### 2.3. Results

#### 2.3.1. Relative duration

Relative fricative duration showed a main effect of height [F(2,20) = 11.8, p < .001]. Across speech modes, relative duration decreased with increasing height from 0.272 for low to 0.258 for high productions. Within normal speech only, the effect was significant, [0.28 versus 0.26: F(1,5, 14.6) = 12.6, p = .001]. There was no interaction of speech mode by height [F(2,20) = 1.6, n.s.]. The results are shown in Figure 1.

![Figure 1: Relative duration by stimulus type and height. Error bars indicate 95% CI of the mean.](image)

There was a main effect of vowel context [F(2,20) = 7.7, p = .003]. In the context of the vowels /i/ and /u/, the relative duration of the fricative was longer than in the context of /a/, as evidenced by a fricative by vowel context interaction [F(2,20) = 16.8, p < .001]. There were interactions of speech...
mode by vowel context \([F(2,20) = 11.5, p < .001]\), and of height by direction \([F(2,20) = 14.9, p < .001]\). In normal speech, showing a marginally significant effect of vowel context \((p = .05)\), relative durations were more comparable between vowel contexts than in whisper, in which the relative duration of fricatives in an \(/\mathrm{a}/\)-context (0.241) was shorter than that of fricatives in \(/\mathrm{i}/\)- and \(/\mathrm{u}/\)-contexts (0.266 and 0.267, respectively, \(F(2,22) = 11.6, p < .001\)). The height by direction interaction indicated that height differences impacted on relative fricative duration mainly in the rising series; in the falling series durations were much more comparable. Finally, the latter interaction was further qualified by an interaction of vowel context by height by direction \([F(4,40) = 3.0, p = .031]\), showing that the main effect of height was followed across vowel contexts in the rising series, but showed a vowel context dependency in the falling series.

### 2.3.2. Intensity in time

Fricative intensity in time showed a main effect of height \([F(1,4,14.0) = 15.5, p = .001]\), and the speech mode by height interaction was just significant \([F(2,20) = 3.6, p = .046]\). As figure 2 shows, the interaction indicated that the absolute intensity differences between different heights were slightly larger in normal speech \([F(2,20) = 13.3, p < .001]\), than in whisper \([F(2,22) = 4.5, p = .023]\). In general, intensity decreased with decreasing height. Post-hoc t-tests for paired samples, (corrected to \(\alpha = .05/3 = .017\)) showed a significant difference between medium (58.6 dB, sd = 3.7 dB) and low fricatives (57.6 dB, sd = 3.8 dB) in normal speech \([t(11) = 3.4, p = .006]\), and between high (55.6 dB, sd = 5.0 dB) and low (54.7 dB, sd = 5.1 dB), and high and medium (54.7 dB, sd = 5.4 dB) fricatives in whisper \([t(11) = 2.8, p = .017]\).

The interaction of time by speech mode, \([F(1,7,17.2) = 7.7, p = .005]\), reflected that in normal speech the intensity of the fricative decreased from beginning to end \([F(1,7,17.1) = 21.6, p < .001]\), whereas in whisper, intensity remained more constant throughout the fricative \([F = 1.0, n.s.]\). Intensity over the course of the fricative furthermore showed main effects of time \([F(2,0,19.6) = 8.5, p = .002]\), of fricative \([F(1,10) = 70.7, p < .001]\), and of vowel context \([F(2,20) = 10.2, p = .001]\). Across speech modes, at 80% into the fricative, intensity decreased relative to earlier in the segment, by about 1 dB. The fricative \([s]\) had more intensity (61.3 dB) than the fricative \([\mathrm{f}]\) (52.3 dB), and in the context of the vowel \(/\mathrm{a}/\), fricatives had less intensity (55.5 dB) than in the context of \(/\mathrm{i}/\) or \(/\mathrm{u}/\) (57.3 and 57.7 dB, respectively).

A time by fricative interaction reflected that the intensity of \([\mathrm{f}]\) hardly changed throughout the segment, whereas that of \([s]\) showed a small fall in intensity at its end, \([F(1,7,16.9) = 11.6, p = .001]\). A fricative by vowel context interaction indicated that intensity differences by vowel context were larger for \([\mathrm{f}]\), with a range of 3.5 dB between lowest and highest intensity, than for \([s]\), with a range of 1 dB \([F(2,20) = 9.3, p = .001]\). An interaction of height by direction, \([F(2,20) = 6.5, p = .007]\), showed that the effect of height was present in the falling direction \([F(2,20) = 18.3, p < .001]\), but not in the rising one \([F < 1]\). Relatively small differences in this interaction between the two fricatives resulted in a fricative by height by vowel context effect \([F(2,20) = 3.5, p = .050]\). There were further interactions of time by fricative by vowel context, reflecting slight differences in intensity development per fricative and vowel context \([F(2,8.27.5) = 7.7, p = .001]\), and of fricative by vowel context by direction, showing that intensity differences by fricative and vowel context were more comparable in the rising than the falling series \([F(2,20) = 4.9, p = .018]\). There finally was an interaction of time by speech mode by fricative by vowel context \([F(6,60) = 3.4, p = .006]\).

### 2.3.3. Centre of gravity

There was a main effect of height \([F(2,20) = 17.0, p < .001]\), showing that CoG varied with height, and in the expected direction with means of 3504 Hz for low and 3720 Hz for high productions (see Fig. 3). Post-hoc analyses showed that the height main effect was present in both whisper, \([F(2,22) = 17.0, p < .001]\), and normal speech \([F(2,20) = 8.7, p = .002]\). The mean difference between high and low productions was 186 Hz in normal speech and 245 Hz in whispered speech. There was no speech mode by height interaction \([F < 1]\). CoG showed further main effects of fricative \([F(1,10) = 88.7, p < .001]\), and of vowel context \([F(2,20) = 44.1, p < .001]\). These indicated that, across speech modes, CoG was higher in \([s]\) than in \([\mathrm{f}]\), and that CoG varied with vowel context, being lowest in between \(/\mathrm{a}/\), and highest between \(/\mathrm{i}/\) and \(/\mathrm{u}/\).

The interaction of height by fricative indicated that the absolute differences between the height steps were somewhat larger for \([s]\) than for \([\mathrm{f}]\) \([F(2,20) = 4.5, p = .024]\). A fricative by vowel context interaction showed that CoG of \([\mathrm{f}]\) seemed to be affected more by vowel context than that of \([s]\), \([F(2,20) = 10.4, p = .001]\). A vowel context by height interaction \([F(4,40) = 3.2, p = .024]\), showed that in the context of \(/\mathrm{i}/\) and \(/\mathrm{u}/\) mean differences in CoG with height seemed clearest, and a height by direction interaction showed a larger effect of height on CoG in the falling versus the rising series \([F(2,20) = 9.8, p < .001]\). The four-way interaction of speech mode by height by fricative by direction was significant \([F(2,20) = 5.6, p = .012]\). In normal speech, but not in whisper, there was a three-way interaction of height by fricative by direction \([F(2,20) = 3.6, p = .046]\), indicating that CoG of \([\mathrm{f}]\) in the rising series followed...
a trend opposite to the main effect. The interaction involving all factors was also significant \([F(4,40) = 3.0, p = .029]\).

3. Discussion and conclusion

The acoustic analysis was aimed, in the first place, at revealing acoustic correlates of height in the fricatives /s/ and /l/ in whispered speech. In the second place, the acoustic correlates were interpreted as secondary or compensatory cues to height through a comparison with normal speech measurements.

Across speech modes, relative fricative duration was influenced by height — longer durations were found for lower productions — but within whisper this was only a trend. As the absence of a main effect of speech mode shows, the fricatives /s/ and /l/ were not systematically longer in whispered than in normal speech. Relative durations in whisper showed longer fricatives in /i/- and /u/-contexts than /a/-contexts. Possibly, the larger movement from the open vowel to the fricative and vice versa (in /a/-contexts) allows less time for the fricative to be pronounced. This would be expected in the normal speech results as well, and such a trend was observable, with fricatives in /u/-contexts having a larger relative duration (0.25) than in /a/- and /i/-contexts (0.27 vs 0.26).

Intensity decreased with decreasing height. Actual differences were somewhat larger in normal than in whispered speech, but in both speech modes, effects were in the order of 1 dB between low and high or high and medium productions. The size and direction of this effect are in line with results reported by Niebuhr (2012) on normal speech. A 1 dB difference in intensity, as found in both studies, may not be a very informative cue, though, as it is at the same level as fluctuations in intensity caused by the speaker changing his/her head direction, and it furthermore fell well within the range of standard deviations obtained. One difference of possible interest between normal and whispered speech was the fact that intensity remained more constant throughout the fricative in whisper than in normal speech. On average, intensity fell over 2 dB in the course of a normal speech fricative, whereas it remained constant in whisper. The maintenance of the energy level throughout the segment may indicate the speaker’s effort to allow the listener to extract information from the segment. Across speech modes, fricatives in the context of the vowel /a/ were produced with less intensity than in the context of /i/ or /u/, which may be due to a trade-off in the amount of energy available for producing the utterance as a whole.

The centre of gravity was also influenced by height, in both speech modes, with productions intended as higher showing a higher CoG. This effect is comparable to that reported in Niebuhr (2012) for normal speech, and in the current study, the effect was found in both speech modes.

The direction of productions (from low to high or from high to low) seemed to affect which acoustic cues differed most with height between the instances in a series. Relative duration was varied more (systematically) in the rising than the falling series, whereas intensity, CoG, and spectral balance showed larger effects of height in the falling than the rising one. A natural downward trend in the course of an utterance is declination, which entails decreases in both f0 and amplitude in the course of a phrase. Possibly, speakers have less difficulty producing spectral changes to signal height when an utterance follows this natural trend rather than an upward one.

In sum, acoustic correlates of height were found for all measures taken, and in highly comparable ways in the two speech modes. Therefore, most correlates were secondary cues; only in the case of intensity development in time can it be argued that compensatory information was provided.

4. Acknowledgements

This work was supported by a VENI grant from the Netherlands Organisation for Scientific Research (NWO).

5. References


Computational modelling for syllabification patterns in Tashlhiyt Berber and Maltese

Anne Hermes, Bastian Auris and Doris Mücke

IFL Phonetik, University of Cologne, Germany
{anne.hermes, bastian.auris, doris.muecke}@uni-koeln.de

Abstract

In the present study, we use a computational model proposed by Shaw and colleagues to evaluate the behaviour of stability indices for word initial clusters. Since phonetic stability indices are known not to be robust against prosodic and contextual variation, this computational model utilises phonetic parameters to predict phonological syllable organisation from phonetic data by increasing the overall variability. We investigate coordination patterns of syllable organisation in two languages: Tashliyiht Berber, a language spoken in Morocco that has been shown to allow for simple onsets only and Maltese, a language with Arabic (Semitic), Romance and English strata, that has not yet been investigated in terms of gestural syllabification patterns. On the basis of kinematic data of one native speaker per language, we first computed heuristic stability indices, which – in both languages – point to conflicting results of simple and complex onset coordination. In a second step, we applied the computational modelling to the data sets, and the simulation results clearly lead to the conclusion that there is underlying simple onset coordination for clusters in both languages, Tashlhiyt Berber and Maltese.

Keywords: syllabification patterns, simple onset, simulation, computational modelling, phonetic stability indices, Tashlhiyt Berber, Maltese

1. Introduction

In a number of studies within the framework of Articulatory Phonology (Brown & Goldstein 1988), it has been shown for languages such as American English, Georgian, Tashlihiyt Berber, Moroccan Arabic, Romanian and Italian, that syllable structure is reflected in the coordination of articulatory gestures (Marin & Pouplier 2010, Goldstein et al. 2007, Hermes et al. 2011, Shaw et al. 2011, Hermes et al. 2013). Patterns of syllable organisation for word initial clusters show either complex onset coordination (COC: e.g. American English, Italian, Romanian and Georgian) or simple onset coordination (SOC: e.g. Tashlihiyt Berber and Moroccan Arabic).

Complex onset coordination (COC): The center of both consonants (i.e. the temporal midpoints of the consonantal targets, or C-center) is timed with the vowel. When a consonant is added to the beginning of the word, the prevocalic consonant is shifted to make room for the added consonant (Fig. 1, left).

Simple onset coordination (SOC): The added consonant does not interfere with the timing of the prevocalic consonant and the vowel. There is no rightward shift of the prevocalic consonant (Fig. 1, right).

However, temporal stability indices are not robust against prosodic and contextual variations and may not always hold. In some cases, phonetic indices can point to conflicting conclusions, resulting in a stochastic backdrop of syllable organisation. Shaw et al. (2011) and Gafos et al. (2014) showed that under increasing variability, phonetic indices change from values associated with simple onset coordination to those associated with complex onset coordination. They used a computational model of temporal organisation to generate temporal patterns of change in Moroccan Arabic consonant clusters by using either simple or complex syllable organisation. They simulated increased variability in the temporal intervals within the syllabic constituents. Surprisingly, the system changed its behaviour by showing crossover points in the temporal stability indices over time.

In this study we assess the phonological syllable organisation for word initial consonant clusters in Tashlhiyt Berber and Maltese by means of quantitative simulation with the ParseEval function in R (cf. Auris 2013, Shaw et al. 2011, R Core Team 2013). We modified recorded kinematic data from both languages and fed this into the computational modelling to test for syllabification patterns.

2. Method

2.1. Kinematic Data for Simulation

For Tashlhiyt we used kinematic data from experimental work reported in Hermes et al. (2011) and Ridoane et al. (2014). Additionally, we analogously recorded lip and tongue kinematics from one native speaker of Maltese with an Electromagnetic Articulograph (AG500).

Figure 2 gives an example of the lip aperture gesture for /p/ in the target word /spina/ from Maltese. From top to bottom the figure displays vertical displacement and vertical velocity. High values indicate that the lips are open (start of the labial plosive) and low values that the lips are closed (target and plateau of the plosive). The following landmarks are shown: onset (blue line), peak velocity (green line), target (red line) and plateau onset and offset (dashed lines). Plateau onset and offset were computed on the vertical velocity signal with a 20% threshold.
Based on the theoretical approach by Shaw et al. (2011), we automatically derived a set of phonetic parameters from the recorded kinematic data (see Figure 3). We computed the plateau duration by subtracting plateau onset from plateau offset of the target consonant(s) and the interplateau duration by subtracting plateau offset of C1 from plateau onset of C2.

Further, we identified the vocalic gesture duration, from the gestural onset to the vocalic target of the tongue body movement for the vowel /i/ in the accented syllable.

### 2.1.1. Speech Material

For Tashlhiyt Berber, the computational modelling was applied to the clusters /kf, lk, kt/ embedded in the carrier phrase “Inna ___ bahra” (‘He said ___ a lot’):

(i)  \( \text{fik} - \text{kfik} \) (‘give yourself’ - ‘give yourself’)
(ii) \( \text{lkf} - \text{lkfif} \) (‘same’ - ‘hashish’)
(iii) \( \text{ktid} - \text{ktid} \) (‘those’ - ‘remember’)

Note that for Tashlhiyt data the anchor for calculating stability indices is the target of the coda consonant (i.e. within the same syllable, e.g. /k/ in fik), since the vocalic target was difficult to identify.

For Maltese, the clusters /sp, sf, tf/ went into the analysis. The target words were embedded in the carrier phrase “Irrepeta l-kelma ___ darbejtin” (‘I repeat ___ twice’):

(iv) \( \text{Pina} - \text{spina} \) (Proper name - ‘backbone’)
(v) \( \text{ilha} - \text{ilha} \) (‘row’ - ‘unthread’)
(vi) \( \text{dfina} - \text{dfina} \) (‘delicate’ - ‘buried’)

Note that there is voicing assimilation in <dfina> - <tf>/

The anchor in the Maltese data set was the target of the vocalic gesture for /i/ within the same syllable.

### 2.2. Parameters for Simulation

The next step to run the simulation is to calculate phonetic stability indices (Relative Standard Deviation, RSD in %) obtained for the latencies of the RE, LE and CC (Figure 3) to the following anchor by dividing the standard deviation by the mean for the single consonant and cluster conditions:

- **Right edge (RE):** Calculates the latencies for the plateau offset of the rightmost consonant relative to the anchor.
- **Left edge (LE):** Calculates the latencies for the plateau onset of the initial consonant relative to the anchor.
- **C-center (CC):** Calculates latencies from the midpoint of the consonantal targets relative to the anchor.

The computational model takes as its input the empirical data (kinematic landmarks and stability indices) and, during the computation, assesses the goodness of fit (using ordinary least squares regression) between the input and the simulated RSD with an increase in overall variability (anchor variability in 15 steps of 5ms). Based on empirical data, this process of simulation (Nsim=1000) offers a convenient method to hint at possible syllabification patterns.

For all simulated clusters, we tested both syllabification patterns: simple onset coordination (SOC) and complex onset coordination (COC). SOC is quantified by the fact that there is underlying right edge stability, i.e. the right edge of a cluster is more stable than C-center to anchor. In turn, the opposite relation quantifies COC. It is expected that there is center stability, i.e. the C-center to anchor distance is more stable than the right edge to anchor.

The applied algorithm produces results based on both of these two different syllabification patterns.

### 3. Results

#### 3.1. Tashlhiyt

##### 3.1.1. Phonetic Indices

Figure 4 provides the phonetic indices of syllable structure (RSD) computed from the kinematic data of one native speaker of Tashlhiyt for the target words /fik-kfik/ , /kif-lkif/ and /tid-ktid/ for the left edge (LE), C-center (CC) and the right edge (RE) variables. The lower the RSD values, the greater the anchor stability.

![Phonetic Indices for Tashlhiyt clusters /fik-kfik/ /kif-lkif/ and /tid-ktid/ for left edge (LE), C-center (CC) and right edge (RE) to anchor.](image-url)
The RSD values for Tashlhiyt do not show a clear pattern toward a certain stability pattern. There is no clear difference between center stability and right edge stability (/fik-kfik/: LE=12.38, CC=7.73, RE=9.56; /kif-kfik/: LE=20.42, CC=17.50, RE=17.00; /tid-ktid/: LE=11.76, CC=10.19, RE=11.36), leading to conflicting results. On the basis of the stability indices, we can only state that in the Tashlhiyt data set – LE is the least stable anchor as it has the highest RSD values.

3.1.2. Simulation

Since results for the kinematic data in 3.1.1 are conflicting, the computational modelling is applied to these clusters in order to investigate the behaviour of the system when anchor variability is increased. Therefore, we applied the ParseEval function to the Tashlhiyt clusters. And indeed, the simulation clearly predicts a simple onset coordination pattern (SOC).

Table 1 presents the ratio of the total number of hits to the number of simulations for the investigated clusters. The larger the numbers of hits, the more stable the coordination patterns observed in the system’s behaviour are. The ratio points towards a SOC pattern, e.g. /kif-kfik/ shows a high hit rate for simple onset coordination at 973/1000 and a low one for complex onsets at 0/1000.

Table 1: Ratio of total number of hits/number of simulations for SOC and COC in Tashlhiyt clusters.

<table>
<thead>
<tr>
<th>Anchor</th>
<th>SOC</th>
<th>COC</th>
</tr>
</thead>
<tbody>
<tr>
<td>/fik-kfik/</td>
<td>163/1000</td>
<td>13/1000</td>
</tr>
<tr>
<td>/kif-kfik/</td>
<td>973/1000</td>
<td>0/1000</td>
</tr>
<tr>
<td>/tid-ktid/</td>
<td>128/1000</td>
<td>23/1000</td>
</tr>
</tbody>
</table>

Figure 5 presents the simulation results for the cluster /kif-kfik/ graphically. The diagnostic plots include the simulation of both tested coordination patterns, i.e. (a) simple onset coordination (SOC, Fig. 5a.1-3 left) and (b) complex onset coordination (COC, Fig. 5b.1-3 right) under increasing anchor variability in 15 steps of 5ms.

Plots a.1 and b.1 display the trend of median RSD values for each landmark (LE, CC and RE) in relation to the respective anchors throughout. These plots are used to assess the approximate relation between RSD values and further infer if the underlying assumption (SOC or COC) holds. However, upon noticing oddities or unexpected relations further simulation analyses, as in a.2-3 and b.2-3, are required.

Plots a.2 and b.2 display the number of hits recorded during the simulation. Whenever the (internal) regression analyses of RSD values result in a statistically significant F-value that is greater than 98.50, a hit is recorded for this landmark and the respective anchor.

In the third row of plots, a.3 and b.3, the median R^2 values are displayed from the regression analyses of the RSD values in relation to the anchor. This simulation measure is applied to quantify the internal regression analyses used during the simulation. Note that the combination of a high number of hits paired with high values of R^2 hints at very reliable results whereas a large number of hits paired with low values of R^2 hints at less reliable results of the simulation.

Taking the hit rate as well as R^2 into account the simulation results for Tashlhiyt clearly support simple onset coordination, SOC. Therefore, as expected, testing the complex onset coordination pattern, COC, leads to the opposite results, in that the overall hit rate is very low and thus, do not indicate a complex coordination pattern at all. These simulation results are therefore in line with the assumption that Tashlhiyt allows for simple onsets only (Goldstein et al. 2007, Hermes et al. 2011, Ridouane et al. 2014).

3.2. Maltese

3.2.1. Phonetic Indices

In Figure 6, the RSD values for the LE, CC and RE are presented as computed from the kinematic data of one native speaker of Maltese for the target words (/pina-spina/, /fila-sila/ and /fina-dfina/ [fina-fina]). Note that high RSD values indicate more anchor stability than low values.

Again, we find conflicting stability patterns in that there is no obvious difference between center and right edge stability (/pina-spina/: LE=19.23, CC=14.92, RE=14.93; /fila-sila/: LE=24.12, CC=17.32, RE=16.06; /fina-dfina/: LE=25.71, CC=20.19, RE=20.41). On the basis of the stability indices computed from the kinematic data, it is difficult to draw a conclusion regarding the coordination pattern in Maltese.
3.2.2. Simulation

We applied the ParseEval function to the Maltese clusters (details of the simulation procedure are given in section 3.1.2). Table 2 provides the ratio of total number of hits in relation to the number of simulations. The larger the numbers of hits ratio, the more stable the coordination patterns observed in the system’s behaviour. The hit ratios already indicate simple onset coordination, e.g. /fila-sfila/ show high values for simple onset coordination with 547/1000 and low values for complex onsets 0/1000.

Table 2: Ratio of total number of hits in relation to number of simulations for SOC and COC in Maltese clusters.

<table>
<thead>
<tr>
<th></th>
<th>SOC</th>
<th>COC</th>
</tr>
</thead>
<tbody>
<tr>
<td>pina-spina</td>
<td>539/1000</td>
<td>3/1000</td>
</tr>
<tr>
<td>fila-sfila</td>
<td>548/1000</td>
<td>0/1000</td>
</tr>
<tr>
<td>fina-dflina</td>
<td>323/1000</td>
<td>8/1000</td>
</tr>
</tbody>
</table>

The simulation results further help to quantify the coordination pattern in Maltese (see Figure 7). The simulation results are similar to the ones presented for Tashlhiyt Berber and Maltese. Although the stability indices could lead to both a simple onset and complex onset analysis, the simulation clearly indicates a simple onset coordination pattern. The RE to anchor is the most stable variable and with respect to hit rate and median R² values, a simple onset coordination pattern can be observed. Thus, the simulation results testing complex onset coordination are clearly not applicable to the Maltese clusters.

Figure 7: Simulation results for Maltese /fina-dflina/ [fina-tfolna] for (a) SOC and (b) COC displaying from top to bottom: Median RSD, Hits and Median R².

4. Discussion and Conclusion

We have demonstrated across the two investigated languages, Tashlhiyt Berber and Maltese, that the computational model of temporal organisation provides a quantitative account as to the stability of the underlying pattern of syllable organisation by increasing its overall variability. The simulation increases anchor variability to generate patterns of change. Even though the analysis of phonetic stability indices leads to conflicting conclusions (cf. Shaw et al. 2011, Gafos et al. 2014), the simulation successfully diagnoses one type of coordination: in our study, neither languages showed a clear pattern with regard to the phonetic indices, but the system’s behaviour in the simulation clearly points in the direction of stable simple onset coordination for these word initial clusters. This is in line with the assumption that Tashlhiyt allows for simple onsets only (Goldstein et al. 2007, Hermes et al. 2011, Ridouane et al. 2014), while for Maltese, this is – to our knowledge – the first study on onset coordination patterns.

5. Acknowledgements

We would like to thank Jason Shaw for providing the original Matlab ParseEval script.

6. References


Resolution of lexical retrieval failures.

Reaction time data in the tip-of-the-tongue paradigm

Nina Jeanette Hofferberth
Goethe-University Frankfurt, Germany
n.j.hofferberth@gmx.de

Abstract

The tip-of-the-tongue phenomenon refers to the experience of having a word on the tip of one’s tongue: There is access to the word’s lemma but only partial access to the word’s lexeme. A common paradigm is to create definitions or questions to induce TOTs and measure the probability of unsuccessful retrieval. Whereas reaction time (RT) data has frequently been used to measure successful retrieval in tasks like picture naming, the present study reports a novel use for RTs in the area of ‘impaired’ lexical access. RT data allows for a more objective measure of TOTs, as opposed to participants’ self-report.

Keywords: tip-of-the-tongue (TOT), reaction times (RT), lexical access, cueing, first syllable

1. Introduction

Lexical retrieval failures, tip-of-the-tongue (TOT) states in particular, can be frustrating. The experience of having an unsuccessful search for a word that exists in the mental lexicon but cannot be retrieved can happen several times a week in everyday life. Speakers have a strong feeling of knowing the word, have access to semantic (concept) and syntactic information (lemma), but the access to the word form (lexeme) is impaired. Often, speakers have partial information in a TOT state, such as the initial phoneme/grapheme, the number of syllables, and the word accent. The target word or name can pop-up imminently or it can take minutes, hours or even days for successful lexical retrieval (A. S. Brown 2012). The resolution of TOTs can encompass external strategies (such as looking up the word in a dictionary or by asking someone) and internal strategies (such as searching in the alphabet or generating similar words).

There are two different explanations for the etiology of TOTs. First, the blocking hypothesis refers to interlopers (that are phonologically and/or semantically similar to the target word) that come to mind while searching for the target word and prevent its retrieval. Second, the incomplete activation hypothesis refers to a weak and incomplete activation of the correct target word. This lack of activation can be explained with a transmission deficit, derived from node structure theory (MacKay 1987): the level of activation from the lemma to the lexeme node is not high enough to support word production. Therefore, the lexeme is incompletely activated, an effect that is exacerbated in old age (Abrams, Trunk, and Margolin 2007; MacKay and Burke 1990) and in speakers of multiple languages (Ecke 2009; Gollan and Acenas 2004). This deficit in transmission of activation also correlates with frequency and recency of word use.

The retrieval fluency, i.e., the ease and speed with which information comes to mind is decelerated in a TOT state. Normally, (picture) naming takes approximately 500 to 800 milliseconds:

When we see a picture or an object, we typically spend the first 150 ms doing visual processing and activating the appropriate concept. We then spend another 125 ms or so selecting the lemma. Phonological encoding starts around 275 ms and we usually start uttering the name from 600 ms (Harley 2001, p. 365). Harley (2001) describes the TOT state as “an extreme form of a pause” (p. 362).

2. Previous studies

Methods in TOT research are, for example, the creation of question or answer sheets (e.g., R. Brown and McNeill 1966; for a replication in German see Hofferberth 2011), diary studies (e.g., Burke, MacKay, Worthley, and Wade 1991), interviews with persons about their TOT experience (e.g., Heinzlinger 1999) as well as fMRI studies (e.g., Maril, Wagner, and Schacter 2001).

For laboratory TOT research, a common method is to create definitions of (low-frequency) words in order to induce TOTs. The definitions or questions can be read aloud or be presented on a computer screen, for example “A cigar-shaped, self-propelled underwater projectile launched from a ship, submarine or aircraft, and designed to explode on reaching a target” for torpedo.

Cueing may help resolving the TOT, especially phonological cues, which are thought to boost activation to the target relative to unrelated words (Abrams, White, and Eitel 2003; Burke, MacKay, Worthley, and Wade 1991; Harley and Bown 1998; James and Burke 2000; Rastle and Burke 1996; White and Abrams 2002).

James and Burke (2000) were the first to use syllable cues. They embedded the first, middle and last syllable of the target cumulatively in a list of cue words, e.g., indigent, abstract, truncate, tradition, locate for the target abdicate (syllables in italics here for clarity). The effect of reading this list after having a TOT was increased TOT resolution relative to an unrelated list. White and Abrams (2002) found that the locus on syllable cueing is the first syllable: the first syllable (embedded in a cue word) facilitated TOT resolution while the middle or last syllable did not. Abrams, White, and Eitel (2003) demonstrated that TOT resolution requires the entire first syllable; the first phoneme or first grapheme alone had no effect. TOT resolution was thereby facilitated by activating the initial syllable via related words sharing that feature.
3. Experimental methodology

Here, the first syllable was presented individually to avoid providing any semantic information, and to prevent interlopers. For a comparison of TOT designs, see Hofferberth-Sauer and Abrams (in press).

A new method presented here is the measurement of reaction times (RTs) to indicate a TOT vs. non-TOT state. RTs belong to one of the oldest and most fruitful paradigms in experimental psychology research used to draw conclusions from mental processes that are not directly observable. RTs are a classic behavioural method for an explicit measurement of the temporal relationship between signal and response. In contrast to off-line methods of language processing (such as elicitation or grammatical judgment task) which give the products after the processing of speech is completed, on-line methods (such as RTs) allow the investigation of language processing while processing is on-going. RTs have been used in the area of lexical access for decades in tasks such as picture naming or lexical decision where words are named successfully, but RT data are not typically reported when lexical access fails.

The experiment presented here used RTs to indicate the degree to which the word was known and able to be retrieved, which does not require a word to be spoken. This is a somewhat atypical use of RTs because on the one hand, it is not only about pressing a button (KNOW, DON’T KNOW, TOT) but also about a productive reaction (typing in the target word). On the other hand, there is no time pressure within cue presentation because the cue got presented for 25 seconds (not milliseconds). This experiment allowed for both RTs and entering the correct/incorrect answer to the word definition. If TOTs are a different kind of retrieval failure from DON’T KNOW responses, then RTs should be longer for TOTs because the lemma has been selected and thus gives a feeling of knowing the word that would encourage participants to keep searching in hopes of the word coming to them.

In previous work, correct cues were used to boost activation to the target word, but incorrect cues have not been used so far. TOT resolution might get hampered with incorrect cues because people’s search for the target begins with misinformation, activating an incorrect syllable, which will transmit activation to other words containing that syllable and not the target.

3.1. Aim and scope

After two pilot studies (Hofferberth 2012), in order to collect definitions and to increase TOTs (by reducing DON’T KNOW answers), a reaction-time experiment was performed to test three hypotheses:

1a. Participants will be faster to report KNOW than DON’T KNOW or TOT,
1b. RTs for TOT responses will be slower than DON’T KNOW responses.
2a. The correct first syllable will facilitate lexical retrieval by speeding up TOT resolution,
2b. and will lead to more accurate TOT resolutions.
3a. An incorrect syllable will inhibit lexical retrieval by slowing down TOT resolution,
3b. and will lead to more inaccurate TOT resolutions.

3.2. Method

3.2.1. Design

The study had a within-subject design. For the analysis involving a cue, the independent variable was the cue condition (correct syllable, incorrect syllable, control). The dependent variables were the RTs to press KNOW and type in the answer (to indicate TOT resolution) after being given a cue as well as the percentage of time a TOT was resolved.

3.2.2. Participants

60 under- and postgraduates (from Heinrich-Heine-University Dusseldorf, Bergische University Wuppertal and Goethe-University Frankfurt) participated in this study. They were native speakers of German and got paid for their participation. Four participants were excluded from data analysis because there were technical problems in data collection, i.e., the participant could not finish the experiment, and eight were excluded because they had not at least five TOTs per cue condition (not ≥ 15 TOTs out of 240 items). For data analysis, the answers of 48 students (30 female, 18 male) between 21 and 35 years (M = 29.5 years, SD = 3.7) were considered.

3.2.3. Apparatus and material

The definitions were presented on a computer screen using the program Presentation. 240 definitions of German nouns (English examples here are only for demonstration purposes) were presented in order to induce TOTs. The definitions were taken from DUDEN dictionary but were adapted (shortened and specified). The correct first syllable was the one of the target word, the incorrect syllable the first syllable of another target used in the study, and the control condition consisted of three x’s. The frequencies of the stimuli and of their first syllables were taken from the DLEX database and were used to ensure that the correct and incorrect syllables were matched in frequency.

3.2.4. Procedure

The participants were asked to press a button as fast as possible to indicate if they KNOW the word, DON’T KNOW the word, or if they experience a TOT. After pressing KNOW, they typed in the answer, and another definition was presented. After pressing DON’T KNOW, the next definition appeared on the screen. After pressing TOT, one of three cues appeared: either the correct first syllable (e.g., tor for torpedo), an incorrect syllable (e.g., gen, first syllable taken of another target), or the control condition (marked by xxx). The participants were given 25 seconds to type in the answer.

3.3. Results

The statistical analysis was done with repeated measures ANOVAS in order to differentiate between conditions. Significant effects were subsequently explored using paired samples t-tests.

3.3.1. TOT rate

The number of TOTs varied between 18 (7.5%) and 132 TOTs (55%). Through 11520 stimuli overall, 2326 TOTs were induced, i.e., the TOT rate was 20.2% with 48 TOTs per person on average (SD = 22 TOTs). Out of the 2326 TOTs, 1216 TOTs (52.3%) were resolved in the given time of 25 seconds, with RTs between 1688 ms and 22252 ms (M = 6725 ms; SD = 3654 ms). Out of these, 862 (37.1%) were accurate TOT resolutions (the answer of the
participant was consistent with the target word), 354 (15.2%) were inaccurate TOT resolutions (the answer of the participant differed from the target word), and 1110 TOTs (47.7%) stayed unresolved in the allowed time of 25 seconds.

3.3.2. Reaction times

The RTs differed between the three types of responses \( F(2, 94) = 41.9, p < .001 \): KNOW responses were faster than DON’T KNOW responses \( t(47) = 4.49, p < .01 \), which in turn were faster than TOT responses \( t(47) = 4.34, p < .01 \). The RTs included the reading times of the definition until a button was pressed to indicate the response. See table 1 and figure 1 for all values.

![Figure 1: Reaction times to definition (in ms).](image1)

<table>
<thead>
<tr>
<th>Cue</th>
<th>( M )</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOW</td>
<td>5265 ms</td>
<td>1085 ms</td>
</tr>
<tr>
<td>DON’T KNOW</td>
<td>5923 ms</td>
<td>1111 ms</td>
</tr>
<tr>
<td>TOT</td>
<td>6474 ms</td>
<td>1122 ms</td>
</tr>
</tbody>
</table>

Table 1: Reaction times after definition.

During a TOT, the RTs to resolve TOTs by typing in an answer were analysed as a function of the cue that was presented, revealing a significant effect \( F(2, 94) = 36.7, p \leq .001 \). The RTs started from cue presentation until the participant pressed the KNOW button to indicate TOT resolution. RTs were faster when the correct cue was presented in comparison with the incorrect cue \( t(47) = 6.92, p < .01 \) or the control condition \( t(46) = 10.42, p < .01 \). There was no difference between the RTs to the incorrect cue and the control condition \( t(46) = .40, p = .69 \). For all values see table 2 and figure 2.

![Figure 2: Reaction times to resolve TOTs after cue presentation (in ms).](image2)

<table>
<thead>
<tr>
<th>Cue</th>
<th>( M )</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct cue</td>
<td>4953 ms</td>
<td>1689 ms</td>
</tr>
<tr>
<td>Incorrect cue</td>
<td>7807 ms</td>
<td>3984 ms</td>
</tr>
<tr>
<td>Control (xxx)</td>
<td>7559 ms</td>
<td>2582 ms</td>
</tr>
</tbody>
</table>

Table 2: Reaction times to resolve TOTs after cue presentation.

3.3.3. Cue analysis

The number of accurate TOT resolutions differed between the three types of cues, \( F(2, 94) = 260.6, p < .001 \). With the correct first syllable, TOTs were accurately resolved more often \( M = 73.5%, SD = 18.6% \) in comparison to the control condition \( M = 24.3%, SD = 16.4%; t(47) = 16.4, p < .001 \), and in comparison to the incorrect syllable \( M = 16.0%, SD = 13.6%; t(47) = 20.1, p < .001 \). The control condition led to more accurately resolved TOTs than an incorrect syllable \( t(47) = 3.7, p = .001 \).

The number of inaccurate TOT resolutions also differed between the three types of cues, \( F(2, 94) = 17.6, p < .001 \). Fewer TOTs were inaccurately resolved with the correct first syllable \( M = 9.2%, SD = 7.8% \) in comparison to an incorrect syllable \( M = 15.8%, SD = 14.4%; t(47) = 3.6, p = .001 \), and in comparison to the control condition \( M = 20.2%, SD = 14.4%; t(47) = 5.9, p < .001 \). The incorrect syllable led to fewer inaccurate TOT resolutions than the control condition \( t(47) = 2.4, p = .023 \).

The number of unresolved TOTs also differed between the three types of cues, \( F(2, 94) = 152.3, p < .001 \). There were fewer unresolved TOTs with the correct first syllable \( M = 17.3%, SD = 16.9% \) in comparison to the control condition \( M = 55.6%, SD = 20.2%; t(47) = 12.3, p < .001 \), and in comparison to the incorrect syllable \( M = 68.2%, SD = 20.4%; t(47) = 15.8, p < .001 \). The control condition led to fewer unresolved TOTs than an incorrect syllable \( t(47) = 4.6, p < .001 \).

4. Discussion

The results offer support for several of the hypotheses. The RTs to report KNOW were faster than DON’T KNOW or TOT, consistent with hypothesis 1a. To report a TOT took significantly longer than reporting DON’T KNOW, consistent with hypothesis 1b. It can be concluded that the search in the mental lexicon for a word that is retrievable (KNOW) is fast and easy. In contrast, it takes longer to go through a mental word list with possible words but without the target word (DON’T KNOW), and even longer to search for a word that exists in memory but is only partially activated (TOT). RTs were useful indicators of dissociating TOT and DON’T KNOW responses. RT data are new in the area of TOT research and allow for a more objective measurement of TOTs instead of relying entirely on participants’ judgments.
Furthermore, RTs were useful to capture the facilitatory effect from a correct cue. Participants responded approximately twice as fast to resolve the TOT when given the correct syllable in comparison to an incorrect syllable or the control condition. With the correct syllable, TOT resolution was also enhanced, as TOTs were resolved about 4.6 times more often in comparison to an incorrect syllable, and about 3 times more often compared to the control condition. These results support hypothesis 2a and 2b, as the correct first syllable facilitated lexical retrieval by speeding up TOT resolution, and led to more accurate TOT resolutions. Results support the Transmission Deficit Hypotheses (TDH) that TOTs occur when activation fails to be fully transmitted from the lemma to the lexeme node. Cuing helps to overcome the transmission deficit: Presenting the first syllable helps to fill the first ‘slot’, and to reduce the number of competitors by strengthening the weakened phonological connections that caused the TOT, facilitating target retrieval.

However, RTs did not detect an inhibitory effect from an incorrect cue, contrary to hypothesis 3a; participants were not slower than in the control condition. An incorrect cue also did not increase inaccurate retrievals of some other word, contrary to hypothesis 3b. In fact, there were fewer inaccurate TOT resolutions following an incorrect syllable than the control condition. The lack of the expected inhibitory effects could be a result of the incorrect syllable not corresponding to another word that would be semantically appropriate for the definition. Therefore, participants may not have activated words starting with the incorrect syllable that would potentially compete with the target. Competition between possible lemmas may be necessary for delaying TOT resolution (slower RTs) and for more inaccurate resolutions to occur. But an incorrect syllable did demonstrate a unique inhibitory effect on TOT resolution, i.e., decreasing accurate retrievals and increasing unresolved TOTs, suggesting that an incorrect syllable may play some role in hampering the resolution process.

In sum, the present results extend the facilitating effect of the correct syllable seen with cue words sharing the first syllable of the target (White and Abrams 2002; Abrams, White, and Eitel 2003) to syllable cues presented in isolation. Furthermore, this experiment is the first to report an inhibitory effect of an incorrect syllable, although limited to one specific measure, i.e., accurate TOT resolutions. Currently, an experiment is being run with an extended syllable in order to investigate if the extended syllable will speed lexical access and facilitate TOT resolution more than the regular syllable. Future research should more fully explore the circumstances under which cues inhibit and facilitate the resolution of TOTs to contribute to better-specified theories of TOTs as well as speech production.

5. Acknowledgements

This Ph.D. project is funded by the Hanns-Seidel-Foundation (Munich).

I would like to thank my supervisor Prof. Dr. Helen Leuninger (Goethe-University Frankfurt) for mentoring and support, and Prof. Dr. Lise Abrams (University of Florida) for comments on an initial version of the paper.

Thanks to Frauke Hellwig (Heinrich-Heine-University Dusseldorf) for programming with Presentation, and to Matthias Hoenen (Heinrich-Heine-University Dusseldorf) for assistance with statistical analysis.

6. References


DLEX database: www.dlex.de

DUDEN dictionary: www.duden.de


Toward a Model of the Pharynx
Kiyoshi Honda¹, Jianguo Wei¹, Jianwu Dang²
¹Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
²School of Information Science, Japan Advanced Institute of Science and Technology, Japan
²khonda@sannet.ne.jp

Abstract

The pharynx expands and contracts during speech not only anteroposteriorly but also mediolaterally. Conventional lateral views of the speech organs suggest that the pharynx changes by movement of the pharyngeal surface of the tongue, whereas the pharynx actually deforms three-dimensionally largely by articulatory movement of the pharyngeal wall. Transverse dynamic MRI data indicates that the pharyngeal cavity changes quasi-concentrically between vowel /i/ and /a/ in a word utterance. Anatomy of the pharyngeal muscles is reviewed in search of the underlying mechanisms of the movement of the lateral pharyngeal wall, and the roles of some of the muscles are discussed in relation to the control of the fundamental frequency (F0) and voice quality in speech.

Keywords: speech physiology, pharyngeal wall, voice control

1. Introduction

The pharyngeal wall is an independent articulatory organ that controls expansion and constriction of the pharynx during speech, and its geometry is largely determined by the muscles located in the pharynx wall. The conventional lateral sketch of the speech organs describes the pharyngeal wall as a fixed structure, while the transverse motion images obtained by cine-MRI demonstrate a rapid motion of the pharyngeal cavity, repeating expansion at inhalation and constriction at utterances to various degrees. The action of the pharyngeal wall during speech has been least explored, and its role needs to be discovered by physiological studies. In this report, transverse deformation of the pharynx is demonstrated by cine-MRI data recorded during a short word utterance with vowels /i/ and /a/.

Then, muscular components for the pharynx wall are described to look into the underlying physiological mechanisms. Finally, we propose two conceptual models of pharyngeal muscle actions for fundamental frequency (F0) control and for voice quality control.

2. Pharyngeal wall movement

The shape of the pharynx has been biased in a researchers’ conception being derived from conventional sketches or lateral X-ray pictures of the speech organs: the pharynx is outlined by the pharyngeal surface of the tongue and the posterior pharyngeal wall. The former changes in position by articulatory movements of the tongue, and the latter is fixed as a solid wall. While our knowledge about the pharynx supports that the lateral wall of the pharynx moves inward and outward during speech, it tends to be underestimated in a general description of speech articulation or in models of the articulatory organs. The pharynx is an active component for articulation, and three-dimensional deformation of the pharynx during speech is best documented with magnetic resonance imaging (MRI). In this section, a cine-MRI dataset is used to show movements of the pharyngeal wall in four transverse slices together with a midsagittal image recorded during a Japanese utterance /kika/.

2.1. Cine-MRI Data

2.1.1. Combined sagittal and transverse scans

The cine-MRI dataset used in this study is a pilot data for a database that was recorded for two sagittal and four transverse planes in a single session. A Japanese male speaker repeated a Japanese word /kika/, while the synchronized cine-MRI technique (Masaki et al., 1999) was used for dynamic imaging. A 1.5-T Scanner (Shimadzu-Marconi, Magnex Eclipse Power Drive 250) at BAIC, ATR-Promotions, Kyoto was employed for the experiment. Sagittal scans were conducted for the midsagittal and parasagittal planes to construct a movie file of 60 fps for each plane, and transverse scans were performed within the same session for four planes in the pharynx to form a movie file of 30 fps. In the word utterance, the vowel /i/ is accented and devoiced according to the rule of Tokyo dialect, and the articulation of this vowel tends to be assimilated to the consonant /h/ for this vowel (Honda et al., 2007). Figure 1 is the midsagittal pre-scan showing the orientation of the transverse planes (Tr1 – Tr4). Each plane is vertically separated with a 15-mm interval.

![Figure 1: Midsagittal reference MRI at rest showing positions of four transverse slices (Tr1 – Tr4).](image)

2.1.2. Midsagittal and four transverse images

Three segments of the utterance were chosen to demonstrate articulatory deformation of the pharynx: rest before utterance initiation (about 200 ms before the first /k/ contact), vowel /i/, and vowel /a/. Midsagittal and four transverse images are shown in Figure 2 for each segment. In rest (Figure 2(a)), the vocal tract takes a neutral shape with the round tongue with the open lips and velopharyngeal port. Cross-sections of the pharyngeal cavity are longer laterally except for the slice of Tr2 with the palatine tonsils bulging into the pharynx. In vowel /i/ (Figure 2(b)), the pharyngeal surface of the tongue advances for this vowel, to a lesser degree in comparison to that in the voiced /i/. Cross-sections of the pharynx show that...
the cavity expands both anteriorly and laterally. The parapharyngeal structures (seen by the vessels) also show movement for lateral expansion. In vowel /a/ (Figure 2(c)), the pharyngeal surface of the tongue retracts to form a constriction in the mesopharynx with the back-tilted epiglottis and narrowed laryngeal cavity. Cross-sections of the pharynx demonstrate constriction of the pharynx in all planes. The hypopharynx (T4) is seen to divide into three passages by the epiglottis in contact with the pharyngeal wall.

3. Anatomy of the pharyngeal muscles

The pharyngeal wall is an independent articulatory organ that contributes to determining of the shape of the pharynx. This notion must firstly be supported by the anatomy of the pharyngeal muscles. The pharyngeal muscles are reviewed in this section in two groups: the constrictor muscles and internal muscles. A functional group of muscles is added to discuss the mechanism of pharyngeal expansion. Figure 3 is a schematic lateral view of the pharyngeal muscles drawn based on anatomical literature (e.g., Zemlin, 1998; Standing, 2008).

3.1. Pharyngeal constrictor muscles

The pharyngeal constrictors are the muscles that semi-circularly surround the pharynx by three parts, the superior, middle, and inferior. They overlap with each other in the manner that the superior one is innermost. Each muscle is thought to have two effects on narrowing the pharynx by shortening the muscle loop and by thickening the muscle bundle.

The superior pharyngeal constrictor controls the opening of the velopharyngeal port from the sides, assisting the closure of the velopharyngeal port. The upper (pterygo-pharyngeal) part of this muscle sometimes produces a bulge behind the velopharyngeal port.

The middle pharyngeal constrictor radiates widely from the hyoid bone to the posterior pharyngeal wall. This muscle does not only constrict the pharynx but also retract the hyoid bone, playing a role of an antagonist to the protractor muscles of the hyoid bone.

The inferior pharyngeal constrictor narrows the hypopharyngeal cavity at glottal constriction in speech (Esling, 1999), and its cricopharyngeal part is a sphincter of the esophageal entrance, which can be involved at least in part in F0 control (Honda & Fujimura, 1991).

3.2. Internal pharyngeal muscles

The internal pharyngeal muscles are those that descend along the pharynx. They elevate the larynx and shorten the pharynx in addition to their own functions on the attachment points.

The stylopharyngeus passes between the superior and middle constrictor muscles and can be a lateral dilator of the pharynx. The palatopharyngeus forms the palatopharyngeal arch behind the palatine tonsils, narrowing the pharynx at the faucies and lower the velum. The salpingopharyngeus opens the entrance of the auditory tube by its contraction.

3.3. Pharyngeal dilator muscles

The protractors of the tongue base and hyoid bone are dilator muscles of the pharynx: The horizontal part of the genioglossus (GGh) and the geniohyoid (GH) expand the pharynx anteriorly. The stylopharyngeus (SP) is known as the lateral dilator muscle because it runs inferiorly and medially.
from the styloid process and reaches the pharyngeal wall passing between the superior and middle constrictor muscles. While this muscle is convex medially at rest, it straightens at contraction by pulling the pharyngeal wall laterally. The styloglossus (SG) may function similarly because it also runs inferiorly and medially to enter the sides of the tongue base passing closely by the middle constrictor muscle near the palatine tonsil.

Figure 3: Pharyngeal constrictor and internal pharyngeal muscles. The constrictors group into the superior, middle, and inferior. They further subdivide into pt: pterygopharyngeal, bu: buccopharyngeal, my: mylo-pharyngeal, gl: glossopharyngeal, ce: ceratopharyngeal, ch: chondro-pharyngeal, th: thyro-pharyngeal, and cr: crico-pharyngeal parts.

4. Models for voice control by the pharyngeal muscles

Modeling the entire structure of the pharynx could be as exhaustive as modeling the tongue in a physiological articulatory model. Biomechanical models of the pharynx are so far limited to certain functions: a model of the velum and nasopharyngeal wall is proposed by Serrurier and Badin (2008), and models of the other regions of the pharynx are related to swallowing (e.g., Mizunuma, et al., 2009). Thus, modeling a part of the pharynx for a certain elementary function in speech may be more practical. In Figure 4, we propose two conceptual models for regulating elements of voice production that involve the structures in the pharynx: a model for supporting the hyoid bone in F0 control, and a model for deforming the laryngeal cavity in voice quality control.

4.1. A model for F0 control

The protractor and depressor muscles of the hyoid bone are thought to contribute to F0 control via external rotational forces indirectly applied to the cricothyroid joint. The geniohyoid (GH) advances the hyoid bone and lengthens the vocal folds involving the thyroid cartilage (Honda, 1983). The sternohyoid (SH) lowers the larynx and shortens the vocal folds via the cricoid cartilage that rotates along the convexity of the cervical spine (Honda, et al., 1999). A recent MRI study reported a case of pseudo-hypertrophy of the middle constrictor in a soprano singer producing high notes (Honda, 2013). This suggests the three-way support model of the hyoid bone for F0 control as shown in Figure 4(a). The hyoid bone is stabilized during speech or singing for finer F0 control in three directions for the rise and fall of voice F0.

4.2. A model for voice quality control

It has been known that the supraglottic laryngeal cavity forms an independent resonator to add an extra formant to speech sound (Kitamura, et al., 2006; Takemoto, et al., 2006). This resonance occurs at around 3 kHz in male speakers contributing partly to individual speaker characteristics. The frequency of the resonance can be altered voluntarily for intended voice quality change or for signaling the end of sentences. The mechanism for manipulating laryngeal cavity shape may be similar to that observed at swallowing (Donner, et al., 1985). Figure 4(b) indicates a mechanism for narrowing the laryngeal cavity. The laryngeal vestibule is constricted minimally by the three muscles: the thyroarytenoid, aryepiglottic, and thyroepiglottic. The simultaneous contraction of those muscles produced the stricture of the vestibule to emanate laryngeal cavity resonance at various frequencies.

Figure 4: Two models of the pharynx in speech. (a) Control of hyoid bone position for F0 adjustment by three muscles: the geniohyoid, middle constrictor, and sternohyoid. (b) Control of laryngeal cavity stricture for voice quality changes by three muscles: the thyroarytenoid, aryepiglottic, and thyroepiglottic.

5. Summary

This report presented cine-MRI evidence of pharyngeal wall movement, reviewed anatomy of the pharyngeal muscles, and proposed two simple models for voice control involving the pharyngeal muscles. Many questions remain regarding the mechanisms to control the pharynx in speech. The pharynx expands laterally for vowel /i/, while the anatomically granted pharyngeal dilator muscle is only one (the stylopharyngeus), and the real mechanism is still unclear. Whether co-
contraction of the tongue muscles and pharyngeal constrictor muscles for vowel /a/ emanates a mechanical saturation effect to stabilize the quantal vowel is an interesting question to answer. Despite the overwhelming complexity of the pharynx muscles, two simple models proposed in this report can be incorporated into existing physiological models for realistic voice control.

6. Acknowledgements

This work is supported in part by the One-Thousand Talent Program of China (No. WQ20111200010), the National Basic Research Program of China (No. 2013CB329301), and grants from the National Natural Science Foundation of China (General Program No. 61175016, and Key Program No. 61233009).

7. References


A cross-language study of laryngeal-oral coordination across varying prosodic and syllable-structure conditions

Philip Hoole\textsuperscript{1}, Lasse Bombien\textsuperscript{1}

\textsuperscript{1}Institute of Phonetics, Munich University, Germany
hoole\textsubscript{\textsuperscript{1}}bombien@phonetik.uni-muenchen.de

\section*{Abstract}
Transillumination and videofiberendoscopic filming was used to investigate the devoicing gesture in German, Dutch and French for material that compared firstly a strong vs. weak prosodic condition, and secondly singletons vs. clusters (stop + /r/ and /l/). The results showed enhancement of the devoicing gesture in the strong prosodic condition and in the segmental context stop + /r/ for German and French, but not for Dutch. This suggests that the voiceless plosives of French have an active specification for glottal spreading. In terms of timing (e.g. timing of peak glottal opening relative to stop release) French was intermediate between German and Dutch. This indicates that static features are not well suited to capturing cross-language differences in voicing typology and changes in voicing specification over time.

\textbf{Keywords:} laryngeal timing, clusters, prosodic strengthening

\section*{1. Introduction}
This study investigates laryngeal-oral coordination for voiceless consonants in languages that are traditionally considered to differ in terms of their voicing typology, namely German on the one hand vs. French and Dutch on the other. The basic rationale was that prosodic variation can be used as a probe to make clearer the nature of the articulatory representations underlying the voicing distinction. In English, the VOT of voiceless plosives is often found to be longer in prosodically strong locations, suggesting that the laryngeal abduction gesture is also longer and/or larger. More intriguingly, Cho & McQueen (2005) observed for Dutch that prosodic strengthening can lead to shortening of VOT. This was interpreted as indicating that the phonologically voiceless plosives are implemented phonetically as \{+ spread glottis\} in English and \{- spread glottis\} in Dutch (these phonetic features then being reinforced as part of prosodic strengthening). However, neither for English nor for Dutch is anything known about what changes in laryngeal movement actually take place. Further phonological accounts of voicing across languages would also appear to predict differential effects of prosodic strengthening. For example, Iverson & Salmons (2007) analyze voiceless plosives as laryngeally specified for spread glottis in German and English but as laryngeally unspecified for e.g. French and Dutch. It was hypothesized that further light could be shed on this general issue by systematically varying the syllable onsets of the target items, specifically by comparing singletons (e.g. \{p\}) with clusters (e.g. \{pr, pl\}). Even though the second element of these clusters would not normally be regarded as underlyingly voiceless, some evidence has been found that the overall duration of voicelessness is longer in the clusters. To the extent that this is an active mechanism involving the glottal devoicing gesture itself, the question arises as to whether the amount of gestural reorganization in the clusters will be sensitive to the underlying voicing typology of the language involved.

\section*{2. Material}
The target words (part of a larger corpus) all had a voiceless consonant (plosive or fricative) in initial position. Two prosodic conditions were compared: a condition where the target word was in focused position vs. an unaccented condition. The syllable-onset was systematically varied by comparing singleton plosives and fricatives with all combinations with /l/ and /r/ available in the language. A point that will be relevant in the discussion is that the rhotic is dorsal in German and French (ranging from voiced approximant to voiceless fricative, depending on context), and apical in Dutch (for convenience the phonemic symbol /r/ is used throughout). To date five speakers of German, and four each of French and Dutch have been analyzed. In most cases five randomized repetitions of each target item were available for analysis (for one French speaker only two repetitions were completed). Details of the material are given below, with the following schematization of prosodic context:

\begin{tabular}{l}
\textbf{xxx}: target; \textbf{xx}: focus; \textbf{xxx}: contrast
\end{tabular}

\begin{tabular}{ll}
\textbf{German} & Plosive onset: p, t, k, pl, kl, pr, tr, kr \\
Fricative onset: f, fs, fl, fr, fr & Focused: Bis sie \textit{piep} sieht, nicht Tisch. \hspace{1cm} [\"Until she sees \textit{piep} not \textit{table}\"]
\end{tabular}

\begin{tabular}{ll}
Deaccented: Bis sie \textit{piep} sieht, nicht \textit{hört}. \hspace{1cm} [\"Until she sees \textit{piep} instead of \textit{hearing it}\"]
\end{tabular}

\begin{tabular}{ll}
\textbf{Dutch} & Plosive onset: p, t, k, pl, kl, pr, tr, kr \\
Fricative onset: f, fs, fl, fr & Focused: Als \textit{t-ie piep} ziet, niet \textit{laat}.
\end{tabular}

\begin{tabular}{ll}
Deaccented: Als \textit{t-ie piep} \textit{leest}, niet \textit{weet}. \hspace{1cm} [\"It was \textit{pipe} that he quoted (not \textit{quelle})\”]
\end{tabular}

\begin{tabular}{ll}
\textbf{Unaccented:} \textit{Voici des pipes} très étroites \hspace{1cm} [\"Here are some very narrow pipes\”]
\end{tabular}

\section*{3. Methods}
Laryngeal activity was recorded by means of transillumination combined with videofiberendoscopy as detailed in Hoole & Bombien (2014). For the present paper we will concentrate on the following measures:

\begin{enumerate}
\item (1) duration of the oral occlusion of C1; (2) voice onset time (from release of C1 to onset of voicing); (3) relative timing of peak glottal opening; (4) magnitude of peak glottal opening.
\end{enumerate}

Relative timing of peak glottal opening was calculated as the time from the onset of oral occlusion to the time of peak devoicing gesture in the strong prosodic condition and in the segmental context stop + /r/ for German and French, but not for Dutch. This suggests that the voiceless plosives of French have an active specification for glottal spreading.
glottal opening, divided by the duration of oral occlusion of C1.

For an aspirated plosive, where peak glottal opening is roughly synchronous with release of the oral occlusion, the relative timing measure gives a value of about 1. For an unaspirated plosive (and also for fricatives), where peak glottal opening occurs at about the midpoint of the oral occlusion, a value of about 0.5 would be expected. Based on previous work, for clusters relatively later timing of peak glottal opening is expected, e.g. values > 1 if peak glottal opening occurs after the end of C1.

Regarding the magnitude of peak glottal opening, since there is no simple way to calibrate the transilluminant signal, and since signal level can vary quite substantially over the course of the experiment, a normalization factor was calculated separately for each block of repetitions. Specifically, this was based on the average glottal opening over all items with fricative onsets (in each block of repetitions). The motivation for this was that we are particularly interested in laryngeal differences for the plosives across languages, whereas there is no previous reason to expect major language-specific differences in glottal opening magnitude for the fricatives (the aerodynamic constraints on voiceless fricative production should be very similar across languages). Thus, in the absence of an absolute measure, this allows us to express glottal opening for plosives as a proportion of glottal opening for fricatives.

4. Results

We start the presentation of the results with occlusion duration, since this gives a straightforward indication as to whether the attempt to contrast prosodic strength on the target-word has been successful. Fig. 1 (bottom panel) shows the results obtained by subtracting the weaker prosodic condition from the stronger one (broken down by syllable-onset type and language; note that in this and the following figures ‘P’ on the x-axis labels stands for ‘plosive’, not /p/, i.e. syllable onset types are averaged over place of articulation of C1). In all cases the values are positive, indicating that as expected from many previous investigations, occlusion durations are longer under prosodic strengthening. Of the three languages, values are lowest for German (about 10ms), but this still represents a statistically significant difference.

The top panel of Fig. 1 shows the actually measured occlusion durations (again broken down by language and syllable-type), but now averaging over prosodic conditions rather than looking at the prosodically-related differences. We see that values increase from German via French to Dutch. There is a not always very large but nonetheless consistent trend for C1 durations to be shorter in the cluster onsets compared to the singleton onset. The latter effect is certainly not unexpected, but will be relevant for the interpretation of the results for laryngeal-oral coordination.

Turning to VOT (top panel of Fig. 2) the first point to make is that the results are in several respects a mirror-image of the occlusion duration results, i.e. they increase from Dutch via French to German, and also increase going from the singleton to cluster syllable-types. Perhaps the most striking result is that French is actually closer to German than Dutch, which was hardly to be expected from traditional descriptions (values for the French singletons are here in the range that would normally be regarded as aspirated, i.e. about 60ms). Another point to be discussed further below is why the rhotic clusters attract particularly high values specifically for German and French.

The mirror-image pattern between occlusion duration and VOT is also interesting from a cross-language perspective. There may be a cross-language tendency for voiceless consonants to have a rather similar total duration of voicelessness (and rather similar glottal gesture duration). Varying the occlusion duration effectively varies the point in the glottal abduction-adduction cycle at which release of the oral occlusion occurs, thus in turn directly affecting voice onset time (specific information on this below; see also Hutters, 1985; Bombien & Hoole, 2013).

The bottom panel of Fig. 2 shows the difference in VOT across the prosodic condition. The main point here is that the values for all languages cluster quite close to zero. For German, following previous findings in the literature for English, an increase in VOT with prosodic strengthening might have been expected. Even though the differences do go in the expected direction, they only amount to about 5ms, which was not significant. Similarly, we can also not confirm the opposite finding for Dutch of Cho & McQueen that VOT may reduce under prosodic strengthening. On average the differences are indeed negative, but in magnitude are even closer to zero than the German results are.

Having set the scene with the acoustic measures we now turn to direct measurements of laryngeal activity, looking first at the main measure of laryngeal-oral coordination, namely the timing of peak glottal opening relative to the oral occlusion (Fig. 3).

The first main point is that, on the background of the values for occlusion duration and VOT, it was to be suspected that French would show a timing pattern intermediate between Dutch and German. This is indeed very clearly the case. Illustrating this for the singleton stops (leftmost in the top panel of Fig. 3), Dutch shows a value of about 0.5, i.e. peak glottal opening roughly in the middle of the oral occlusion, as expected for an unaspirated stop. For the aspirated stops of German a wide-open glottis at the release of the oral occlusion is indicated by the values of 1 or greater. The intermediate value of about 0.75 for French indicates that, unlike German, the glottis is already closing by the time of oral release, but is still far enough from actual closure to allow for a substantial period of voicelessness after release. A general trend over all languages is that peak glottal opening is timed later in the clusters than the singletons. This is particularly striking for the rhotic clusters of German and French (values well above 1 in both cases), and it will be recalled that it was precisely these items that had the longest VOTs.

What actually leads to the very clear differences in laryngeal-oral relative timing over languages and syllable conditions? As discussed in Hoole & Bombien (2014), glottal gestural duration tends to vary less than, for example, oral occlusion duration. Thus changes in relative timing can in effect fall out from the differences in occlusion duration outlined above. German and French indeed had the shortest occlusions for the rhotic clusters. But evidence was also found for longer gestural durations in the rhotic clusters (not shown here, but discussed in detail for German in Hoole & Bombien, 2014). This indicates that German and French speakers may actively enhance the amount of voicelessness in these clusters. Regardless of the various studies of glottal timing differences related to the prosodic condition (see bottom panel of Fig. 3): The difference values are generally lower for all conditions and languages, in other words peak glottal opening is timed slightly earlier in the stronger prosodic condition. This in turn is an indication that the glottal gestural duration does not lengthen as much as the occlusion duration does, and is also a further indication that German speakers are not aiming for an active lengthening of VOT in the prosodically strong
condition. If they were, they would need to ensure that the glottal gesture is lengthened sufficiently to keep at least the same and preferably later relative timing of peak glottal opening as the occlusion lengths.

Results for the magnitude of peak glottal opening are shown in Fig. 4 (top panel). Clearly, French once again occupies an intermediate position between German and Dutch. Recall that a value of 1 indicates a comparable glottal opening to the fricatives. For German (aspirated), the singletons are only slightly below this value, whereas for Dutch (unaspirated), they are well below 0.5. The other main point of interest is that German and French have in common a particularly large glottal opening in the rhotic clusters. Taken together with the timing measurements this indicates that speakers are actively aiming for substantial glottal opening over a substantial part of the rhotic segment, i.e. they are aiming to ensure its realization as a clear voiceless fricative (amplitude close to 1). As argued in Hoole & Bombien (2014) there may thus be a more active pattern of reorganization in the rhotic compared to the lateral clusters: devoicing of the lateral may be a simple passive coarticulatory effect, falling out from the proximity to the devoicing gesture of the voiceless plosive.

The final result concerns prosodically-related differences in peak glottal opening, clearly a key area given our initial hypotheses. Strikingly, French patterns together with German, rather than lying between German and Dutch: German and French both show a clear increase in movement amplitude in the prosodically stronger condition, whereas the change for Dutch is absolutely negligible (Fig. 4, bottom panel).

In fact, it is not quite clear why German and French speakers increase the magnitude of glottal opening. Even if the finding is not unexpected, at least for German, given the many articulatory correlates of prosody that have been found (more background in Hoole & Bombien, 2014), the greater glottal opening cannot be part of a set of adjustments to increase the duration of voicelessness since we observed above that VOT was only weakly influenced by prosody. Future work will need to look in detail at the acoustic properties of the burst and aspiration phase of the plosives for prosodically related differences that could be useful to the listener in recovering the prosodic structure of the utterances.

5. Discussion

The results for German conformed to expectations in that the magnitude of the laryngeal abduction gesture increased under prosodic strengthening. Interestingly, this was the case for French as well. In fact, many of the French voiceless plosives in our material would be regarded as aspirated, with peak glottal opening usually located well into the second half of the oral occlusion. This suggests that French voiceless plosives have an active specification for glottal spreading, with a timing pattern that is still different from German but nonetheless results in substantially positive VOT values. This active glottal spreading can then be targeted by the phonetic reinforcement processes forming part of prosodic strengthening. The results for Dutch were different, since there was no tendency towards an increase in the magnitude of the glottal abductory movement in strong prosodic contexts. Accordingly, while French and Dutch are traditionally regarded as typologically similar with regard to the voicing distinction, the articulatory representation of the voiceless consonants may well have started to diverge.

This interpretation is confirmed by the syllable-structure condition in the corpus. For German, clusters with /t/ (e.g. /pr, tr, kr/) typically showed a longer and/or larger glottal gesture compared to the corresponding singleton aspirated plosives (confirming Jessen, 1999). At first sight this is an unexpected result since /t/ would not normally be regarded as having an active specification for glottal abduction (and so a combination of e.g. /p/ + /t/ should not result in two smaller glottal gestures blending into one larger one). However, this may be quite a natural process in cases where /t/ is realized with a dorsal constriction, which gives conditions that are very unfavourable for voicing at the release of the plosive (see Hoole & Bombien, 2014). Thus speakers reinforce the tendency towards voicelessness by enhancing the glottal abduction already present for the plosive itself. Once again, the more striking result was that a very similar pattern occurred for French. The presence of a strong glottal abduction gesture in these clusters in French is mysterious if the voiceless plosives are assumed to be laryngeally unspecified or even specified as [-spread glottis] at the oral release, since such a representation would hardly predict an enhancement of the glottal abductory movement in specific contextual conditions. But the problem dissolves if French is assumed to represent voiceless plosives in terms of an active glottal abductory movement that is simply timed somewhat differently from German and English. (The results here for French also have an interesting parallel to recent work of Beckman et al. (2011) on Swedish, which indicated that representation of voicing in terms of a single privative feature may not be appropriate for all languages.)

Assuming a representation in terms of coordination patterns, rather than in terms of discrete atemporal features (cf. Löfqvist & Yoshioka, 1981), also gives a much more natural account of how languages may diverge as a result of subtle shifts in intergestural timing.

6. Acknowledgements

Work supported by DFG Grant HO3271/3-2 in SPP1234 to P. Hoole. Thanks to René Bombien and René Schramm for performing the endoscopy, and to Elizabeth Heller for help with segmentation and kinematic analysis.

7. References


Figure 1. Duration of stop occlusion phase. Top panel: averaged over prosodic conditions and subjects. Bottom panel: Difference between prosodic conditions (strong-weak) averaged over subjects.

Figure 2. Duration of voice onset time. Other details as for Fig. 1

Figure 3. Relative position of peak glottal opening in stop occlusion phase. Other details as for Fig. 1

Figure 4. Magnitude of peak glottal opening, normalized by glottal opening in fricatives. Other details as for Fig. 1
Simulating a state feedback model of speaking
John F. Houde¹, Caroline Niziolek¹, Naomi Kort², Zarinah Agnew¹, and Srikantan S. Nagarajan²

¹Dept. of Otolaryngology – Head and Neck Surgery, University of California San Francisco
²Dept. of Radiology & Biomedical Imaging, University of California San Francisco
houde@phy.ucsf.edu, cniziolek@ohns.ucsf.edu, naomi.kort@ucsf.edu, zarinah.agnew@ucsf.edu, sri@ucsf.edu

Abstract
An important part of understanding the neural control of speaking is determining how sensory feedback is processed. The role of sensory feedback in speaking suggests a paradox: it need not be present for intelligible speech production, but if it is present, it needs to be correct or speech output will be affected. For this reason, current models of speech motor control relegate sensory feedback to a more indirect role, with an inner feedback loop within the CNS that directly controls speech output, and a slower outer feedback loop where the possibly delayed and intermittent sensory feedback updates the internal feedback loop. Such models can be described as variations on a more general theory of state feedback control (SFC). Here we show, via numerical simulations, how the SFC model can account not only for what is known about the behavioral role of sensory feedback in speaking, but also many of our recent findings about neural responses to auditory feedback.

Keywords: speech motor control, sensory feedback, numerical simulation

1. Introduction
The paradoxical role of sensory feedback in speaking is that it is not necessary for intelligible production, but if it is present, it needs to be correct or speech will be affected (Houde & Nagarajan, 2011). For these reasons, current models of speech motor control relegate sensory feedback to a more indirect role, with an inner feedback loop within the CNS that directly controls speech output, and actual sensory feedback (both auditory and somatosensory feedback) forming slower, possibly delayed and intermittent, external loops that update the internal feedback loop (Guenther & Vladusich, 2012; Price, Crinion, & Macsweeney, 2011; Tian & Poeppel, 2010). Such models can be described as variations on the general theory of state feedback control (SFC), developed in the domain of modern control engineering theory (Houde & Nagarajan, 2011). SFC models have become more prevalent in many domains of motor control research, and we have previously described the hypothesized applicability of SFC to modeling speech motor control (Houde & Nagarajan, 2011).

Here, we construct a numerical simulation of an SFC model and show how it accounts not only for behavioral phenomena associated with the roles of sensory feedback in speaking, but also several of our past experimental findings concerning the neural phenomena involved in auditory feedback processing during phonation.

2. Structure of the SFC model
To facilitate comparison with our experiments, we illustrate our model by focusing on the production of pitch, although our model can easily be generalized to other aspects of speech production.

In this model, production of an utterance involving phonation begins in the CNS with higher frontal cortex (IFG) activating a phonation control network (blue arrow in Figure 1). This network controls phonation via state feedback control (SFC): During phonation, vPMC maintains a running estimate of the current dynamic state of the larynx (i.e., the estimated laryngeal state; orange in Figure 1); this state carries information about current sub-glottal pressure, vocal fold position, tension, and any other parameter the network has learned is important to monitor for achieving phonation. M1 generates laryngeal controls based on this state estimate, using a state feedback control law (state fb ctrl law in Figure 1) that keeps the larynx tracking a desired state (e.g., one that maintains a desired pitch). While the larynx responds to these controls, vPMC uses a copy of these controls (“reference copy”) to predict the next laryngeal state. It feeds this prediction forward (green arrows) to the higher sensory areas (SII/IPL in somatosensory feedback and vSMG/pSTS for auditory feedback) that use it to predict sensory feedback. At the primary sensory cortices (S1 in somatosensory cortex and A1 in auditory cortex), feedback from the larynx is compared with the feedback predictions, resulting feedback prediction errors. The higher sensory areas convert these feedback predictions into state corrections (state corr somato, state corr auditory in Figure 1) that are fed back (red arrows) to vPMC and added to the original state prediction, resulting in a refined estimate of the next laryngeal state (orange). This in turn is fed back to M1 to generate the next laryngeal controls.

Figure 1: SFC model of how the CNS controls phonation. See text for description.
3. A simulation of an SFC model of pitch control

To verify and illustrate the claims we have made about the SFC model, we have developed a simulation of SFC-based control of speech. For simplicity, the model controls a one-dimensional “speech output” which we have likened to pitch. However, it is straightforward to extend the simulation to control more realistic, multi-dimensional speech output (e.g., loudness, pitch, formants, friction). The simulation was implemented in Matlab (The Mathworks, Inc., Natick, MA), and consists of two parts: the “larynx” and the “phonation control network”.

3.1. The “larynx”

The “larynx” to be controlled is modeled as a single damped spring-mass system with a variable rest length of the spring. Position of the mass of this system is taken to be the current pitch output of the vocal tract. This model is based on the idealization (admittedly incomplete) that vocal fold length (i.e., position of the mass) determines pitch, and that the muscles controlling vocal fold length (e.g. the cricothyroid muscle (Titze, Jiang, & Drucker, 1988)) can be modeled as damped spring-mass systems with variable spring rest length (Hill, 1925). This model is not intended to simulate the rich range of laryngeal behaviors captured in multidimensional models (e.g. (Story & Titze, 1995)), but rather to act as a system with dynamics that the controller (i.e., the “phonation control network”) must contend with to control pitch. Rest length of the muscle controlling pitch is, in turn, controlled by “brainstem/spinal cord” lower motor system that in turn integrates descending cortical control into a constantly updated rest length of the muscle (Shalit, Zinger, Joshua, & Prut, 2012). In this way, the simulation assumes that cortical motor output codes only desired changes in the current pitch output.

Based on the findings of prior motor control studies, the descending cortical control signal is also subject to “signal dependent noise” (Harris & Wolpert, 1998). This means that the control signal actually seen by the lower motor system is the cortical motor output plus noise that scales with the magnitude of the cortical motor output.

This simulated “larynx” produces two types of sensory “feedback”. First, an “auditory” output is idealized as linear conversion to Hz from muscle position to a reasonable value for a speaking pitch, and is assumed to be corrupted by additive white Gaussian “observation” noise that has a feedback delay of 150 ms (Houde & Nagarajan, 2011). Second, a “somatosensory” output is included, reflecting the current position of the mass of the muscle spring/mass system also corrupted by white Gaussian noise, with a feedback delay of 15 ms.

3.2. The “phonation control network”

The simulated “phonation control network” for controlling pitch is made up of two parts: (1) an observer: a system that estimates the current state of the larynx and (2) a state feedback control law that uses the state estimate to generate controls of the vocal tract. Most of the phonation control network is engaged in implementing the observer via interaction of feedforward predictions and feedback corrections (i.e., the green and red arrows in Figure 1). Here, we focus on the observer system. Although in principle, we could use optimal control theory to implement the state feedback control law where the forward model is first used to estimate the current, undelayed auditory output of the larynx from the current state estimate. This estimated current output is then compared with the current desired output, with the difference passed through a control gain G to generate control applied to the laryngeal simulation on the next time step.

3.2.1. The Kalman Filter Based Observer

The observer estimates state via a recurrent prediction-correction process where a prediction of next state is used to generate sensory predictions that are compared with incoming feedback. The resulting feedback prediction errors are converted by observer gains (state corr somato, state corr auditory in Figure 1) into corrections of the state prediction. When these gains are computed optimally (i.e., based on the noise characteristics of the sensory feedback), the observer is referred to as a Kalman filter (Houde & Nagarajan, 2011). We therefore implemented the observer as a Kalman filter, reflecting the assumption that the CNS would also seek optimal values for the observer gains.

The heart of the Kalman filter observer simulation is a forward predictive model of the current state of the lower motor/muscle “vocal tract”. For the simulations, the forward model is simply a copy of the parameters of the state space model used to simulate the vocal tract. Our simulations here do not include the process of learning all parameters of the forward model, and instead concentrate on how the speech motor system behaves after some parameters of the forward model have been learned. In particular, the model estimates the feedback delay and the covariances of the state and observation noise that determine the Kalman gain, by cross-correlating auditory feedback with somatosensory feedback.

The predicted feedback output of the forward model is delayed by the estimated feedback delay before being compared with the incoming feedback. The resulting delayed feedback prediction error is multiplied by a Kalman gain function to compute a correction to the state prediction of the forward model. We approximated this gain by first computing the steady-state Kalman gain assuming zero feedback delay (Houde & Nagarajan, 2011), then computing the effect of this gain after it has been propagated through the forward model N time steps, where N is the estimated feedback delay. Ultimately, this way of calculating the Kalman gain quantifies the intuition that a feedback prediction error from N time steps in the past (the feedback delay) becomes less and less informative about the current laryngeal state as N increases.

4. Results

We simulated two different auditory feedback experiments we have previously conducted. Both of these experiments contrast responses to auditory feedback from the subject’s ongoing speech (the speaking condition) with passive re-listening to playback of auditory feedback from the speaking condition (the listening condition). To facilitate comparison of simulation results with experimental data, our simulation includes a simplified model of evoked response potential (ERP) generation (David, Harrison, & Friston, 2005).

4.1. Speech Onset

Figure 2 shows one trial from a simulated 100-trial speech onset experiment. The top panel shows the behavioral outputs of the simulation, showing that onset of a 120 Hz phonation target (y) results in a pitch output (y) that initially slightly deviates from (undershoots) the target. This undershoot is due
to signal-dependent noise added to the large initial laryngeal control.

Figure 2: single trial from simulated speech onset experiment. 1st (top) panel: yt: target pitch, y: output pitch, ydel: auditory feedback received by phonation control network, with 150ms auditory processing delay. 2nd panel: deviation of this single trial from the median across trials, showing initial undershoot and subsequent “centering”. 3rd panel: feedback prediction error (ye) in the speak and listen conditions of this trial. 4th panel: ERPs of the feedback prediction errors (ERP(ye)), showing SIS as the ERP difference between the speak and listen conditions.

Figure 3: The SIS falloff effect. Scatterplot of the relation between SIS (y-axis) and initial deviation from median output pitch in the simulated speech onset experiment

This undershoot generates both auditory and somatosensory feedback prediction errors that, via their Kalman gains, generate state estimate corrections resulting in small, corrective laryngeal controls that counteract the initial undershoot as the utterance continues. The 2nd panel shows how this initial undershoot and subsequent correction can also be seen by measuring how much pitch output for a single trial deviates from the median pitch output over all trials. Such analysis avoids explicit reference to the pitch target and duplicates similar analysis done in our experimental studies (Niziolek, Nagarajan, & Houde, 2013), where we refer to the subsequent correction as “centering”. The 3rd panel shows auditory feedback prediction errors (ye) for both the speaking condition (red), compared to the listening condition (blue), while the 4th panel shows the simulated ERPs corresponding to these prediction errors. The prediction error in the listen condition is very large because the speech onset is not predicted when passively listening to an external speech source, whereas the prediction error in the speak condition is smaller because a speaker is able to predict his/her own speech onset, via efference copy of his/her own laryngeal controls. Thus, the ERP in the speak condition is smaller than the ERP in the listen condition. This replicates the speaking-induced suppression (SIS) effect we commonly see in speech onset experiments (Kort, Nagarajan, & Houde, 2014).

In the speak condition, the size of the ERP is related to the unpredicted deviation of auditory feedback (ydel) from the target pitch. Figure 3 shows in a scatterplot across all trials that this means that initial deviation from the across-trial median (a surrogate for the target pitch) is closely related to the size of the speak – listen ERP (SIS) difference, which replicates the “SIS falloff” effect we have recently documented experimentally (Niziolek et al., 2013).

4.2. Speech Feedback Perturbation

Figure 4 shows one trial from a simulated 100-trial auditory feedback perturbation experiment. The top panel shows the behavioral outputs of the simulation, showing (in light blue) the auditory feedback, perturbed for 400 msec by a 100 cent (one semitone) shift down in pitch, and (in green) the effect this has on output pitch: in this trial, it induces 65% compensation.

Several factors influence how much compensation will be expressed on each trial. First, compensation for the auditory feedback perturbation is moderated by conflicting information conveyed by somatosensory feedback, which remains unaltered. The next panels of Figure 4 show these conflicting influences on compensation. The 2nd panel shows that the auditory feedback perturbation creates an auditory feedback prediction error that is then reduced by the compensatory response. But the 3rd panel shows the compensatory response itself then creates a somatosensory feedback prediction error that results
an opposing influence on the realized compensation. The strength of this opposing influence is regulated by the Kalman gain on somatosensory feedback prediction errors, which in turn is determined by the estimated somatosensory observation noise – i.e., the estimated reliability of somatosensory feedback. In the simulation shown here in the plots, mean compensation was 31.6%, but if we increase somatosensory noise from 0.005 to 0.05, mean compensation rises to 36.9% due to the decreased reliability (i.e., the "numbing") of somatosensory feedback. This is consistent with prior findings about the effect of numbing somatosensory compensation for pitch feedback perturbations (Larson, Altman, Liu, & Hain, 2008).

Other factors cause measured compensation to vary from trial to trial around the mean. State noise and observation (sensory feedback) noise cause feedback to fluctuate over the course of the simulation, which directly affects the measurement of peak compensation used to gauge compensation on individual trials. But the fluctuating feedback also indirectly affects compensation, because we hypothesize that the Kalman gain on sensory feedback is continually re-estimated from current sensory feedback over the course of the experiment. This re-estimation causes the Kalman gain to vary slightly from trial to trial, and since the size of the Kalman gain determines magnitude of compensation on each trial, compensation therefore fluctuates because of this.

Evidence that fluctuation in the Kalman gain contributes to compensation variability in real experiments comes from consideration of additional outputs of the simulation. The 4th panel of Figure 4 shows ERPs generated from the auditory feedback prediction errors in the speak (red) and listen (blue) conditions of the experiment. At speech onset, these two responses differ greatly, demonstrating the SIS effect, but here the two responses are identical, since unlike speech onset, the externally-applied pitch perturbation is equally unexpected in both the speak and listen conditions. The 5th panel, however, shows that these equal feedback prediction errors nevertheless result in unequal state estimate corrections. The panel shows ERPs generated from state estimate corrections in the speak (red) and listen (blue) conditions, with larger ERPs (corresponding to larger state corrections) in the speak condition. This speech perturbation response enhancement (SPRE) matches that seen in our prior studies (Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Kort et al., 2014), and in the simulations is the result of the Kalman gain on auditory feedback being larger in the speak condition than in the listen condition (because of the inability to ascribe the total variance to anything but observation noise).

\[ R(0.050), p(0.6210) \]
\[ R(0.864), p(0.0000) \]

**Figure 5: Correlation with compensation.** Scatterplots comparing regression of percent compensation with ERPs from feedback prediction errors (ye) (left) which exhibit only SIS, and ERPs from state corrections (xe(1)) (right) which exhibit SPRE.

SPRE, therefore, is due to action of the Kalman gain on feedback prediction errors, and so activity in the parts of the model expressing SPRE will reflect trial-to-trial variability in the Kalman gain not seen in the feedback prediction errors. Thus, since Kalman gain determines compensation magnitude, ERPs from SPRE-expressing parts of the model (i.e., ERPs from state corrections) will be more correlated with trial-to-trial variation in compensation than ERPs from model components expressing only SIS (i.e., ERPs from feedback prediction errors). Figure 5 shows that this is the case in our simulations, which matches what we have found in previous pitch feedback perturbation experiments (Chang et al., 2013).

5. Conclusions

The concept of state feedback control (SFC) is a powerful and flexible model of motor control, and many current models of speech motor control can be described as examples of SFC. Here, we have considered an SFC model of speech motor control with a very general form, and found it can account for many of the known characteristics of the role of auditory feedback in the control of speech, as well as many of the phenomena observed in our previous studies of the neural processing of auditory feedback.

6. Acknowledgements

Support provided by NSF grant BCS-1262297 and NIH grant R01-DC010145.

7. References


An EMA Examination of the Czech Alveolar and Post-Alveolar Fricatives
Phil Howson and Alexei Kochetov
The University of Toronto, Department of Linguistics, Canada
phil.howson@mail.utoronto.ca, al.kochetov@utoronto.ca

Abstract
This paper examines the alveolar and post-alveolar fricatives in Czech using articulography. The placement of sensors along both the mid-sagittal plane and the coronal plane allowed for data collection of both the mid-sagittal contour and the cross-sectional morphology of the alveolar and post-alveolar fricatives. Six native speakers of Czech read nonce words in three environments: word-initial, intervocalic and word-final. The results indicate that the primary difference between the alveolar and post-alveolar fricatives on the mid-sagittal plane is the absence or presence of the tongue body raising, respectively. The raising of the tongue body for the post-alveolars creates a longer channel, further back in the mouth. The results also indicate that the post-alveolars are produced with a laminar articulation. The cross-sectional morphology of the alveolars is deeply concave, while the post-alveolars are slightly concave. This is contrary to previous findings on post-alveolars in other languages which have a flat or domed cross-sectional morphology.

Keywords: Czech, fricatives, EMA, grooving

1. Introduction
Alveolar and post-alveolar sibilant fricatives are produced with a tight constriction which generates channel turbulence (Fant, 1960; Stevens, 1971). The emerging jet stream then impinges on the incisors, generating wake turbulence (Catford, 1977). This plays a major role in generating the high frequency associated with sibilants. The acoustic characteristics of the different fricatives are also dependent on the length of the front cavity (Hughes and Halle, 1956).

The cross-sectional morphology and groove length has also been shown to be an important characteristic for sibilant production (Shadle, 1985). The alveolar fricatives are typically characterized by a concave cross-sectional morphology. In contrast, the post-alveolars have been shown to have a flat or slightly convex cross-sectional morphology and raising of the tongue body, creating a small degree of palatalization (Catford, 1977; Narayanan, Alwan, and Harker, 1995; Ladefoged and Maddieson, 1996; Dixit and Hoffman, 2004).

Linguistic characteristics in consonant articulations have, however, been shown to be wide across different languages (Marchal, Farnetani, Hardcastle, and Butcher 1988). Therefore, the aim of this study is to compare the linguistic characteristics and cross-sectional morphology of the alveolar and post-alveolar consonants in Czech through the use of electromagnetic articulography (EMA).

2. Method
2.1. Participants
Four female natives speakers and one male native speaker of Czech participated in this study. The speakers’ age ranged from 18-29, with a mean age of 26. They were recruited from the Czech community in Toronto, Ontario.

2.2. Stimuli
The target stimuli were nonce words with four target phonemes, /s/, /z/, /ʃ/, /ʒ/, in the word-initial (#_a; e.g. sap), intervocalic (a_a; e.g. bafajap), and word-final contexts (a_; e.g. paz). Distracter stimuli were also presented. The stimuli were randomized and presented using PowerPoint. The participants repeated each set 15 times, for a total of 45 productions of each token. Each set contained the same order.

2.3. Instrumentation
Articulatory data was recorded with an AG500 electromagnetic articulograph (EMA; Carstens Medizinelektronik, GmbH) at the Oral Dynamics Lab, Speech Language Pathology Department, University of Toronto. The AG system has 12 channels which record horizontal (x), lateral (y), and vertical displacement (z) for sensors attach to points of interest. Data was recorded with a sampling rate of 200 Hz, and audio data with a sampling rate of 16 kHz.

2.4. Procedure and Analysis
The participants had 5 EMA sensors glued to the tongue with periAcryl blue glue. The 5 sensors had the following configuration: tongue tip (TT), approximately 1 cm from the tongue tip; tongue body (TB), approximately 3 cm from the tongue tip; tongue dorsum (TD), approximately 4 cm from the tongue tip; tongue right (TR) and tongue left (TL), approximately 1 cm from the side of the tongue and 1 cm from the center sensor; TT, TB and TD were along the mid-sagittal plane (Figure 1).

Figure 1: Sensor Configuration
A microphone was placed approximately 6 inches in front of the participants and a computer monitor with the PowerPoint presentation approximately 2 feet in front of the participant. The participants first read Aesop’s *The North Wind and Sun* in Czech (from Dankovičová, 1999) twice to habituate themselves to the EMA sensors.

Raw articulatory data was corrected for head movement by rotating and transposing the data to a reference position, such that the x-y plane coincided with the subject’s occlusal bite-plane at a bubble level (Westbury, 1995). Data was filtered with an 11-point Butterworth filter with a cut-off frequency of 35 Hz to remove high-frequency noise. In order to extract the
Grooving data for each consonant, the boundaries of each were determined by examining the spectrogram of the acoustic data, setting the onset at the end of the preceding vowel and the offset at the beginning of the following vowel. The presence of clear glottal pulses (or lack thereof) were taken to indicate the onset or offset of the vowel. Tongue grooving was operationalized as the angle $\gamma$ between two imaginary lines extended from the TR and TL sensors to the TB sensor (Figure 2). A low angle ($\gamma$) represents a more grooved cross-sectional morphology, while a higher angle ($\gamma$) represents a flatter cross-sectional morphology. 180 degrees would represent a completely flat contour. The average degree of grooving for each consonant was used for comparison using the ez package (Lawrence, 2011) in R (2008), which allows for repeated measures ANOVA to be run easily in R. It also performs Mauchly’s Test for Sphericity and the Greenhouse-Geisser Correction for factors with more than two levels. If Mauchly’s Test was significant, the corrected p-values were used. The ANOVA had factors Consonant (4 levels: $s$, $z$, $ʃ$, $ʒ$) and Environment (3 levels: #_a, a_a, a_#). Post-hoc pairwise t-test with Bonferonni correction were performed.

### 3. Results

In the first section, the results of the grooving analysis will be presented. In the second section, the results of the gestural analysis will be presented.

#### 3.1. Grooving Results

Recall that a low angle ($\gamma$) correlates to a more concave cross-sectional morphology and a high angle ($\gamma$) correlates to a more convex cross-sectional morphology. 180 degrees would indicate a flat cross-sectional morphology.

The repeated measures ANOVA revealed a main effect of Consonant [$F(3,12) = 10.37$, $p = 0.0309$] (Figure 3), but no main effect of Environment [$F(2,8) = 0.01$, $p = 0.9914$]. An interaction between the factors Consonant and Environment [$F(6,24) = 4.16$, $p = 0.0053$] was also revealed (Figure 4). Post-hoc tests revealed a significant difference between /s/ and /ʃ/ ($p < 0.0001$) and /ʒ/ ($p < 0.0001$). There was also a significant difference between /z/ both /ʃ/ ($p < 0.0001$) and /ʒ/ ($p < 0.0001$). The angle of /s/, /z/ was approximately 26 degrees lower than /ʃ/, /ʒ/, indicating a more concave cross-sectional morphology for the alveolars. There was no significant difference between /s/ and /z/ ($p = 1$) or between /ʃ/ and /ʒ/ ($p = 1$).

#### 3.2. Gestural Results

##### 3.2.1. X-axis (horizontal plane) Results

The examination of the TT sensor's horizontal movement revealed a main effect of Consonant [$F(3,12) = 3.65$, $p = 0.0443$] (Figure 5), but no main effect of Environment [$F(2,8) = 0.74$, $p = 0.5067$]. There was no interaction between the factors Consonant and Environment [$F(6,24) = 0.49$, $p = 0.8122$]. Post-hoc tests revealed a significant difference between /s/ and /ʃ/ ($p = 0.0018$) and /ʒ/ ($p < 0.0001$). /z/ was significantly different from /ʃ/, /ʒ/ ($p < 0.0001$). /s/, /z/ was approximately 2 mm further forward than /ʃ/, /ʒ/. There was no significant difference between /s/ and /z/ ($p = 1$) or between /ʃ/ and /ʒ/ ($p = 1$).
The examination of the TB sensor's horizontal movement revealed a marginally significant main effect of Consonant [F(3,12) = 3.40, p = 0.0536], but no main effect of Environment [F(2,8) = 0.20, p = 0.8209]. There was no interaction between factors Consonant and Environment [F(6,24) = 1.28, p = 0.3027]. The post-hoc tests revealed no significant difference for the factor Consonant (p = 1). The mean of /s/ was 82.75 mm, /z/, 83.02 mm, /ʃ/, 81.97 mm, and /ʒ/, 81.53 mm, suggesting that the tongue body may be slightly further forward for the alveolar fricatives.

The examination of the TD sensor's horizontal movement revealed a main effect of Consonant [F(3,12) = 16.38, p = 0.0002] (Figure 6), no main effect of Environment [F(2,8) = 0.32, p = 0.3278] and no interaction between the factors Consonant and Environment [F(6,24) = 1.01, p = 0.4371]. Post-hoc comparisons revealed a significant difference between /s/ and /ʃ/ (p = 0.0544), but not /ʃ/ (p = 0.1421). /z/ was significantly different from both /ʃ/ (p = 0.0002) and /ʒ/ (p < 0.0001). /s/, /z/ was further forward in the mouth than both /ʃ/, /ʒ/ by approximately 2 mm. There was no significant difference between /s/ and /ʃ/ (p = 0.3311) or between /ʃ/ and /ʒ/ (p = 1).

3.2.2. Z-axis (vertical plane) Results

The examination of the TT sensor's vertical movement revealed no main effect of Consonant [F(3,12) = 2.79, p = 0.0859], no main effect of Environment [F(2,8) = 0.38, p = 0.6984] and no interaction between the factors Consonant and Environment [F(6,24) = 1.74, p = 0.1550].

The examination of the TD sensor's vertical movement revealed a main effect of Consonant [F(3,12) = 45.85, p < 0.0001] (Figure 7), but no main effect of Environment [F(2,8) = 0.49, p = 0.6299]. There was no interaction between factors Consonant and Environment [F(6,24) = 0.47, p = 0.8272]. Post-hoc tests revealed a significant difference between /s/ and both /ʃ/ (p < 0.0001) and /ʒ/ (p < 0.0001). There was also a significant difference between /z/ and both /ʃ/ (p < 0.0001) and /ʒ/ (p < 0.0001). /s/, /z/ was also approximately 9 mm lower than /ʃ/, /ʒ/. There was no significant difference between /s/ and /ʃ/ (p = 1) or /ʃ/ and /ʒ/ (p = 1).

Figure 7: The tongue body (TB) on the z-axis (vertical plane).

The examination of the TD sensor's vertical movement revealed a main effect of Consonant [F(3,12) = 24.09, p < 0.0001] (Figure 8), but no main effect of Environment [F(2,8) = 0.76, p = 0.4964] and no interaction between the factors Consonant and Environment [F(6,24) = 0.78, p = 0.5936]. The post-hoc tests revealed a significant difference between /s/ and both /ʃ/ (p < 0.0001) and /ʒ/ (p < 0.0001). There was also a significant difference between /z/ and both /ʃ/ (p < 0.0001) and /ʒ/ (p < 0.0001). /s/, /z/ was also approximately 8 mm lower than /ʃ/ and 7 mm lower than /ʒ/. There was no significant difference between /s/ and /ʃ/ (p = 1) or /ʃ/ and /ʒ/ (p = 1).

Figure 8: The tongue dorsum (TD) sensor on the z-axis (vertical plane).

4. Discussion

The grooving analysis revealed a deep cross-sectional channel for the alveolar fricatives in Czech, which has been described as a characteristic feature of alveolar fricatives in general (Ladefoged and Maddieson, 1996; Narayanan et al., 1995). However, the post-alveolar fricatives in Czech exhibited a somewhat unexpected characteristic given previous studies of fricatives: the tongue body was not found to be flat/domed as some previous analysis have shown to be the case, at least for English (Ladefoged and Maddieson, 1996; Narayanan et al., 1995). The grooving channel may possible act as a funnel, directing the airflow to the constriction location. It is possible that this is a more general mechanism of fricative production, as Narayanan et al. (1995) found that even post-alveolar fricatives in English exhibit some degree of grooving in the posterior portion of the tongue.

The gestural analysis revealed there was a significant difference between the height of the tongue body and dorsum and the frontness of the tongue tip and dorsum when comparing the alveolar and post-alveolar fricatives. The alveolars are relatively flat, produced with a tongue body and dorsum much lower than the post-alveolars. The Czech post-alveolars have a tongue body which is much more raised and a
slightly further back tongue tip. This suggests a more laminal articulation and a longer groove. The groove length likely contributes to the acoustic differences between the alveolar and post-alveolar fricatives, as groove length has been shown to be critical aspect in sibilant production (Shadle, 1985). The length of the front channel has also been shown to have a significant effect on the different acoustic characteristics of fricatives (Hughes and Halle, 1956). This conclusion is supported here because the alveolar and post-alveolar fricatives both have different front channel lengths. The groove length, front cavity and place of articulation appear to be the primary features distinguishing the alveolar and post-alveolar fricatives.

5. Conclusion

In this paper, an analysis of the cross-sectional morphology and gestural characteristics of the alveolar (s, z) and post-alveolar (ʃ, ʒ) in Czech were presented. The position of the tongue sensors allowed for simultaneous data collection of the mid-sagittal and coronal plane. The angle given during the analysis of the cross-sectional morphology was taken to reflect the degree of grooving.

The alveolar fricatives, /s, z/, were found to exhibit a deep groove, while the post-alveolars, /ʃ, ʒ/, were found to exhibit a slight groove. This is contrary to some examinations of English which have shown a flat/domed tongue body for post-alveolars. The gestural analysis showed a primary difference in the height of the tongue body, which was significantly higher for the post-alveolars, /ʃ, ʒ/. The causes the post-alveolars to have a small degree of palatalization. This is likely the primary contrast between the two fricatives because it creates a longer groove length and larger front cavity for the post-alveolar fricatives. The findings further support studies which have shown groove length and place to be important in the production of fricatives.

6. Acknowledgements

The authors would like to thank Dr. Pascal van Lieshout, Chris Neufeld and James Le for their assistance and support with the EMA data collection and analysis. This work was supported by a SSHRC grant to the second author.

7. References


Neural correlates of temporal coding of somatosensory-auditory interactions in speech

Takayuki Ito\textsuperscript{1}, Vincent L. Gracco\textsuperscript{2,1}, David J. Ostry\textsuperscript{2,1}

\textsuperscript{1}Haskins Laboratories, New Haven, CT, USA
\textsuperscript{2}McGill University, Montréal, QC, Canada

\texttt{taka@haskins.yale.edu, vincent.gracco@mcgill.ca, david.ostry@mcgill.ca}

Abstract

The temporal relationship between speech movements and sounds is presumably the primary source of the somatosensory-auditory interaction that occurs in perceptual processing of speech. We assessed to what extent sensory evoked potentials reflect somatosensory-auditory interaction during speech perception in terms of relative timing differences between the onsets of the two sensory stimuli. We recorded event-related potentials in response to somatosensory stimulation alone, auditory stimulation alone, and paired somatosensory and auditory stimulation. In the paired condition, the timing of the two stimuli was either simultaneous or offset by 90 ms. We found the magnitude of the event-related potential difference varied as a function of the relative timing of stimulation. Changes in the event-related response were reliably correlated with participants’ judgments of speech sounds. The results show neural correlates of the temporal coding of somatosensory-auditory interactions in speech and suggest the interaction process is dependent on the specific temporal order of sensory inputs in speech production.

Keywords: speech production, speech perception, ERP, orofacial skin sensation.

1. Introduction

Multiple sensory inputs seamlessly interact in the process of speech perception and production. Recent findings of an orofacial somatosensory influence on the perception of speech sounds suggest a crucial role for the somatosensory system in the linkage between speech production and perception system. Whereas psychophysics experiments have examined somatosensory-auditory interactions in behavioral terms (Fowler and Dekle, 1991; Gick and Derrick, 2009; Ito et al., 2009), the neural correlates of orofacial somatosensory cortical processing in the perception of speech sounds are not yet understood.

The interaction between sensory inputs can vary depending on the context of sensory events (Stein and Meredith, 1993). The temporal relationship between multiple sensory inputs is one important factor for the tuning of multi-sensory interactions (Vroomen and Keetels, 2010). At a behavioral level, multiple sensory inputs are not required to arrive exactly at the same time, but some level of temporal proximity is needed to induce an interaction. In audio-visual speech, the visual inputs influence speech perception even when the visual input leads the auditory input by as much as 200 ms (Munhall et al., 1996; van Wassenhove et al., 2007). Temporal relationships have also been examined in somatosensory-auditory interactions [see review for Occelli et al. (2011)], but only in non-speech processing.

The current study investigates event-related potentials (ERPs) in conjunction with speech sounds and patterns of facial skin deformation that would ordinarily be associated with articulatory motion for these speech sounds. A robotic device was used to generate the somatosensory stimuli, that is, patterns of facial skin deformation that are similar in timing and duration to those experienced in speech production. We assessed the extent to which evoked potentials reflect somatosensory-auditory interaction during speech sound processing. We also examined the dynamic modulation of multisensory integration that occurred as a result of relative timing differences between the onsets of the two sensory stimuli. ERPs using electroencephalography have benefits for the investigation of temporal asynchronies because of their better temporal resolution in comparison with the other brain imaging techniques. The findings reveal neural correlates of temporal coding in multisensory convergence and suggest a dynamic modulation of multisensory interaction during speech.

2. Method

2.1. Participants

12 native speakers of American English participated in the experiment. The participants were all healthy young adults with normal hearing and all reported to be right-handed. All participants signed informed consent forms approved by the Yale University Human Investigation Committee.

2.2. Experimental stimulation and task

We examined interaction effects between speech sound processing and orofacial somatosensory processing. Event-related potentials (ERPs) were recorded in response to either individual or paired somatosensory and auditory stimulation. The somatosensory and auditory pairs used in the current test have been previously found to induce perceptual modulation in the context of both speech sound perception (Ito et al., 2009) and somatosensory judgment (Ito and Ostry, 2012) in behavioral tests.

We programmed a small robotic device (SenSable Technology, Phantom 1.0) to apply skin stretch loads for the purpose of orofacial somatosensory stimulation (Figure 1a). The details of the somatosensory stimulation device have been described in our previous studies (Ito et al., 2009; Ito and Ostry, 2010). Briefly, the skin stretch was produced using two small plastic tabs attached bilaterally with tape to the skin at the sides of the mouth. The skin stretch was applied upward. We applied a single cycle of a 3-Hz of sinusoidal pattern with 4 N maximum force. This temporal pattern has successfully
induced somatosensory event-related potentials in a previous study (Ito et al., 2013). Audio stimulation was delivered binaurally through plastic tubes and earpieces (Etymotic research, ER3A). We used a single synthesized speech utterance that was midway in a 10 step sound continuum between “head” and “had”. The speech continuum was created by shifting the first (F1) and the second (F2) formant frequencies in equal steps (Purcell and Munhall, 2006). The original sample sounds of “head” and “had” were produced by a male native speaker of English. These same sounds were used in the previous study that demonstrated modulation of speech perception due to facial skin stretch (Ito et al., 2009). We chose the center point of the continuum as an example of a perceptually ambiguous sound. In the current test, participants reported 68.5 % of stimuli as “head” due to the ambiguity of the stimulus sound.

We tested three somatosensory-auditory conditions that varied according to the time lag between the two stimuli. These were 90 ms lead, simultaneous, and 90 ms lag of the somatosensory onset relative to the auditory onset. A 90 ms temporal asynchrony was chosen because a 90 ms somatosensory lead reliably induced a modulation of speech perception in a previous study (Ito et al., 2009). Figure 1b shows three temporal relationships between somatosensory and auditory stimuli (lead, lag and simultaneous). Two unisensory conditions (somatosensory alone and auditory alone) were also assessed in the same test. The five conditions were presented in random order. The intervals between trials, after a participant’s judgment response, varied between 1000 and 2000 ms in order to avoid anticipation and habituation.

The participant’s task was to indicate whether the sound they heard was “head” or not. The participants’ answer was recorded by key press. In the somatosensory alone condition, the participants were instructed to answer not “head” since there was no auditory stimulation. Participant judgments and the reaction time from the onset of the stimulus to the key press constituted the behavioral measures. The participants were also asked to fixate their gaze on a cross without blinking in order to eliminate artifacts during ERP recording. The cross mark was removed every ten trials for a short break and then the recording was resumed along at a participants’ pace.

2.3. EEG acquisition and data processing

Event-related potentials (ERPs) were recorded from 64 electrodes (Biosemi ActiveTwo) in response to five stimulus conditions: somatosensory stimulation alone (soma), auditory stimulation alone (aud), and paired somatosensory and auditory stimulation (pair: lead, simult, lag). 100 ERPs per condition were recorded. Trials with blinks and eye movement were rejected offline on the basis of horizontal and vertical electro-oculography (over ±150 mV). More than 85% of trials per condition were included in the analysis. EEG signals were filtered with a 0.5-50 Hz band-pass filter and re-referenced to the average across all electrodes. Bias levels were adjusted using the average amplitude in the pre-stimulus interval (-200 to -100 ms).

We reconstructed auditory-like potentials by subtracting somatosensory potentials (soma) from the ‘pair’ potentials.

The results were compared with typical auditory event-related potentials (aud) and with participants’ behavioral performance, that is, the probability that the stimulus was identified as “head” during the test. We expected that ‘pair’ ERPs with the removal of the somatosensory potentials would be equivalent to the auditory-alone ERP, if neural responses to each of the unisensory stimuli are independent in the ‘pair’ condition. The subtracted potential should be different from the auditory responses if there is a nonlinear interaction.

The ERPs in the ‘pair’ condition were aligned at auditory onset and somatosensory responses were subtracted at the corresponding temporal shift in each condition. We focused on the first negative peak (N1) and the following positive peak (P2) at Fz and Cz because as a general tendency the maximum amplitude of the auditory ERP is observed at these electrodes and this was true of the current responses. A 60-ms time window was used to calculate the response amplitude. The analysis window was centered at the ERP peak location for each participant and each condition. Repeated-measures ANOVA was applied to assess differences in the four conditions (three ‘pair’ potentials and one auditory potential). Pairwise comparisons with Bonferroni correction followed.

2.4. Behavioral performance

Behavioral performance was evaluated using reaction time and judgment probability separately. Reaction time was calculated as the period between auditory onset and the behavioral response (key press for the speech sound identification). Repeated measures ANOVA was used to assess differences in reaction time across five conditions: three ‘pair’ and two unisensory conditions. We also calculated the judgment probability that the participant classified the sound as “head”. The somatosensory alone condition was not included in this analysis. Note that in more than 95% of somatosensory trials participants responded not “head” as instructed. Repeated measures ANOVA was used to compare judgment measures across conditions.

We also examined the extent to which the perceptual judgments were correlated with ERP amplitude change that were observed in response to changes in the relative timing of somatosensory-auditory stimulation. The correlation analysis was carried out between the participants’ judgment probability and the auditory ERP amplitude that was obtained when the somatosensory response was subtracted from ‘pair’ responses. For the purpose of this analysis, both variables were transformed into z-scores in order to remove differences in amplitude variability between individuals.
3. Results

We examined whether the timing difference between somatosensory and auditory stimulation induced changes in ERP activity. In order to compare paired auditory and somatosensory processing with that involved in speech perceptual processing, we examined these paired effects in relation to auditory related processing on its own. For this analysis, all of the data were aligned at auditory onset. We extracted auditory-related responses in the various paired conditions by subtracting the somatosensory-alone response from that obtained in the ‘pair’ conditions. The logic is that if there is a nonlinear interaction between somatosensory and auditory processing, the response after the subtraction should be different from the auditory alone response.

The subtracted potentials and the auditory-alone potentials showed a typical N1-P2 pattern with the first negative peak (N1) between 100 and 200 ms after auditory onset followed by a second positive peak (P2) between 200 and 300 ms (See Figure 2A). The maximum response was observed along midline electrodes near Cz (vertex electrode).

The peak amplitude at the Cz and Fz electrodes was quantified using 60-ms temporal window in each of the three ‘pair’ timing conditions (lead, simultaneous and lag) and for the auditory response alone (Figure 2B). Each color represents a different condition. Error bars represent the standard error across participants. We found a clear N1 response at Cz in all four conditions (lead, simultaneous, lag and auditory). The peak amplitudes were not statistically different for the four conditions \[F(3,33) = 0.122, p > 0.9\]. The peak amplitude of the P2 response showed a graded change according to the stimulus timing (lead, simultaneous and lag), although the change was statistically marginal as follows. Whereas repeated measures one-way ANOVA showed reliable difference across the four conditions \[F(3,33) = 5.95, p < 0.01\]. Post-hoc tests with Bonferroni correction showed that the lead condition was reliably different from lag condition \(p < 0.02\). The activity change in terms of ERP amplitudes was dependent on the relative timing of auditory stimuli for auditory processing. The activity change in terms of ERP amplitudes that was dependent on the relative timing of somatosensory processing was reliably correlated with judgment probabilities in the perception task. The results demonstrate a graded change in ERP amplitudes that was dependent on the relative timing of somatosensory-auditory interaction was quantified using event related potentials. We found a graded change in ERP amplitudes that was dependent on the relative timing of stimuli for auditory processing. The activity change in terms of auditory processing was reliably correlated with judgment probabilities in the perception task. The results demonstrate a clear multisensory convergence and suggest a dynamic modulation of multisensory interactions during speech. The two taken together suggest that somatosensory-auditory interaction may be important for speech perception.

Correlation analysis showed that the judgment probabilities in the four conditions were reliably correlated with N1 amplitude at Fz \((r = 0.3, p < 0.05)\) and marginally correlated with P2 amplitude at Fz \((r = 0.28, p = 0.05)\). The peak amplitude of N1 and P2 at Cz were not reliably correlated with the judgment probabilities \(N1: r = 0.18, p > 0.2, P2: r = 0.078, p > 0.6\). Thus, overall, although the magnitude of the correlation was relatively low, the results suggest that the perceptual modulation as measured behaviorally may be represented to some degree in the cortical response at Fz.

Reaction times across the five conditions: three ‘pair’ conditions and two uni-sensory conditions ‘soma’ and ‘aud’ were evaluated. We did not find any reliable differences across all five conditions \(F(4,44) = 0.532, p > 0.70\). This is inconsistent with typical responses due to multisensory stimulation conditions. Reaction time to respond to stimuli typically becomes shorter when two sensory modalities were stimulated simultaneously than when single sensory modalities are stimulated. The difference from the typical multisensory reaction may presumably be because the current task involved identification only.

4. Discussion

This study assessed the neural correlate of the temporal interaction between orofacial somatosensory processing and speech sound processing. The cortical activity associated with orofacial somatosensory-auditory interaction was quantified using event related potentials. We found a graded change in ERP amplitudes that was dependent on the relative timing of stimuli for auditory processing. The activity change in terms of auditory processing was reliably correlated with judgment probabilities in the perception task. The results demonstrate a clear multisensory convergence and suggest a dynamic modulation of multisensory interactions during speech. The two taken together suggest that somatosensory-auditory interaction may be important for speech perception.

In contrast, a reliable change was observed at Fz electrode in both N1 and P2 amplitude (see right two panels in Figure 4B). N1 responses at Fz were reliably different across the four conditions \[F(3,33) = 4.80, p < 0.01\]. Comparing the auditory alone responses with the other ‘pair’ condition yielded a reliable difference from the lead condition \(p < 0.05\) and a marginal difference from simultaneous condition \(p = 0.10\). The difference for the lag condition was not reliable \(p > 0.9\). Overall, the results reveal that auditory event-related potentials show a change when combined with temporally offset somatosensory stimulation. The largest change occurs when somatosensory stimulation leads for the speech sound. On the other hand, when somatosensory stimulation lags speech onset, the amplitude of the auditory potentials are not different from the potentials for auditory stimulation alone.

We also examined the behavioral results and their relationship with EEG activity. There was no reliable change of judgment probability in the three paired conditions in comparison to the auditory alone condition \[F(3,33) = 1.128, p > 0.3\].
interactions in perception are related to the specific temporal order of somatosensory-auditory inputs during speech production.

The timing of sensory stimulation is a key factor in multisensory interaction. The effective time-window for multisensory integration is known to be as long as 200 ms (Meredith et al., 1987; van Wassenhove et al., 2007). At a behavioral level, this is consistent with the results of our control test in which the participants perceived the skin stretch perturbation and the speech sound “head” as simultaneous in a comparable temporal range. Although the neural correlates of audio-visual interaction including that involving speech stimuli has been previously investigated (Pilling, 2009; Vroomen and Stekelenburg, 2010; Liu et al., 2011), the temporal range was larger than 200 ms, and hence it is not known the extent to which multisensory interactions occur at shorter temporal asynchronies. In the present study, dynamical modulation at an electrocortical level was found at a range of 100 ms. The current finding suggests cortical processing is sensitive to temporal factors even within the time range at which events are behaviorally judged simultaneous.

In audio-visual speech, the effective temporal range between auditory and visual stimulus onsets for effective multisensory interaction is asymmetric in terms of onset timing. While audio-visual speech phenomena, such as the McGurk effect is induced with up to a 240 ms of visual lead, while for visual lag the time window is much shorter (up to 40 ms) (Munhall et al., 1996; van Wassenhove et al., 2007). Our ERP findings may be comparable. N1 and P2 potential amplitudes in the 90 ms somatosensory lag relative to auditory onset were not different from those in the auditory alone response, whereas the lead and simultaneous condition showed a difference between the ‘pair’ and ‘sum’ responses, indicating that the somatosensory lead condition has affected audio processing, but not in the lag condition. This can probably be attributed to the temporal relationship between orofacial somatosensory inputs and acoustic output in speech production, since articulatory motion mostly precedes acoustic output in speech production [e.g. Mooshammer et al. (2012)].

The linkage between speech production and perception processing has been a topic of interest for over five decades (Liberman et al., 1967). Whereas the idea has been previously tested from the viewpoint of speech production and motor function (Fadiga et al., 2002; Watkins et al., 2003; Meister et al., 2007; D’Ausilio et al., 2009), the role of somatosensory function in speech perception has been overlooked. Previous psychophysical findings have showed that orofacial somatosensory inputs can influence speech processing (Ito et al., 2009). The current findings further suggest that somatosensory stimulation has access to cortical areas associated with speech processing. One intriguing possibility is that somatosensory information may be an important component in establishing the neural representations for both speech production and speech perception. Further investigation of the manner in which orofacial somatosensation modulates speech perceptual processing may provide some important clues to understanding the development of the linkage between speech perception and production.

5. Acknowledgements

We thank Alexis Johns and Joshua H. Coppolla for data recording and processing. This work was supported by the National Institute on Deafness and Other Communication Disorders Grants (R03DC009064 and R01DC012502), and the Natural Sciences and Engineering Research Council of Canada.

6. References


Linguistic Influences on Diphthong Realization of /ɔɪ/ in Hood German

Stefanie Jannedy¹, Melanie Weirich²

¹Zentrum für Allgemeine Sprachwissenschaft (ZAS), Berlin
²FSU Jena, Jena

jannedy@zas.gwz-berlin.de, melanie.weirich@uni-jena.de

Abstract

Multi-ethnic youth varieties of languages have been observed in many urban centers of Europe. Our previous work showed that speakers of Hood German have a tendency to centralize the diphthong /ɔɪ/ in comparison to speakers of a standard Berlin variety. In this study, we are investigating the linguistic factors that influence diphthong centralization of /ɔɪ/ in Hood German. We coded all words containing /ɔɪ/ that were elicited from a spontaneous speech corpus for the preceding and following segmental environment. Moreover, we coded if the diphthong occurred in an open or closed syllable, if it carried lexical stress and if it occurred in an accented position in the utterance. In addition to F2 values, the duration of the diphthong and the entire word was measured.

Results indicate that the centralization of /ɔɪ/ in Hood German (mirrored by a raised F2) is influenced by the segmental environment of the diphthong, with voiceless obstruents favoring centralization. Moreover, temporal differences in diphthong realizations between female Hood German and Berlin speakers are affected by the syllable structure the diphthong occurs in (closed vs. open).

Keywords: diphthong centralization, Hood German, multi-ethnolect, segmental factors, prosodic factors

1. Introduction

The variety of German spoken by many adolescents in the multi-ethnic and multi-cultural Berlin neighborhoods Kreuzberg, Wedding and Neukölln displays several morphosyntactic alternations that makes Hood German (see Jannedy & Weirich 2014 for a discussion of this term) differ from more standard varieties (see Dirim & Auer 2004, chapter 6 for a description of these alternations). Hood German is predominantly spoken by adolescents and young adults from neighborhoods that have a multi-ethnic composition but also by mono-lingual and mono-ethnic German speakers that live in these areas.

Recently, we have begun investigating the phonetic-phonological alternations found in this urban multi-ethnolect. In Hood German, the palatal fricative /ʃ/ as in ich ‘I’ is often realized as the unrounded version of the postalveolar fricative /ʃ/ or also as /ʃ/. We have argued that this alternation is part of a more general sound change pattern that is driven by two forces: by the spread of this feature from the central German dialect belt into Brandenburg and Berlin and by the pronunciation variant used by younger populations in Berlin. Moreover, this alternation may also (have) spread to Berlin German from the Leipzig area (Thuringian) during the 18th and 19th century (Auer 2013).

We have also noticed and described a vocalic alternation (Jannedy & Weirich 2013, 2014a) whereby the diphthong /ɔɪ/ is realized more centralized (higher F2) in the speech of females speaking Hood German compared to the speech of females who speak a more standard variety of Berlin German (left panel Fig. 1). The same tendency is observable for male speakers, too (right panel, Fig. 1).

Labov (1972) describes the social motivation of diphthong centralization on the island of Martha’s Vineyard and shows it to be group specific behavior but also dependent on linguistic influences: especially the vineyard born local fishermen on Martha’s Vineyard centralized their diphthongs /aʊ/ and /aɪ/ in the context of following obstruents – something that the descendants of the Portuguese settlers did not do.

We have described elsewhere (Jannedy & Weirich, 2014b) that some speakers of Hood German show hiatic tendencies in some of their productions of /ɔɪ/. We assume these to be expressions of their individual speaking styles and with expansion of our spontaneous speech database, we plan to investigate the social motivation for this alternation. In our current work, we are trying to determine the segmental and prosodic factors that facilitate the centralization process (Labov 1972). To that effect, we have begun to investigate the linguistic conditioning environment of this centralization process. While the nucleus of the diphthong /aʊ/ is centralized at least for the first two time points (start and early) for both male and female speakers, the offglide seems to keep the high front target. Thus, the onset of the nucleus here is the most relevant point for investigation. Therefore, we will focus in our investigation on the F2 value (as a reliable measure of the front – back dimension) at the onset (start) of the diphthong.

2. Method

2.1. Database

The spontaneous speech data was collected through sociolinguistic interviews. They were orthographically transcribed and added to the LaBB-CAT (Fromont & Hay 2012) database that allows for searching for all occurrences of a particular sound which then can be extracted. All the /ɔɪ/ tokens were extracted from our Hood German and from the Berlin German corpus (Jannedy & Weirich 2013). The Hood German database contains the speech of adolescents and young adults speaking this multiethnic variety of German, whereas the Berlin German corpus contains data from speakers with a native Berlin regional standard dialect background.

At this point, we have investigated and coded 1009 tokens of /ɔɪ/ as produced by 31 speakers: 8 Berlin German speakers (5 female, 3 male) and 23 Hood German speakers (13 female, 10 male). The Hood German speakers were recruited from schools and youth centers located in multiethnic neighborhoods. While almost all of them were born and raised in Germany, most of them have a second strong language such as
Turkish or Arabic. However, Hood German data was also obtained from mono-ethnic monolingual German speakers who live in these neighborhoods. Speakers in the Hood German group ranged in age between 14 and 24 years. Berlin speakers were also all born and raised in Berlin (and had a larger age spread; there was one 14 year old, others ranged between 25 and 53) but most importantly, did not display any of the grammatical features that are characteristic of Hood German (see Auer & Dirim 2004). Table 1 gives a summary of the number of tokens separated by gender (f, m) and corpus (Berlin, Hood German) included in our analyses.

Table 1: Number of /ɔɪ/ token per corpus by gender.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Corpus</th>
<th>Berlin</th>
<th>Hood German</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>Berlin</td>
<td>171</td>
<td>383</td>
</tr>
<tr>
<td>m</td>
<td>Berlin</td>
<td>113</td>
<td>342</td>
</tr>
</tbody>
</table>

2.2. Annotation and analyses

For each token, F2 values were extracted at five equidistant time points throughout the diphthong using PRAAT (Boersma & Weenink 2012). Thus, measurement points were at the start, middle and end of the diphthong and in between start and middle (early) and in between middle and end (late) (see Figure 1). In our analysis we will concentrate on the values of the start point, since here, differences between Berlin and Hood German speakers were significant for both male and female speakers. Moreover, the duration of each diphthong and the duration of the word it occurred in were logged.

We also coded the preceding and following segmental environment for each diphthong and grouped the individual sounds into broader phonological sound classes such as sonorants and obstruents. Obstruents are defined as those sounds where the airstream is orally obstructed in some way (stops, fricatives, affricates, nasals) whereas sonorants are sounds that have an unobstructed airstream (liquids and glides). Moreover, we coded the voicing status of the segments preceding and following the diphthong. Additionally, stops were also grouped by place of articulation (bilabial, alveolar, velar), however, the data was not equally distributed into these three categories. Due to data sparsity in some categories, we could not do adequate statistics.

We also coded prosodic characteristics of the diphthong such as the type of syllable it occurred in (open vs. closed), number of syllables of the word the diphthong occurred in, if the diphthong occurred in stressed or unstressed position in the word and if the word occurred in accented or unaccented position of the utterance.

3. Results

For statistical analyses linear mixed models (LMMs) as implemented in the lme4 package (Bates et al. 2011) were run in R (version 2.14.1, R Development Core Team 2008) with either the duration of the diphthong or the starting point of the F2 formant value of the diphthong as dependent variable. Speaker and word were included as random factors. Likelihood ratio tests were used for model comparisons to find the model with the best fit to the data. Comparisons were done in an additive stepwise fashion by including more factors (or interactions) in each step of the model and comparing the outcome model with the model that did not include the particular factor (or interaction) in question. P-values were obtained by likelihood ratio tests.

3.1. Duration of diphthong

To control for differences in speech rate, the duration of each diphthong was normalized by dividing its duration by the duration of the word. Note that many more diphthongs occurred in stressed than in unstressed syllables: only 9% of the Berlin and 7% of the Hood German diphthongs were unstressed, thus, for the durational analysis, we only considered the stressed instances of /ɔɪ/.

The peaks of the histogram in Figure 2 differ slightly between the groups and point to a generally longer diphthong duration for the Berlin speakers compared to the Hood German speakers. The distribution for the Hood German speakers is skewed towards the shorter durations whereas the Berlin diphthongs appear to be normally distributed. The data shown in Figure 2 contains accented and unaccented tokens.

Following additional prosodic factors were also considered: the sentence accent, the type of syllable the diphthong occurred in (closed or open) and the number of syllables of the carrier word. Prosodic prominence such as sentence accent can influence temporal patterns in speech (van Bergem 1993), and open syllables are expected to show longer vowel durations (Maddieson 1985). Also, the shorter the word is (i.e. the fewer syllables it has), the longer the proportion of the diphthong. Figure 3 shows the durations for the two genders (top: females, bottom: males) separately for closed and open syllables. The number of tokens for each group is given in Table 2.

Table 2: Number of /ɔɪ/ tokens by syllable type, corpus and gender.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Syllable</th>
<th>Berlin</th>
<th>Hood German</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>closed</td>
<td>78</td>
<td>303</td>
</tr>
<tr>
<td></td>
<td>open</td>
<td>78</td>
<td>65</td>
</tr>
<tr>
<td>m</td>
<td>closed</td>
<td>48</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td>open</td>
<td>57</td>
<td>108</td>
</tr>
</tbody>
</table>

The male speakers in both groups show rather similar productions of the diphthong, there was not much inter-group variability. The female speakers in the two groups however behave differently when producing diphthongs in open versus closed syllables: only the Berlin German females seem to produce longer diphthongs in open syllables than in closed syllables (see Figure 3).

For the statistical tests, separate analyses were carried out for the male and female data set. We first tested whether any of the potential factors (corpus, syllable, sentence accent) had an effect on duration by performing likelihood ratio tests between a model containing the particular factor in question and a model without this factor. If the model containing one of these factors showed a better fit to the data we further tested if including an interaction between this factor and the test variable corpus improves the fit to the data.
Figure 3: Distribution of normalized diphthong duration in open and closed syllables, separated by gender (above: females, below: males) and corpus (Berlin, Hood German)

For males, the number of syllables of the carrier word had a significant effect on the diphthong duration ($\chi^2(1)=46.6$, $p<0.0001$), with words containing more syllables decreasing the proportional length of the diphthong. Also, for males, the syllable structure had a significant effect on diphthong duration ($\chi^2(1)=10.4$, $p<0.002$), with diphthongs being longer in the open syllable than in the closed syllable.

For females, the number of syllables of the carrier word also had a significant effect on the diphthong duration ($\chi^2(1)=42.4$, $p<0.0001$). Here, however, we found an interaction between syllable structure and corpus ($\chi^2(2) = 6.04$, $p<0.05$), revealing that the effect of syllable structure is twice as high for the Berliners than for the Hood Germans.

A possible explanation for this finding might originate in the fact that a large number of speakers in the Hood German group has Turkish as another strong language. In an acoustic analysis of near minimal pairs, Jannedy (1994) investigated the effect of syllable structure on vowel duration in Turkish and found longer vowel duration in closed syllables than in open syllables. However, in this case we would assume that this effect is not due to language interference (otherwise males would show this effect, too), but that increased vowel duration in closed syllables may be a prosodic marker that was adopted by female speakers and may be another feature of Hood German.

3.2. Effects of segmental and prosodic factors on diphthong centralization

We investigated the potential interactions between the diphthong realization (i.e. centralization of the nucleus in Hood German reflected in a raised F2 starting value) and several linguistic factors, such as the segmental environment (following and preceding the diphthong), the prosodic structure (position of sentence accent), and - as for the diphthong duration - the syllable structure (open vs. closed syllable). The number of tokens carrying sentence accent versus not carrying sentence accent was equally distributed in both corpora (Berlin: 104 vs. 180, Hood German: 324 vs. 401). The distribution of tokens separated by syllable type and segmental environment are given in Tables 2 and 3 respectively.

<table>
<thead>
<tr>
<th>Table 3: Number of tokens separated by gender, and preceding and following sound class:</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
</tr>
<tr>
<td>f</td>
</tr>
<tr>
<td>f</td>
</tr>
<tr>
<td>m</td>
</tr>
<tr>
<td>m</td>
</tr>
</tbody>
</table>

The graphs in Figure 4 show the expected effect of a raised starting value of the F2 of the diphthong /ɔɪ/ for the Hood German speakers. However, this effect seems to be influenced by the preceding sound class environment. Generally, obstruents seem to have a greater effect on raising the F2 start value in the Hood German corpus than sonorants. This is true for male and female speakers. Yet, female speakers of Hood German show a greater effect than males. The same pattern holds true for the following environment (graphs not shown here).

Figure 4: Effect of the preceding segmental class (voiced sonorants; voiced obstruents; voiceless obstruents) and corpus (Berlin vs. Hood German) on the F2 starting value.

For statistical tests separate analyses were calculated for each gender, and LMMs and likelihood ratio tests were run analog to the durational analysis. For males, a main effect of corpus was found ($\chi^2(1) = 4.88$, $p<0.05$), with F2 values being 179 Hz higher for the Hood Germans than for the Berlin Germans. In addition, we found an effect of the following environment ($\chi^2(2)=7.9$, $p<0.05$), with sonorants having lower F2 values than voiced and voiceless obstruents. Given that stops built the biggest group of the obstruents, and within them nearly all were alveolar plosives, the raising effect of this place of articulation on the F2 value might explain this finding. However, since no effect of the preceding environment (with also many alveolar stops) was found, this explanation might not hold/b be sufficient. There was no effect of sentence accent or syllable structure on diphthong centralization.

For the females, an interaction between corpus and preceding sound class ($\chi^2(4)=16.94$, $p<0.002$), and between corpus and following sound class ($\chi^2(4)=43.4$, $p<0.001$) was found. In both cases the largest difference between F2 values of Hood German and Berlin speakers was found for the voiceless obstruents as mirrored in Figure 4 (top panel, for the preceding voiceless obstruents the difference was about 240 Hz, for the following about 250 Hz). Again, neither sentence accent nor
syllable structure was found to affect F2 values or interact with corpus. To investigate whether a particular sound class within the group of obstruents carries the effect of diphthong centralization in Hood German speakers, Figure 5 shows the F2 values separated by preceding sound class independent of voicing status (for the female speakers).

![Figure 5: Effect of individual sound classes on the F2 starting value.](image)

While it is obvious and expected that the sonorants (the liquids /l/ and /ʃ/), do not enhance the diphthong centralization in Hood German, there seems to be no individual sound class in the group of consonants that carries the functional load of the raising effect. However, due to an imbalance of the data with some classes containing only few tokens (see Table 4) running statistics is not sensible.

**Table 4: Number of tokens for the female speakers, separated by preceding sound class and corpus**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Berlin</th>
<th>Hood German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liquids</td>
<td>67</td>
<td>137</td>
</tr>
<tr>
<td>Nasals</td>
<td>17</td>
<td>35</td>
</tr>
<tr>
<td>Plosives</td>
<td>36</td>
<td>173</td>
</tr>
<tr>
<td>Affricates</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Fricatives</td>
<td>34</td>
<td>19</td>
</tr>
<tr>
<td>Glottal stops</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

Our analysis of over 1000 tokens of the diphthong /əʊ/ gathered from our spontaneous speech data base showed that speakers of Hood German and of Berlin regional standard German vary in their spatial and temporal realizations of the diphthong. However, these differences are facilitated by several segmental and prosodic factors.

First, while we found an influence of syllable structure on diphthong duration in males irrespective of corpus (with longer durations in open than in closed syllables), for females, only the Berlin speakers showed this effect. The female Hood German speakers in contrast, did not differ in the diphthong duration depending on syllable structure. We suggest that a missing lengthening effect of vowel duration in open syllables might be a prosodic feature of Hood German and point to a different timing pattern that is developing among females in this speech community.

Second, while male Hood German speakers showed a raised F2 value irrespective of segmental environment, for the female speakers, the centralization of the diphthong was enhanced by following and preceding obstruents contrary to Labov’s 1972 results where preceding liquids, glides and nasals facilitated centralization. While obstruents may intensify diphthong raising in Hood German, we believe that diphthong centralization is mainly due to social group membership and the degree of centralization to individual speaking style. Syllable structure or sentence accent seem to have less of an effect on diphthong centralization. Further analyses may reveal whether also word frequency effects or lexical specificities might play a role.

5. Acknowledgements

This work was supported by the Federal Ministry for Education and Research (BMBF) under Grant Nr. 01UG0711. We gratefully acknowledge the work of our RA Luisa Helmeke. Gratitude is owed to our conversation partners who shared their stories with us.

6. References


Acoustic Analysis of Dysarthric Speech after a Traumatic Brain Injury
Mathew Karia, Martine Grice & Doris Mücke

Abstract
The study focuses on two aspects: consonant production (incomplete oral closures and deterioration in coordination of the oral and glottal system in the production of plosives) and intonation (prosodic phrasing and its tonal marking). We carried out acoustic recordings of 10 subjects with dysarthric speech following TBI from Kenya and a healthy control group matched for age and language background. Subjects produced a fast syllable repetition task and recited long telephone numbers with natural phrasal breaks. The study revealed articulatory undershoot in oral and glottal gestures and restricted coordination between them. We found longer closure durations, high variation in VOT and reduced intensity differences in the plosive production of the TBI group. Furthermore, this group shows prosodic deficits in terms of reduced pitch modulation and rather slow speaking rates, making their speech sound monotonous. They failed to produce rises at the end of non-final phrases. In conclusion, we highlight the need for speech therapists to move from the traditional approach to a dynamic approach, integrating prosody and supralaryngeal articulation in developing diagnostic tests for dysarthria.

Keywords: dysarthria, traumatic brain injury, articulatory gestures, incomplete closure in plosives, prosody

Introduction
Dysarthria is a sensorimotor impairment of the articulatory movements (Ziegler, 2002). McAuliffe et al. (2010) report that dysarthria occurs as a result of damage to the central and peripheral nervous system, and may be characterised by impaired speech motor control and reduction in speech intelligibility. Disturbances in muscular control affect the speech processes through the resultant weakness, slowness, lack of coordination, or altered muscles (Darley et al., 1975). According to Kent et al. (2000) weak movements of the articulators, which have been observed in both kinematic and acoustic studies, result in a slow speaking rate. Furthermore, dysarthria due to TBI is related to a combination of respiratory, phonatory, articulatory and/or resonatory impairments (Cahill et al., 2000).

Traumatic brain injury frequently occurs when an external force neuroatomically injures the brain (McDonald, 1998). Theodoros et al. (2001) note that TBI may result in reduced ability to communicate, which in turn may have a serious effect on an individual’s social, family and academic life, among other things. Studies have demonstrated that TBI may result in impaired articulator movements, which in turn cause impaired speech production (Stierwart et al., 1996; Cahill et al., 2000). One aspect of this impairment is evident in the production of oral stops, or plosives.

Plosives are sensitive to effects of articulatory undershoot resulting from motor impairments. A complete closure of the active oral articulators is required to build up pressure in the supraglottal cavities, which is released in a burst (Ladefoged, 2001; Keith, 2003). This fine coordination between the oral and glottal system requires a full range of motor actions of the active articulators, which may cause problems to people with motor impairments (Kent & Kim, 2003).

Prosodic impairments in dysarthria have also been identified in dysarthric individuals (Darley et al., 1969a,b, Rosenbek & La Pointe, 1978; Vance, 1994; Schlenck et al., 1997; Ackermann & Hertrich, 2000). These studies have reported deviant prosody, especially slow speaking rate.

In the present study, we investigate both deterioration in plosive production and deviant prosody in the production of TBI subjects. Subjects and healthy controls are from Kenya and speak the English spoken in Nairobi as a second language. We use a dynamic approach, describing the observed phenomena in terms of articulatory gestures (Browman & Goldstein, 1989; Browman & Goldstein, 1992). We concentrate on target undershoot of the oral and glottal systems, and limitations of the temporal coordination between the systems in plosive production. For the prosodic realisation of phrasing, we concentrate on problems with pitch modulation that makes speech sound monotonous and may result from poor breath control.

2.1. Methods

2.2. Participants

Ten subjects with dysarthric speech following TBI were selected from Neurology and Speech/Occupation Therapy departments of Kenyatta National Hospital, Kenya, following consultation with a neurosurgeon. The subjects were compared with 10 subjects of a healthy control group matched in terms of age, gender and level of education. They all spoke Swahili as the first language and Kenyan English as the second language, which is also the language of instruction in school.

2.3. Speech materials

The target language of the present study was Kenyan English. Two articulation tests were administered. One was a fast syllable repetition task, Diadochokinesis (DDK) for 6 plosives (/pa/, /ba/, /ta/, /da/, /ka/, /ga/). Both the TBI subjects and their control subjects carried out a DDK test. A total of 600 tokens were recorded (6 plosives x 5 repetitions x 20 participants). The second test served to investigate prosodic phrasing. Fictitious phone numbers were presented as three chunks of three or four digits, i.e. 0721-222-989 or 0721-989-222) at normal and fast speaking rates. Phone numbers are prosodically structured in terms of phrasing and accented (Baumann & Trouvain, 2001). A total of 80 tokens were recorded (2 digit sequences x 2 speech rates x 20 participants).
2.4. Procedure
All recordings were carried out at the Kenyatta University (KU 99.9 FM) soundproof recording studio. The participants (TBI subjects and controls) were acoustically recorded using an Edirol (R-09HR) wave 16-bit/44.1 kHz digital recorder. The headset microphone, (AV-JEFE TCM 141 condenser microphone) was set at a mouth-to-microphone distance of 10cm.

2.5. Measurements
All data annotation and analysis were performed in PRAAT (Boersma & Weenink, 2011). The following acoustic measurements A-C were used for analysis of plosives (DDK task), and the measures D-E for the prosodic analysis (telephone number task).

(A) Closure duration (ms): Calculated intervocally as the interval from the offset of the initial vowel (a substantial decrease in periodic energy of F2) to the consonant’s release (voiceless stops: burst onset of noise; voiced stops: abrupt increase in periodic energy of F2).

(B) VOT (ms): The time from the release of the stop consonant to the onset of the following vowel. Negative VOT (prevoicing) was measured from a drop of periodic energy during the consonantal closure to the beginning of the following vowel.

(C) Intensity Difference; IntDiff (dB): Intensity of the initial vowel (V_max) minus intensity of the following consonant (C_max). This is a measurement to capture the degree of consonantal constriction (Parrel, 2010). Higher IntDiff values are expected for the voiceless plosives, lower values for voiced ones.

(D) F0, rises and falls: Frequency of occurrence of rising and falling F0 movements in the three prosodic phrases of each telephone number.

(E) Speaking rate (ms): Dividing the number of syllables in the utterance by the total duration of the utterance for each participant. The duration of the utterance was calculated by the duration of the complete digit number, either 0721-222-198 or 0721-989-222. The pauses are included in this measure.

Results

2.6. Plosive production
Table 1 provides means and standard deviations for the acoustic measures closure duration and VOT, separately for the control group and the TBI subjects.

Closure duration: The closure durations were on average 34 ms longer for TBI subjects than for the controls. In the control group, closure durations ranged from 73-90 ms, and in the TBI group from 98-121 ms. Results from independent t-tests revealed a significant difference between controls and TBI subjects for the plosive types /b/, /d/, /g/, /p/, /t/, /k/.

VOT: Subjects of the control group produced voiceless plosives with short voicing lags (mean VOT for /p/, /t/, /k/ = 15 ms), and voiced plosives with negative voicing lead (mean VOT for /b/, /d/, /g/ = -73 ms). This is because Kenyan English distinguishes between voiceless, non-aspirated stops and fully voiced stops. In contrast, individuals with TBI showed a deterioration in the plosive production when comparing the VOT for the intended voiceless and voiced plosives. The intended voiced labial and velar plosives, /b/ and /g/, were produced with negative voicing lags (mean VOT = -49 ms). However, the intended voiceless counterparts, /p/ and /k/, were also produced with negative voicing lags (mean VOT = -10.5). The voicing contrast for labial and velar plosives was merged in the production of the TBI subjects. Even more striking was the amount of variation in the TBI group. For all plosive types except /b/, they showed standard deviation values higher than 58 ms. In contrast, the control group showed VOT standard deviation values lower than 18 ms.

Table 1: Means and standard deviations (in parenthesis) for controls and TBI subjects in the measures closure duration and VOT.

<table>
<thead>
<tr>
<th></th>
<th>Con.</th>
<th>TBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>89(14.2)</td>
<td>110(35.9)</td>
</tr>
<tr>
<td>/t/</td>
<td>75(11.1)</td>
<td>111(34.9)</td>
</tr>
<tr>
<td>/k/</td>
<td>73(13.5)</td>
<td>121(38.6)</td>
</tr>
<tr>
<td>/b/</td>
<td>80(11.2)</td>
<td>111(21.4)</td>
</tr>
<tr>
<td>/d/</td>
<td>73(14.8)</td>
<td>98(26.5)</td>
</tr>
<tr>
<td>/g/</td>
<td>73(15.9)</td>
<td>114(46.1)</td>
</tr>
</tbody>
</table>

Intensity Difference (IntDiff): Figure 1 displays means for the Intensity Difference measurement in the production of controls and TBI subjects. This captures the difference between the maximum intensity of the vowel and the preceding consonant. High values indicate that plosives were maximally distinct from vowels, at best with a silent gap during the complete oral closure. When comparing the Intensity Difference for voiced and voiceless plosives within one group, the following picture arises: for the controls, the Intensity Difference was high for voiceless plosives (on average of 30 dB across all data) and low for the voiced ones (17 dB). In contrast, TBI subjects showed only a small difference between voiced and voiceless plosives (voiceless plosives: 28 dB; voices plosives: 23 dB). To sum up, the differences between voiced and voiceless plosives was considerably higher in the control group (amounting to 13 dB), than in the TBI subjects (5 dB).
2.7. Prosody task

\(F_0\) rises and falls: Figure 2 shows the analysis of the number of rising \(F_0\) movements in final position for the three prosodic phrases within one utterance in the telephone number task. The controls displayed similar results at both fast and normal rates. They had 85% rises in the initial phase, 95% rises in the medial phrase and all produced a final fall at the end of the utterance. In comparison, the TBI subjects presented different results for fast and normal rates. They had 55% and 45% rises in the initial phrase for normal and fast rates respectively, and 45% and 35% rises in the medial phrase for normal and fast rates respectively. Just like the controls, they all marked the final phrase with a fall at the end of the utterance.

As illustrated in Fig. 3, the controls marked the utterance beginning with a rise and the utterance end with a fall. However, as illustrated by a typical production in Fig. 4, the TBI subjects often had low endings at the end of every non-final phrase boundary. This factor, as well as \(F_0\) range effects, may have resulted in monotonous sounding pitch for the TBI subjects. The fall at the end of each phrase is interpretable as being at least in part due to subglottal pressure. Moreover, impaired neuromuscular control following traumatic brain injury would inhibit the proper adjustment and tension of the vocal folds to counteract the tendency for the vocal folds to slow down at the end of the phrase if there is a reduction in pressure.

![Figure 2: Frequency of occurrence for \(F_0\) rises (in %) at the end of three phrases of one utterance (initial, medial, final), separately for controls and TBI.](image1)

![Figure 3: \(F_0\) and spectrogram for digit sequence 0721-222-989 at a fast speaking rate, for a healthy control.](image2)

Speaking rate: Our results confirmed that controls were able to produce two different speaking rates in the telephone number task; normal and fast rate (paired t-test, p<0.05). However, this was not the case for the TBI subjects (paired t-test, p>0.05). Moreover, speaking rates were on average 3.15 syllables per seconds slower in the production of TBI subjects compared to controls. In contrast to the healthy controls, the TBI subjects showed large variation in their speaking rates. Some averaged close to the means of the healthy controls while others were considerably below the TBI group’s average (controls’ standard deviation amounts to 0.89 syllables per second and TBI subjects’ to 1.9 syllables per second). This may be explained by the fact that the individuals sampled suffered from different types and severity of dysarthria. As reported by Darley et al. (1969a), dysarthria presents different clusters of deviant properties, depending on its nature, and thus, speech rate may vary across the different cases. Therefore, the speaking rate may have differed due to the severity of each individual’s dysarthria.

2.8. Discussion

We found deterioration of plosive production and deviant prosody in speech production of subjects with dysarthric speech following TBI. The analysis of plosive production in the DDK task concurs with previous research findings that plosives are particularly liable to undergo articulatory undershoot of the oral system as a result of motor speech disorders. Our results indicate that impaired neural muscular control reduces the ability of the articulators to make the proper constriction, i.e. it leads to a weakening of oral gestures for the plosives. In the temporal domain, TBI subjects produced longer closure durations than healthy controls, and in the spatial domain they frequently undershot the target for a full oral closure, leading to a considerably high amount of aperiodic and periodic energy during the closure phase. We found the Intensity Differences between voiced and voiceless plosives to be critically reduced. The articulation was also imprecise when looking at the coordination of the glottal and oral control. The variation in the VOT measure was about four times higher in the TBI subjects than in the healthy controls. The TBI subjects often showed negative VOTs in the production of intended voiceless plosives indicating a weakening of the glottal abduction gesture for voicelessness. Figure 5 and 6 show examples for incomplete closures and normal high variation in the production of intended voiced plosives and voiceless plosives respectively. Kenyan English distinguishes between voiceless, unaspirated plosives and fully voiced plosives, but TBI subjects show serious limitations in producing the relevant acoustic cues for voicing and manner of articulation.
In the prosodic task, (producing phrasing in telephone numbers) the majority of TBI subjects show prosodic disturbances, probably as a result of poor breath control. In particular, they have reduced pitch modulation and rather slow speaking rates, making their speech sound monotonous. Whereas controls produced rises at the end of non-final chunks of the telephone number, the TBI subjects used a predominantly falling intonation in this position, thus failing to signal at the end of each chunk that more digits are to come.

Our study showed that effects of speech impairments of the oral and glottal systems are continuous, and are best described within dynamic frameworks. Furthermore, we have examined oral control (constriction in plosives), oral-glottal coordination (VOT) and glottal control (intonational rises phrase finally). It is important to highlight the need for speech therapists to treat these aspects of speech production in an integrated way when developing diagnostic tests for dysarthria.

References


Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data

Jangwon Kim, Naveen Kumar, Sungbok Lee, Shrikanth Narayanan

University of Southern California, Los Angeles, CA, U.S.A.

Abstract

This paper introduces an algorithm for robust segmentation of airway-tissue boundaries in the upper airway images recorded by the real-time magnetic resonance imaging. Compared to the previous method by Proctor et al. [1], the present algorithm performs image quality enhancement, including pixel sensitivity correction and grainy noise reduction, followed by robust estimation of airway path between the vocal tract walls. The airway path as well as the locations of the lips and the top of the larynx are found using the Viterbi algorithm. The tissue-airway boundaries are found for each grid line by searching the closest pixel of higher intensity than a threshold from the estimated airway path. The accuracy of the tissue boundary segmentation was evaluated in terms of root-mean-squared-error as well as statistics (mean and standard deviation) of error for specific region in the vocal tract. Results suggest that the proposed algorithm shows significantly less estimation error than the previous method [1], especially for the front cavity and the lower boundary.

Keywords: real-time magnetic resonance imaging, image segmentation, automatic tracking, vocal tract analysis

1. Introduction

Real-time Magnetic Resonance Imaging (rtMRI) [2] is an important tool for studying human speech production. The rtMRI provides the entire mid-sagittal view of the upper airway of a subject. The airway-tissue boundary segmentation in the MR images is often required as a pre-processing for the analysis and modeling of the vocal tract movements [3] and of the morphological structure of the vocal tract [4]. Performing this segmentation automatically is essential for analyzing rtMRI data of speech production, that typically comprise hundreds or thousands of video frames; the complex structure of the vocal tract, non-uniform field sensitivity of the tissues in head and neck, grainy noise, magnetic resonance (MR) image artifact, and the rapidly varying irregular vocal tract shape, however, make this problem challenging. This paper presents an algorithm for more robust segmentation of the MR images, which includes (1) retrospective pixel intensity correction of the MR images, (2) detection of the front-most edge of the lips and the top of the larynx, (3) segmentation of airway-tissue boundary in the vocal tract, and (4) measurement of the distance between the upper and lower boundaries. The current method improves the robustness of the airway-tissue boundary estimation over the previous method [1] by using a combination of data-driven way of pre-processing of the MR images, robust airway path estimation, and model-based weighted linear curve fitting.

2. Methods

2.1. Pre-processing of MR images

The MR images in rtMRI data often suffer from grainy noise and non-uniform field sensitivity of the tissues, depending on recording configuration [2]. Figure 1 (a) shows an example of the MR images in the USC-EMO-MRI corpus [5], which was recorded at an image frame rate of 23.18 frames/sec and a spatial resolution of $68 \times 68$ pixels. The present algorithm uses a multi-resolution approach to minimizing the effects of the noise, artifacts, and non-uniform field sensitivity of the tissues.

1. Create a field sensitivity map, denoted by $S$, of an original MR image using a morphological closing operation, followed by 2-dimensional median filtering. Figure 1 (b) shows the sensitivity map of the image in Figure 1 (a).

2. Create the set of edge points, as in Figure 1 (c), of the sensitivity map using the Canny edge detector [6] implemented in MATLAB. Likewise, create the set of edge points of the original image. Let $E_G$ and $E_{SM}$ denote the sets of edge points of the sensitivity map and the original image, respectively.

3. Create the head and neck boundary line $E_H$ by finding the left-most points of $E_G$ and $E_{SM}$. Then, create a binary image, denoted by $B$, of the head-neck region by setting the pixel intensity to be 1 for pixels in the right side of $E_H$ in each row and setting the pixel intensity to be 0 otherwise, as in Figure 1 (d).

4. Multiply the pixel intensity of the original image and the inverse of the pixel intensity of $S$ for non-zero elements in $B$, while setting the non-tissue pixel intensity to be zero. Figure 1 (e) shows the result image, denoted by $C$.

5. Perform a sigmoid warping of the pixel intensity in $C$ for suppressing grainy noise as well as highlighting tissue. Figure 1 (f) shows the final image.

3. Construction of grid lines

In order to detect the lips, the larynx, and the airway-tissue boundaries, the present algorithm constructs grid lines, adopting from the previous method [1]. The previous method is motivated by the analysis of the upper airway image by Ohman [7]. The grid construction method requires four manually chosen anatomical landmarks near the larynx, the highest point on the palate, the alveolar ridge, and the center of the lips, in one of the MR images. See [1] for the details of grid construction that we follow. The differences from the previous method are (i) that...
a user chooses the distance between the center of adjacent grid lines in the present algorithm, not by the number of grid lines as in the previous method, and (ii) that the origin of the reverse polar grid lines is placed at the top of the image on the horizontal coordinate of the labial landmark point. This point offers more smooth transition from the forward polar grid lines to the reverse polar grid lines. Note that such method of the grid line construction assumes that the head and neck are aligned such that the subject faces the left side of the image and the neck is vertically straight.

4. Lips and Larynx detection

For each frame, the initial and the final grid lines correspond to the locations of the top of the larynx and the front-most edge of the lips, respectively. Since these articulatory positions vary slowly and smoothly over time, the present algorithm finds each of their optimal positions by constraining rapid change of the estimated locations of them.

Assume \( q_i \) is a state at instance \( t \). \( N \) denotes the number of states. \( S_{q_i}^T \) denotes the transition score from \( q_i \) to \( q_j \). \( S_{q_i}^L \) is the likelihood score (of the observation) for \( q_i \). \( P_t \) is the prior score of \( q_i \). \( K \) is the number of instances. \( Q \) denotes a sequence of states \( q_1, q_2, \ldots, q_K \), one state for each instance. The objective score \( J \) of \( Q \) is defined as follows:

\[
J = \left( P_t S_{q_1}^L + w S_{q_2}^T + \sum_{u=2}^{K-1} S_{q_u}^L + w S_{q_{u+1}}^T \right)
\]  

(1)

where \( w \) is a weighting factor for \( S_{q_i}^T \). The optimal sequence \( Q^* \) is obtained by finding \( Q \) associated with the minimum \( J \):

\[
Q^* = \arg \min_{\{q_1, q_2, \ldots, q_K\}} J
\]  

(2)

For detection problem of the edge of the lips, \( q_i \) corresponds to the \( i \)-th grid line, where \( q_{N/2} \) is placed on the grid line of the labial landmark (the 77-th grid line in Figure 2). Also, \( S_{q_i}^T \) is the Euclidean distance between the centers of grid lines \( x \) and \( y \). \( S_{q_i}^L \) is the maximum pixel intensity of all pixels in the grid line \( x \). Note that the length and width of searching region for the lip detection are specified by users.

For the top of the larynx, \( q_i \) corresponds to the \( i \)-th grid line where \( q_{N/2} \) is placed on the grid line of the larynx landmark (the first grid line in Fig. 2). \( S_{q_i}^T \) is the same as defined for lips detection. Let \( D_{q_i}^L \) be the mean of the first-order derivatives of pixel intensities of \( x \), computed along the grid lines. Then, \( S_{q_i}^L = D_{q_i}^L \times W_x \), where \( W_x \) is an optional weighting term which gives more weight on higher grid line. \( W_x \) often helps for better estimation, especially when the low part of the larynx in MR images protrudes. This algorithm detects the point where the pixel intensity increases the most, searching from the top grid line.

The length and the width of searching regions (grid lines) for the lip detection and the larynx detection are specified by users. \( w \) is set to be 1 for these problems. One example of lips and larynx detection results is shown in Figure 3 (a).

5. Airway-path detection

The key idea behind improving airway-tissue boundary segmentation is to find an accurate and possibly approximate airway path in the upper airway first, from which the optimal airway-tissue boundaries can be determined easily and more robustly. The optimal airway paths passing through all grid lines in an MR image are determined by finding the paths of the minimum score, using the Viterbi algorithm. For this problem, each possible path in a grid line corresponds to a state, while each grid line corresponds to an instance. \( q_i \) corresponds to the \( i \)-th bin, where \( q_{N/2} \) is located at the center of the grid line; \( S_{x,y}^T \) is the Euclidean distance between bins \( x \) and \( y \), where the bins are located in the adjacent grid lines, one bin for each grid line. \( S_{x,y}^L \) is the pixel intensity (observation) of the bin \( x \), determined for each instance. Then, the optimal airway path is found by minimizing the score of possible bins as in the equation 2. The reason of using all bins, not only local minima as in the previous method [1] is that all local minima are sometimes found outside the upper airway when some regions in the vocal tract is fully closed. The estimated airway path in our method can still stay within the region of interest during full contact in the upper airway, restricted by the transition costs between states.
6. Airway-tissue boundary segmentation

Two airway-tissue boundaries, i.e., the upper and lower boundaries of the vocal tract walls, are determined at the first bins whose pixel intensity is over a certain threshold in the upper direction and lower direction, respectively. The threshold was set to be 0.5, where the maximum pixel intensity of each MR image is 1.

The estimated airway-tissue boundary points are smoothed by the robust local regression using weighted linear least squares and a 1-st degree polynomial model, implemented in MATLAB, for each image frame. Figure 3(c) illustrates the smoothed airway-tissue boundaries. Finally, a distance function for the airway-tissue boundaries is obtained by computing the Euclidean distance (in pixel unit) between the upper and lower boundaries or between the upper boundary point and the closest lower boundary point regardless of their grid line. It was observed that the later (green line in Fig. 4) is less erroneous, in particular near the lip region, than the former (blue line in Fig. 4). The initial boundary point for computing the distance function is in the grid line of the estimated larynx. The final boundary point is in the grid line of the first local minimum distance from the final grid line. Figure 4 illustrates a distance function in the upper airway. The software package which contains the MATLAB codes for the present algorithm and the subsets of data for demonstration is freely available at http://sail.usc.edu/old/software/rtmri_seg.

7. Evaluation of estimated airway-tissue boundaries

The estimated airway-tissue boundaries are evaluated against manually annotated tissue boundaries. For this purpose the annotators were instructed to sketch the lower and upper vocal tract walls using a continuous curve. For each of lower and upper boundaries, the Euclidean distance between each estimated boundary point and the closest point in the reference boundary for the estimated point is measured.

The statistics (mean and standard deviation) of the distance values are computed for each sub-region in the vocal tract and each phone. The sub-regions of the present algorithm are (1) grid lines $1 \sim 19$ for pharyngeal region, (2) grid lines $20 \sim 52$ for velar and dorsal constriction region, (3) grid lines $53 \sim 67$ (alveolar ridge landmark) for the hard palate region, and (4) grid lines $68 \sim 77$ for labial constriction region. The sub-regions of the previous algorithm are also determined in a similar way. The previous algorithm does not include the lip detection, thus large estimation error is observed in the grids after the lip landmark. For a fair comparison, the final grid line for analysis is fixed to the lip landmark point.

The palatal and dental corrections, and the mean pharyngeal wall were used as pre-processing for the baseline system [1]. See [1] for more details. For the present algorithm, the mean of the estimated boundary in the palatal region and the vertical position of the palate landmark is used in the final upper boundary. The reason for the palatal corrections in both algorithms is that the soft tissue in the hard palate region often shows significantly lower pixel intensity than other tissues, thus not sufficiently contrasted to the airway.

The list of phones used for evaluation is \{B, F, G, IY, K, M, N, NG, P, UW, V\}. Producing speech sound for these phones involves highly constricted or fully closed articulatory gestures, where the error of the estimated airway-tissue segmentation tends to be high. For each phone, 10 phone instances were randomly selected in a male subjects' data in the USC-EMO-MRI corpus. The acoustic phone boundary of each phone instance is obtained using an adaptive speech-text alignment tool, SalAlign [8]. The image frames within the starting and final times with one marginal frame in each side were selected. In
Figure 5: Errorbar of the distance (in pixel unit) between manual airway-tissue boundary and estimated airway-tissue boundary. From left to right in each phone, each errorbar is for pharyngeal region (black color), velar and dorsal region (red color), palatal region (green color), and labial region (blue color).

total, 492 image frames were used for evaluation.

Figure 5 shows the errorbar (as standard deviation) of the distance for each region and each phone for each estimated boundary. For lower boundary, the mean and standard deviation of the proposed algorithm is significantly smaller in all four regions than those of the baseline algorithm. Especially, the larger error in the front cavity (the palatal and labial regions) is significantly suppressed in the proposed algorithm. For the upper boundary, the baseline algorithm performs significantly better in the regions from the pharynx to the palate regions than for the lower boundary, presumably partially by the dental and palatal correction. However, the labial region still shows significantly large error. The amount of error in the labial region is significantly suppressed in the proposed algorithm. Table 1 shows the root-mean-squared-error (RMSE) for all estimated boundary points in each of the lower and upper boundaries. The proposed algorithm shows significantly lower RMSE for both lower and upper trajectories than the baseline algorithm. These results suggest that the proposed algorithm generates significantly more accurate airway-tissue boundaries than the baseline algorithm. In sum, the proposed algorithm generates more robust airway-tissue boundaries regardless of the phone and the vocal tract region than the baseline algorithm.

Table 1: RMSE between the estimated and manually-labeled boundaries in pixel unit.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>lower</td>
<td>2.56</td>
<td>0.71</td>
</tr>
<tr>
<td>upper</td>
<td>2.13</td>
<td>0.93</td>
</tr>
</tbody>
</table>

8. Conclusion and future work

The present algorithm estimates the airway-tissue boundaries from a robustly estimated airway path in each enhanced MR image. According to the quantitative evaluation on the estimated boundaries, the estimation error is significantly reduced by the present algorithm than the previous method [1] in terms of RMSE (2.56 to 0.71 for the lower boundary; 2.13 to 0.93 for the upper boundary). A major advantage of the proposed method over the baseline is robustness across different regions in the vocal tract . The proposed algorithm also extracts the positions of the front-most edge of the lips and the top of the larynx automatically. This helps constrain the search space of the airway-tissue boundaries, resulting more robust boundary estimation. In addition, with the algorithm one can estimate the length of the vocal tract above the larynx.

Automatic head movement correction for each MR image is an on-going work that we would like to use for more robust and convenient tissue boundary estimation. In addition, this approach also calls for a preprocessing technique that is better suited to this imaging modality.

9. Acknowledgements

This work is supported by NSF IIS-1116076 and NIH DC007124.

10. References


USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging

Jangwon Kim, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Sungbok Lee, and Shrikanth Narayanan

University of Southern California, Los Angeles, CA, U.S.A.
jangwon@usc.edu, toutios@sipi.usc.edu, yoonckim@usc.edu, yinghuaz@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu
http://sail.usc.edu/span, http://mrel.usc.edu

Abstract

This paper introduces a new multimodal database of emotional speech production recorded using real-time magnetic resonance imaging. This corpus contains magnetic resonance (MR) videos of five male and five female speakers and the results of evaluation for the emotion quality of each sentence-level utterance, performed by at least 10 listeners. Both speakers and listeners are professional actors/actresses. The MR videos contain MR image sequences of the entire upper airway in the mid-sagittal plane and synchronized speech audios after noise cancellation. The stimuli comprises the “Grandfather” passage and seven sentences. A single repetition of the passage and five repetitions of the sentences were recorded five times, each time with a different acted emotion. The four target emotions are anger, happiness, sadness and neutrality (no emotion). Additionally one repetition of the Grandfather passage was recorded in a neutral emotion and fast speaking rate, as opposed to a natural speaking rate for the rest of the recordings. This paper also includes a preliminary analysis of the MR images to illustrate how vocal tract configurations, measured in terms of distances between inner and outer vocal-tract walls along the tract, vary as a function of emotion.

Keywords: emotional speech database, magnetic resonance imaging, emotional speech production, emotion evaluation.

1. Introduction

Previous studies in emotional speech production have reported evidence that different emotions affect articulatory movements differently in a systematic way. Erickson et al. [1, 2] reported that the positions of the tongue tip, the tongue dorsum, the jaw and the lips, can be characterized by emotion type of speakers. Lee et al. [11] also found that emotional speech articulation exhibits more peripheral or advanced tongue positions, especially for sadness, which is in the line with the finding of Erickson et al. Lee et al. [3] also found that the movement range of the jaw is largest for anger, which is also in the line with the finding of Erickson et al. Kim et al. [4, 5] reported that the position, movement range and speed of articulators are important cues for distinguishing emotional state of speakers of their dataset. Kim et al. [6] also found that angry speech involves more emphasis on articulatory variation than happiness, while happiness involves more emphasis on the pitch variation; the degree of relative emphasis varied depending on speakers. Although the findings of these preliminary studies are limited to small lexical contents and small number of subjects, they hint that emotional information is encoded in a systematic variation of articulatory movements and prosodic behaviors.

The articulatory information used in the aforementioned preliminary studies was obtained using ElectroMagnetic Articulography (EMA). EMA allows monitoring the movements of a few flesh-point sensors attached on articulatory points of interest. Although EMA provides information of articulatory motions at a relatively fast sampling rate (100, 200 and 400 Hz for the NDI Wave system) with reliable accuracy overall [7] the number of articulatory points (maximum 6 sensors for a single NDI Wave system), and the vocal-tract region that the EMA can monitor are limited.

Recently, real-time magnetic resonance imaging (rtMRI) [8] with simultaneous speech recording [9] were utilized to study emotional variation in the articulatory domain. The rtMRI offers the entire view of the upper airway of a speaker in any scan plane of interest with no need of repetition, and therefore dynamic information on the movements of the lips, the jaw, the velum, the tongue (including in the pharyngeal region), and the larynx in the mid-sagittal plane can be captured. This information provided by the rtMRI allows us to study their patterns of articulatory coordination and global vocal tract parameters, such as the vocal tract length. Analyzing the rtMRI data of emotional speech of a male speaker, Lee et al. [10] found that angry speech is characterized not only by wider and faster vocal tract shaping, but also by more usage of the pharyngeal region than other emotions (neutrality, happiness and sadness). They also reported that happy speech exhibited shorter vocal tract length than the other emotions in the paper.

One of the important characteristics of emotional speech production is inter- and intra-speaker variability. Such a largely heterogeneous nature of emotion expression and perception has been raised in literature (e.g., [11, 6, 12, 13]), but the knowledge in orchestrated articulatory control for emotion encoding on top of linguistic articulation is still limited. Another interesting aspect is the mismatch between perceived emotion of listeners and the emotional state of speakers.

The present paper introduces a new speech production corpus of emotional speech, namely the USC-EMO-MRI corpus, which was collected at the University of Southern California using rtMRI. This corpus comprises the entire mid-sagittal vocal tract images and simultaneously recorded speech audios of 10 speakers, and emotion quality labels by at least 10 listeners for each speaker’s data. The articulatory and acoustic data are collected with three basic emotions (happiness, sadness, anger) and neutrality (no emotion presented).

The USC-EMO-MRI corpus is designed as a resource for systematic analysis for the inter- and intra-speaker variability...
of emotional speech in the articulatory movements and prosodic behaviors. This corpus also aims at assisting more comprehensive modeling of the vocal tract shaping in the upper airway and eventually the joint modeling of articulatory and acoustic behaviors with human-like emotion coloring. In addition, this corpus could be equally useful for other speech production studies, such as articulatory-to-acoustic forward mapping [14] or acoustic-to-articulatory inversion [15, 16] of emotional speech.

2. Data collection

Sequences of mid-sagittal images of the upper airways of five male and five female subjects were recorded in a 1.5 Tesla MRI scanner using a custom upper airway receiver coil and a real-time MRI acquisition protocol that has been described in detail elsewhere [17]. The resolution of the MR image is 68 × 68 pixels, where the pixel size is 3 mm × 3 mm. Hence, the field of view is 204 mm × 204 mm. All subjects are professional actors or actresses who have had theatrical vocal training. Figure 1 shows MR images extracted from the tMRI videos for one male and one female speaker in the database.

![MR images of two subjects, M1 and F1](image)

Figure 1: Example MR images of two subjects, M1 and F1 in Table 1

The frame rate of the reconstructed MRI video is 23.180 frames/sec. The details of MR image reconstruction that we followed for this database have been described in [18]. Speech audio was simultaneously recorded, using a custom fiber-optic microphone at a sampling rate of 20 kHz during the MR imaging. Noise cancellation was performed on speech audio using a custom adaptive signal processing algorithm [9], the audio and video signals were synchronized. See [9] for the details of noise cancellation and audio-video synchronization that we followed.

The subjects were asked to read stimuli after immersing themselves in one of four target emotions, such as neutrality, happiness, anger, and sadness, in the scanner. The stimuli consist of the “Grandfather” passage and seven sentences.

- The “Grandfather” passage: You wished to know all about my grandfather. Well, he is nearly ninety-three years old; he dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, “Banana oil!” Grandfather likes to be modern in his language.

- 7 sentences:
  1. John bought five black cats at the store.
  2. The Leopard, skunk and peacock are wild animals.
  3. Charlie, did you think to measure the tree?
  4. The queen said the KNIGHT is a MONSTER.
  5. Hickory dickory dock, the mouse ran up the clock.
  6. 9 1 5 (short pause) 2 6 9 (short pause) 5 1 6 2.
  7. Ma Ma MA (short pause) Ma Ma Ma (short pause) Ma Ma Ma Ma.

The subjects uttered the passage in normal speaking rate for each emotion and a second time in fast rate only for neutrality. The subjects also uttered six or seven sentences with seven repetitions in only normal speaking style for each emotion. The order of the sentences was randomized at each repetition of data collection, except for the sentences 6 and 7. The subjects were asked to repeat their intonation of the sentence 6 when they read the sentence 7. The sentence 7 was always presented right after the sentence 6. Subjects were asked to emphasize “KNIGHT” and “MONSTER” when reading the sentence 4. Table 1 shows the sentences uttered by each subject. The set of sentences is designed to investigate the effects of emotion expression on the syntactic, prosodic and rhythmic structure in corresponding spoken utterance, including one reiterant speech.

3. Emotion evaluation

The emotion quality of the data was evaluated for each sentence-level utterance with perceptual tests from at least 10 actors/actresses. After listening to speech audio, the evaluators were asked to give their opinion on three questions: (1) the best representative emotion among five categories, such as neutrality, anger, happiness, sadness, and ‘other,’ where ‘other’ was for the case that none of the listed four emotions was the best, (2) confidence in their evaluation, and (3) the strength of emotion expression. Confidence and strength were evaluated on a five-point Likert scale. The best emotion was determined by majority voting. If there were multiple emotions with the same evaluation score, the one of higher mean of confidence scores was chosen.

Table 1 shows the number of evaluators, sentences, the average and standard deviation of matching ratio between target and evaluated emotions. ‘Sent’ 1-7 denotes all seven sentences were recorded, while ‘Sent’ 1-6 denotes the sentence 7 was not recorded. The matching ratio refers to ‘the number of utterances whose target emotion and perceived emotion match’ over ‘the number of all utterances.’ The matching ratios and associated standard deviations reflect the perceptual goodness of speech emotions portrayed by the speakers.

Table 2 shows an example of confusion between target and perceived emotions. The evaluation result of one female evaluator on M1 is used for Table 2. A large number of ‘other’ implies that they are not pure target emotions, but possibly mixed emotions (e.g., happiness and sadness) and slightly different emotions (e.g., annoyed, fear, excited, nervous, worried) from the four target emotions. The USC-EMO-MRI corpus includes the evaluation results of every individual combination of speaker and evaluator.
Table 1: Summary of evaluation results of all evaluators. ‘#Eval’ denotes the number of evaluators. ‘Sent’ indicates the sentence ID included. ‘AVE’ and ‘STD’ denotes average and standard deviation of the matching ratio (%) between target emotion and evaluated emotion for sentence-level utterances, respectively.

<table>
<thead>
<tr>
<th>Subject ID (M: male, F: female)</th>
<th>#Eval</th>
<th>1-7</th>
<th>63</th>
<th>11</th>
<th>12</th>
<th>12</th>
<th>12</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>10</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>M2</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>M3</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>M4</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>M5</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>F1</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>F2</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>F3</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>F4</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>F5</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 2: An example of confusion between the target emotion and the perceived emotion by a single evaluator for sentence-level utterance of M1’s data.

<table>
<thead>
<tr>
<th>Target</th>
<th>Perceived</th>
<th>Neutrality</th>
<th>Anger</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutrality</td>
<td>58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Anger</td>
<td>2</td>
<td>42</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Happiness</td>
<td>1</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Sadness</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>42</td>
<td>39</td>
<td>39</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

4. Preliminary analysis of articulatory variation for different emotions

This section provides a preliminary analysis result on the MR images for emotion-dependent vocal tract movements. Vocal tract shapings of different emotions are compared in terms of cross-distances between the inner and outer vocal-tract walls from the larynx to the edge of the lips. In order to obtain these cross-distances, semi-automatic tissue-airway boundary segmentation was performed using a recently introduced MATLAB software [19]. This software performs (i) pixel sensitivity correction, (ii) noise suppression on the MR image, (iii) tracking of the lips and the larynx, (iv) segmentation of the airway-tissue boundary, and (v) computation of the distance function. The processes (ii) and (iv) are performed automatically based on the semi-automatically constructed gridlines. The cross-distances are obtained by measuring the Euclidean distance between the top-right and left-bottom airway-tissue boundary in each gridline from the larynx to the lips. The distance of the gridlines outside the upper airway (from the larynx to the lips) is considered to be 0. See [19] for the details of the method.

Fig. 2 illustrates the each step of parameter extraction process. Fig. 2 (a) shows an original MR image of the subject M1. Fig. 2 (b) shows the enhanced image by the pixel sensitivity correction and noise suppression. Fig. 2 (c) shows the gridlines. Finally, Fig. 2 (d) shows the tissue-airway boundary segmentation results.

Fig. 3 shows the mean of the range of the distance during the word ‘five’ in the sentence 1 for each emotion. In the plot, anger and happiness show wider movement range than sadness in the grid lines 49 ~ 68. The wider range of tongue movement for high arousal emotions (anger and happiness) in the palatal region than low arousal emotion (sadness) is well captured in the MR images. Fig. 3 also shows the difference between anger and happiness in terms of the tongue movement range; on average, anger shows wider movement range than happiness near the grid lines 13 and 23, while anger shows lower movement range than happiness in the grid lines 49 ~ 68. This indicates that on average, the pharyngeal constriction and releasing were more emphasized for anger than for happiness, while the palatal constriction and releasing were more emphasized for happiness than for anger, during producing ‘five.’

5. Discussion and Future works

Although this corpus relies on the emotion elicited from actors/actresses in laboratory conditions, the emotional variability exhibiting in the data still allows us to study the controls over linguistic variables such as prosodic context, phonological environment and neutrality of utterances. Xu et al. also argues that “lab speech” is not necessarily inadequate for human speech production research [20].

We are also working on collecting EMA data from the same speakers using the same stimuli. The EMA data and the MRI data provide complementary information for articulatory dynamics. More specifically, EMA provides 3-dimensional sensor trajectories of certain articulatory surface points, while the collected rMRI data provides image sequences of the full airway in the mid-sagittal plane represented by pixel intensity.

Development of robust (semi-)automatic MR image analysis tools for this corpus is another important on-going work of our research group. This task is important for improving the usability of this corpus in broader research communities, such as linguistics and psychology, with interests in speech production and perception.
The emotion evaluation for the USC-EMO-MRI corpus is still on-going process. To complement the emotion perception data from professional actors/actresses, the emotion evaluation from naive listeners will be also collected. Busso et al. [21] have reported the difference between professionals and naive listeners in terms of their emotion perception. It is interesting to investigate how the emotion perception of the two groups contrasts and to see whether the observation is a general trend in emotion perception.

6. Acknowledgements

This work is supported by NSF IIS-1116076 and NIH DC007124.

7. References


Acoustic transitions in Khmer word-initial clusters

James P. Kirby

Linguistics and English Language, University of Edinburgh, Scotland, UK

j.kirby@ed.ac.uk

Abstract

Onset clusters in Khmer (Cambodian) often appear with an acoustic transition between consonants, but the phonological status of these elements is indeterminate. If transitions result from gestural separation, they may disappear in fast speech. Acoustic analysis of data from 10 speakers shows that vocalic transitions in Khmer are found in largely predictable set of consonantal contexts. While their presence is modulated by speech rate, they never disappear completely, in some cases becoming more rather than less frequent in fast speech. Clusters containing transitions are generally longer in duration than those that do not, and are also longer than monosyllables containing a lexical schwa, but the transitions do not show any spectral evidence of a distinct gestural target. The possible interpretations of these findings are discussed in the context of the range of articulatory variation known to occur in the implementation of speech rate.

Keywords: Khmer, intrusive vowel, speech rate, onset cluster

1. Introduction

The presence of excrescent, intrusive or otherwise 'transitional' elements in consonant clusters is common in many of the world’s languages. However, research suggests that not all of these elements have the same status: in some cases they are argued to be epenthetic, involving the insertion of some type of phonological unit (e.g., Dutch: Warner et al., 2001) while in other languages they are claimed to be fundamentally phonetic effects, e.g. byproducts of gestural retiming (e.g., German: Jannedy et al., 2008). Teasing these two possibilities apart is not always straightforward: for some languages, evidence in favour of both positions has been advanced (e.g., Tashlhiyt Berber: Coleman, 2001; Ridouane & Fougeron, 2011).

Khmer, the national language of Cambodia, has a rich inventory of no less than 70 onset clusters (Table 1). Huffman (1972) proposed that Khmer onsets can be separated into three classes based on the type of transition they contain: (a) ‘Class 3’ clusters such as /k/ and /ph/, which are separated by a brief voiced transition (hereafter abbreviated /u/); (b) ‘Class 2’ clusters such as /ph/ and /kh/, which are separated by a brief voiceless transition (i.e., aspiration); and (c) ‘Class 1’ clusters such as /pt/ and /ks/, which are articulated without audible transitions. This suggests that the set of conditions giving rise to intrusive vowels in Khmer may differ from that documented for languages such as Tashlhiyt Berber (Ridouane & Fougeron, 2011).

Recently, Butler (2012, 2014) conducted an acoustic study of 20 Khmer onset clusters. Her data suggest that the presence of an acoustic transition can vary with cluster type: for some clusters (e.g., /bl/) /u/ is consistently present, while others (e.g., /ml/) may occur without an audible transition. Butler also compared spectral and durational properties of transitional vocoids in words like ឈ្នាួ /hute/ > [m/teh] ‘pepper’ to short lexical vowels in words like ឃ្ន /niu/ ‘dash away’. She reports that clusters containing transitional vocoids were significantly shorter than monosyllables with lexical short vowels, and the formant structure of the vocoids was more variable than that of corresponding lexical vowels. Butler interprets her findings in the context of Articulatory Phonology (Browman & Goldstein, 1992) as indicating that transitional elements in Khmer onset clusters lack an associated articulatory gesture and are simply surface-level effects of the phasing of the consonant gestures, as has been argued for languages like Tashlihyy or German.

However, the fact that transitional vocoids differ acoustically from lexical short vowels does not necessarily mean that these elements do not involve a phonological target of some kind (gesture or segment), only that this target is not the same as that of a nuclear vowel in a lexical monosyllable. Many Khmer monosyllables with CC-onset clusters are known to have developed from the loss of the initial syllable of disyllabic iambs, e.g. /paj/ ‘husband’ < Sanskrit pati, /k’kae/ ‘dog’ < Old Khmer kae (Huffman, 1972, p. 61) as part of a historical process common to many Southeast Asian languages (Matisoff, 2003). It is thus conceivable that these elements are still part of the phonological specification of lexical items where they (predictably) appear, or that they are historically intrusive but have become lexicalized with their own unique gestural target(s).

Here, I look for additional evidence bearing on the phonological status of acoustic transitions in Khmer onset clusters by considering acoustic data from 35 cluster types in the context of a speech rate manipulation. Disappearance of intrusive or excrescent vowels in fast speech has been observed in a number of languages (see Hall, 2006 for a review), arguably as the result of an increase in the relative overlap of extant articulatory gestures (Gay, 1981; Munhall & Lòfqvist, 1992; Byrd & Tan, 1996). If acoustic transitions in Khmer onset clusters are found to disappear in fast speech, this would lend further support to the idea that they are phonetically excrescent and do not involve phonological specification of a segment or gesture.

Table 1: Onset clusters in Khmer (after Huffman, 1972). Class 1 are dark, Class 2 are light, and Class 3 are not shaded.

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>pt</td>
</tr>
<tr>
<td>t</td>
<td>tp</td>
</tr>
<tr>
<td>k</td>
<td>kp</td>
</tr>
<tr>
<td>pb</td>
<td>wp</td>
</tr>
<tr>
<td>pm</td>
<td>wp</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>pmb</td>
<td>wpb</td>
</tr>
<tr>
<td>pmn</td>
<td>wpn</td>
</tr>
<tr>
<td>bmn</td>
<td>bmn</td>
</tr>
<tr>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>p</td>
<td>p</td>
</tr>
</tbody>
</table>

Comparison with other languages:

Khmer, the national language of Cambodia, has a rich inventory of no less than 70 onset clusters (Table 1). Huffman (1972) proposed that Khmer onsets can be separated into three classes based on the type of transition they contain: (a) ‘Class 3’ clusters such as /k/ and /ph/, which are separated by a brief voiced transition (hereafter abbreviated /u/); (b) ‘Class 2’ clusters such as /ph/ and /kh/, which are separated by a brief voiceless transition (i.e., aspiration); and (c) ‘Class 1’ clusters such as /pt/ and /ks/, which are articulated without audible transitions. This suggests that the set of conditions giving rise to intrusive vowels in Khmer may differ from that documented for languages such as Tashlihyy (Ridouane & Fougeron, 2011).

Recently, Butler (2012, 2014) conducted an acoustic study of 20 Khmer onset clusters. Her data suggest that the presence of an acoustic transition can vary with cluster type: for some clusters (e.g., /bl/) /u/ is consistently present, while others (e.g., /ml/) may occur without an audible transition. Butler also compared spectral and durational properties of transitional vocoids in words like ឈ្នាួ /hute/ > [m/teh] ‘pepper’ to short lexical vowels in words like ឃ្ន /niu/ ‘dash away’. She reports that clusters containing transitional vocoids were significantly shorter than monosyllables with lexical short vowels, and the formant structure of the vocoids was more variable than that of corresponding lexical vowels. Butler interprets her findings in the context of Articulatory Phonology (Browman & Goldstein, 1992) as indicating that transitional elements in Khmer onset clusters lack an associated articulatory gesture and are simply surface-level effects of the phasing of the consonant gestures, as has been argued for languages like Tashlihyy or German.

However, the fact that transitional vocoids differ acoustically from lexical short vowels does not necessarily mean that these elements do not involve a phonological target of some kind (gesture or segment), only that this target is not the same as that of a nuclear vowel in a lexical monosyllable. Many Khmer monosyllables with CC-onset clusters are known to have developed from the loss of the initial syllable of disyllabic iambs, e.g. /paj/ ‘husband’ < Sanskrit pati, /k’kae/ ‘dog’ < Old Khmer kae (Huffman, 1972, p. 61) as part of a historical process common to many Southeast Asian languages (Matisoff, 2003). It is thus conceivable that these elements are still part of the phonological specification of lexical items where they (predictably) appear, or that they are historically intrusive but have become lexicalized with their own unique gestural target(s).

Here, I look for additional evidence bearing on the phonological status of acoustic transitions in Khmer onset clusters by considering acoustic data from 35 cluster types in the context of a speech rate manipulation. Disappearance of intrusive or excrescent vowels in fast speech has been observed in a number of languages (see Hall, 2006 for a review), arguably as the result of an increase in the relative overlap of extant articulatory gestures (Gay, 1981; Munhall & Lòfqvist, 1992; Byrd & Tan, 1996). If acoustic transitions in Khmer onset clusters are found to disappear in fast speech, this would lend further support to the idea that they are phonetically excrescent and do not involve phonological specification of a segment or gesture.
2. Method
2.1. Materials and subjects
10 native Khmer speakers (3 female) of the Phnom Penh dialect were recorded reading 79 lexical items 3 times each at two self selected speech rates. Participants were instructed to produce the first set of items slowly and carefully, and the second set as rapidly as possible. All items in both conditions were embedded in the frame sentence /kprohm taa/ ‘I say ___ again’, and randomized and counterbalanced across blocks and participants. Here, we focus on the 61 items that include one of 35 onset clusters /pd ph pk pn pr pb th tk tl tn t\,\,t\,\,k/.

2.2. Segmentation
Acoustic data were segmented and annotated using Praat 5.3.57 (Boersma & Weenink, 2013) to determine total syllable duration along with durations of C₁ release, C₂ closure and release, transitional vocoid (if present), and syllable rime. A combination of spectral and acoustic cues was used to determine presence of a transitional vocoid, including presence of a periodic waveform, increase in signal energy at C₁, onset of a transitional vocoid, including presence of a periodic waveform, increase in signal energy at C₁ release, and/or a region of formant structure (see Figure 1).

Figure 1: Segmentation of token /piv/ ‘pumpkin’, speaker 6, slow condition, repetition 3 showing C₁ (o)subset, (v)ocoid, C₂ closure and release (b)urst, and syllable rime (m).

2.3. Analysis
Results were analysed with hierarchical (‘mixed-effects’) logistic or linear regressions as appropriate, with random intercepts for subjects and items. Use of random slopes, where appropriate (and where models would converge), is noted below.¹

3. Results
3.1. Rate and distribution of voiced transitions
The rate of appearance of voiced transitions by speech rate is given in Table 2. In general, either of the members of the cluster being voiced was sufficient to condition \( \exists \) (cf. Butler, 2014:94). There was just one instance in 3,640 tokens where a transitional vocoid appeared in a completely voiceless cluster, a fast speech token of /\( \text{th} /\)khaa/ ‘prosperous’. Overall rates of vocoid appearance were relatively stable across subjects and conditions (slow: \( \mu = 0.39, \sigma = 0.03 \); fast: \( \mu = 0.27, \sigma = 0.08 \)).

A hierarchical logistic regression predicting presence of an intrusive element from covariates C₁, C₂ and RATE confirms the main effect of RATE (\( \beta = -1.53, SE = 0.12, p < 0.001 \)), but while the appearance of \( \exists \) is modulated by rate, it never disappears entirely. For Class 3 items, \( \exists \) occurred less frequently in fast speech (\( \beta = -2.26, SE = 0.21, p < 0.001 \)), although this effect was greater for some clusters than others (compare e.g. /md/ vs. /ml/). For Class 2 (C + nasal) sequences, on the other hand, \( \exists \) actually occurred more frequently in fast speech (\( \beta = 1.24, SE = 0.25, p < 0.001 \)).²

3.2. Duration of voiced transitions
If transitional elements are part of the phonological specification of the syllable, we might expect to observe an increase in the duration of onset clusters containing such elements. Figure 2 shows the duration of C₁C₂ sequences by RATE according to the presence or absence of \( \exists \) (defined as when duration of \( \exists > 0 \) ms). Clusters containing \( \exists \) were significantly longer than those that did not in slow speech (\( \beta = 13.5, SE = 2.8, p < 0.001 \)) and marginally so in fast speech (\( \beta = 9.8, SE = 3.13, p = 0.057 \), with random by-subject and by-item slopes for RATE).

---

¹The treatment of aspirated stops as sequences (clusters) of plain stops + /h/ is usually argued for on morphological grounds (see e.g. Henderson, 1952), but cf. Section 3.5.

²Data and code may be found on the author’s website at http://www.lel.ed.ac.uk/~jkirby/khmer/.

---

Table 2: Rate of appearance of voiced transitions (\( \exists \)). Dark is Class 1, light is Class 2, white is Class 3. For each cluster the top row gives the rate in slow speech; bottom row in fast speech.

<table>
<thead>
<tr>
<th>C₁</th>
<th>C₂</th>
<th>( \beta )</th>
<th>p</th>
<th>t</th>
<th>k</th>
<th>b</th>
<th>d</th>
<th>n</th>
<th>( \eta )</th>
<th>l</th>
<th>r</th>
<th>s</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0</td>
<td>0</td>
<td>0.83</td>
<td>0.13</td>
<td>0.05</td>
<td>1.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>t</td>
<td>0</td>
<td>0</td>
<td>0.82</td>
<td>0.19</td>
<td>0.02</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.70</td>
<td>0.27</td>
<td>0.13</td>
<td>0.27</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>m</td>
<td>1.00</td>
<td>0.97</td>
<td>0.93</td>
<td>1.00</td>
<td>0.90</td>
<td>0.90</td>
<td>1.00</td>
<td>0.90</td>
<td>0.90</td>
<td>0.43</td>
<td>0.87</td>
<td>0.63</td>
<td>0.27</td>
</tr>
</tbody>
</table>

| 1.00 | 0.60 | 0.73 | 0.66 |

---

Figure 2: Duration (in ms) of C₁C₂ with and without \( \exists \) by subject and speech rate. Error bars show 95% confidence intervals.
3.3. Comparison with lexical vowels

Butler (2014) observes that the duration of the transitional vo-
coid in C∃C syllables was shorter than that of corresponding
(short) lexical vowels. This was tested with the present data by
comparing durations of C∃C clusters to the durations of eight
CαC and CαN monosyllables (e.g. ʰth/‘hak/ ‘granary’, ʰpθ/‘poun/ ‘clever’) also recorded as part of the production task. C∃C on-
set clusters (where C1 is voiceless and C2 is voiced, the primary
environment where ɜ occurs) were longer than CαC monosylla-
bles at slow (β = 32.94, SE = 5.1, p < 0.001) but not fast (β = 6.35, SE = 13.25, p = 0.65) speech rates. How-
ever, CαN onset clusters were shorter than CαN monosyllables
at both speech rates (slow: β = 73.77, SE = 9.61; fast: β = 82.11, SE = 10.76; both p < 0.001).

3.4. Evidence from formant transitions

The formant transitions of the transitional vocoids were also ex-
amined. If these elements have an articulatory target, one might
expect to see this reflected in their height relative to the sur-
rrounding consonants. This effect would be most obvious in
cases where ɜ is flanked by two consonants sharing the same
place of articulation. In such cases, if ɜ is truly targetless, the
difference between F2 at the midpoint and F2 at onset or off-
set should not differ significantly from zero. Indeed, this is
what we find: even when including subject- and item-specific
slopes, neither F2on−F2mid nor F2mid−F2off is significantly
different from zero (F2on−F2mid: β = 16, SE = 13, p = 0.31;
F2mid−F2off: β = 68.6, SE = 38.3, p = 0.12; see Figure 3).

3.5. Voiceless transitions

Huffman’s class system suggests that the durational properties
of the C1 release may vary predictable with the manner of C2.
Figure 4 summarizes the durations of voiceless C1 obstruents
(.esp k/l) by class and manner of C2. The release duration of
C1 in (Class 1) C+ḥ/ clusters (i.e., aspirated stops) was not
distinguishable from that of (Class 2) C + sonorant clusters at
either slow or fast speech rates; durations of C1 release when
C2 was a sibilant, obstructive, or fricative, or when the cluster
was /kŋ/, were similarly indistinct. In both conditions, how-
ever, the duration of the first group was significantly longer
than that of the second (slow: β = −35.33, SE = 2.1; fast: β = −21.86, SE = 2.19), and the presence of ɜ reduced the
duration of the C1 release (slow: β = −19.1, SE = 1.65; fast: β
= −14.65, SE = 1.29; all p < 0.001).

3.6. Cluster length as proportion of syllable

It is also instructive to consider the duration of onset clusters
as a proportion of total syllable length by class and cluster type
(Figure 5). Clusters where C1 = /p t k/ appear to fall into one of
two classes: the total duration of aspirated stops (C+ḥ/ ‘clus-
ters’) is not distinct from that of /Cv/ clusters at either speech
rate, nor are the durations of the remaining clusters types from
other another. Proportion of total syllable duration remains
extremely stable across speech rates for all cluster types (including
sonorant-initial clusters, not shown here).

4. Discussion

The goal of the present study was to shed new light on the
phonological status of transitional elements in Khmer through
the use of a speech rate manipulation. The most unequivoc-
ally finding is that the presence of a single voiced element in a
Khmer onset cluster is sufficient to license an intrusive vocoid.
However, even when this basic condition is met, the appearance
of ɜ is highly variable: it appears reliably when C1 is a sonorant
(l/m or /ŋ/) or C2 is /t/, and remains frequent when C2 is voiced
(l/ or /d/), but is notably less frequent when C2 is a sonorant
(l/m or /ŋ/). The exception to this last generalisation is the cluster
/kŋ/, which always occurs with ɜ in slow speech. This does not
seem to merely be an effect of homorganicity (ɜ is much less
common in homorganic clusters like /l/m or /ŋ/m) but may instead
reflect an articulatory strategy aimed at maximising perceptual
salience, given that l/ and /t/ may be especially confusable in
syllable onsets (Narayan, 2008).
The spectral characteristics of ∃ do not give any indication that it has an associated spatial target or tongue body gesture (consistent with the findings of Butler, 2014), and with few exceptions, overall cluster duration was found to be stable across speech rates for a given cluster type. However, for at least some speakers, ∃ appears to contribute significantly to total onset length (CxC clusters were longer than CC clusters in slow speech, and marginally so in fast speech) and the resulting clusters are longer than lexical CxC (but not CaN) monosyllables (cf. e.g. Lebanese Arabic: Gouskova & Hall, 2009; Tashlhiyt Berber: Ridouane & Fougéron, 2011). These latter findings are consistent with an account on which ∃ is phonologically specified in some way.

One might be tempted to view the fact that ∃ generally fails to disappear in fast speech as additional evidence that it has a phonological target. However, this finding is consistent with a number of other interpretations. The use of the speech rate manipulation paradigm was motivated by the idea that if vowel intrusion in slow speech is a byproduct of gestural separation, it may be prone to disappear in fast speech as the relative overlap of the gestures increases. This is based on the assumption that the timing between gestures is pliable. Another possibility is that speakers shorten the duration of each consonantal gesture while maintaining the relative timing between them (Byrd & Tan, 1996); if two consonantal gestures are phased such that the onset of the second is initiated during the release phase of the first, this may result in an audible transition regardless of speech rate (Gafos, 2002). Moreover, individual speakers may differ in whether or not an increase in speech rate results in an increase in gestural overlap due to their employing different gestural implementations of the same cluster types (Tjadjen & Weismer, 1998; Davidson, 2006). Such individual differences seem likely to underlie at least some of the variation observed in the present study. While fast speech impacted the appearance of ∃, it did not affect all cluster types in the same way. In particular, it is not clear why intrusion becomes more rather than less likely in fast speech for Class 2 clusters where \( C_2 = \eta \), nor why /kŋ/ should pattern with (Class 3) /pf/ and /hθ/ rather than with (Class 2) /kn/ and /kθ/ (though as noted above, there may be a perceptual explanation for the behaviour of this particular cluster). These findings raise the possibility that clusters in Khmer may be to some extent individually specified for gestural coordination and timing.

Finally, while the disappearance of a transitional vocoid at fast speech rates may constitute good evidence that it is intrusive, the converse is not necessarily true (Hall, 2006). This means that the persistent appearance of intrusive vocoids at fast speech rates in Khmer does not permit us to conclude they are phonologically specified. Instrumental articulatory data may help build a more accurate picture of how Khmer onset clusters are implemented and whether or not there is a gestural or segmental target associated with transitional elements.

5. Summary

This study has examined the effect of consonantal context and speech rate on interconsonantal acoustic transitions in Khmer onset clusters. Both context and speech rate were found to have variable effects on the appearance of intrusive material. While the majority of the findings are consistent with an account on which transitions are fundamentally phonetic effects of gestural timing, additional study will be necessary to fully understand the status of acoustic transitions this language.

6. Acknowledgements

Special thanks to Mr Sor Sokny of the Buddhist Institute, Phnom Penh for his invaluable assistance with this project, and to Patrycja Strycharczuk for her insightful comments. This study was funded in part by the University of Edinburgh Hayter Fund and the Carnegie Trust for the Universities of Scotland.

7. References


Vowel Coarticulatory Effects on Kannada Retroflex Stops

Alexei Kochetov¹, N. Sreedevi²
¹University of Toronto, Canada
²All India Institute of Speech and Hearing, India
al.kochetov@utoronto.ca, srij_01@yahoo.co.in

Abstract

Ultrasound tongue imaging data were collected from 10 speakers of Kannada with the goal to investigate effects of adjacent vowels on the articulation of retroflex consonants. The speakers produced multiple repetitions of geminate retroflex, dental, and labial stops in the symmetrical high front and back vowel contexts. Tracings of sagittal tongue contours were made during each consonant’s closure and compared between the two contexts. The results revealed consistent fronting/backing of the tongue next to /ɨ/, compared to /ɯ/. The magnitude of the coarticulatory effects was much greater for the labial than for the two lingual articulations, and somewhat greater for the retroflex than for the dental. These results are consistent with predictions of theories of coarticulation, and namely the Degree of Articulatory Constraint model (Recasens, Pallarès, and Fontdevila 1997). While the coarticulatory displacement of the tongue body for retroflex stops is relatively small, it may be sufficient to produce marked differences in the constriction location, noted in previous palatographic studies of retroflexes.

Keywords: ultrasound, retroflex, coarticulation, Kannada

1. Introduction

Theories of coarticulation, such as the Degree of Articulatory Constraint model (DAC: Recasens, Pallarès, and Fontdevila 1997), predict that the extent to which vowels affect consonants is inversely correlated with the involvement of the tongue body in the consonant production. Specifically, the raising of the tongue body for posterior coronals (alveopalatals or palatals) is predicted to render them particularly resistant to coarticulatory effects from high peripheral vowels /i/ and /u/. In contrast, the lack of active tongue body control for labials predicts them to show extensive vowel coarticulation, fronting or backing. Dentals or alveolars are expected to exhibit intermediate behaviour, as the tongue body is not directly involved in their constriction formation, yet mechanically coupled with the tongue tip.

These theoretical predictions have been confirmed in a number of studies investigating lingual and non-lingual consonants in a variety of languages (see Recasens 1999, Farnetani and Recasens 2010 for reviews). Much less work, however, has been done on retroflexes – consonants articulated with the tip or the underside of the tongue against the hard palate. In principle, some lowering of the tongue body, which accompanies the curling of the tongue tip, should constrain the retroflex articulation. This would make retroflexes more coarticulation-resistant than dentals or alveolars. Yet, a few articulatory studies of retroflexes in South Asian languages have reported the opposite – greater vowel coarticulatory effects on these consonants compared to dentals (Dave 1977 on Gujarati; Dixit and Flege 1991 on Hindi; Krull and Lindblom 1996 on Hindi and Tamil; Khatiwada 2007 on Nepali). As these effects were mostly based on the location of the tongue tip constriction obtained using static palatography or electropalatography, it remains unclear whether the same effects would hold for the tongue body posture.

This study addresses the question of coarticulatory behaviour of retroflexes by examining tongue shapes for Kannada (Dravidian) geminate stop /ʈ/ in two different vowel contexts.

2. Method

2.1. Participants

Ten native speakers of Kannada from Mysore, India participated in the experiment. Five of them were female and the other five were male, all in their early 20s. The participants reported no speech or hearing problems.

2.2. Materials

The materials included real Kannada words with labial /p/, dental /ʈ/ (IPA [ʈ]), and retroflex /ʈʈ/ occurring as geminates in two symmetrical vowel contexts, /u u/ and /i i/, as shown in Table 1. Each item was produced 10 times.

Table 1: Stimuli used in the study.

<table>
<thead>
<tr>
<th></th>
<th>u_u</th>
<th>i_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>labial</td>
<td>/muppʊ/</td>
<td>’age’</td>
</tr>
<tr>
<td>dental</td>
<td>/mutʊ/</td>
<td>’pearl’</td>
</tr>
<tr>
<td>retroflex</td>
<td>/muṭʊ/</td>
<td>’arrive’</td>
</tr>
</tbody>
</table>

2.3. Equipment

The participants wore an Articulate Instruments stabilizing headset (Scobbie, Wrench, and van der Linden 2008), with a PI 7.5 MHz SeeMore probe (Interson Corporation) fixed under the chin and connected via USB to a laptop computer. The probe depth was set to 10 cm. Videos of the sagittal cross-section of the tongue were recorded a sampling rate of 15 frames per second.

2.4. Analysis

Tongue contours during the consonant closure were traced using the Articulate Assistant Advanced software (AAA: Articulate Instruments 2011). Individual token tracings for each consonant were used to create mean splines, based on means at each of 42 fan lines. Pairs of mean splines (/u_u/ vs. /ippi/ vs. /uttu/ vs. /itti/, and /u_i/ vs. /ฏฏ/ vs. /itti/) were further evaluated using the function Diff. in AAA. This function compares the two splines by calculating a 2-tailed t-test (using the Welch-Satterthwaite equation) for each spoke and outputting a Root-Mean-Square (RMS) distance weighted by confidence. The resulting RMS distance values (1 per each consonant, 3 per speaker) were input into a Repeated Measures ANOVA with
the between-subjects factor Consonant (labial, dental, and retroflex) and the within-subjects factor Gender (female and male). Bonferroni post-hoc tests were used to evaluate differences between pairs of consonants.

3. Results

The results of the Repeated Measures ANOVA revealed a main effect of Consonant $F(2,16) = 126.916, p < .001$). A Bonferroni post-hoc test showed that RMS differences between the /u_u/ and /i_i/ contexts were greater for labials than for dentals and retroflexes (both $p < .001$); among the latter two consonants, differences for dentals were marginally higher than for retroflexes ($p = .059$). These differences can be observed in Figure 1, which plots means and standard deviations for /p/, /t/, and /ʈ/ splines in two vowel contexts, as produced by one of the speakers.

Figure 1: Sample splines, means and standard deviations for the tongue shape during the closure in (a) /uppu/ – /ippi/, (b) /uttu/ – /itti/, and (c) /uʈʈu/ – /iʈʈi/, speaker KM4.

Notice in Figure 1 that for all three consonants, the posterior tongue body is fronted and the anterior tongue body is raised in the /i/ context. Yet, these differences are much greater in magnitude for /p/ than for /t/ and /ʈ/. Also, the posterior tongue body is fronted in the /i/ context to a greater extent for /t/ than for /ʈ/.

The analysis also revealed a main effect of Gender ($F(1,8) = 9.872, p = .014$) and a significant Consonant * Gender interaction ($F(2,16) = 4.647, p = .026$). These indicated that RMS vowel context differences for dentals and retroflexes were smaller and not significantly different from each other for the female speakers, compared to the male speakers. Mean RMS differences and standard errors by consonant and gender are plotted in Figure 2. It can be seen that the two groups are very similar in their coarticulatory patterns with labials and somewhat different in the patterning of dentals and retroflexes, resulting in either two- (labial > dental/retroflex) or three-way displacement patterns (labial > dental > retroflex).

A closer examination of individual results revealed that the gender difference was largely due to the reverse patterns for dentals and retroflexes produced by two female speakers, KF4 and KF5. All the other female and all male speakers showed higher RMS difference values for dentals than retroflexes (see Table 2). A Repeated Measures ANOVA rerun without KF4 and KF5 produced a main effect of Consonant $F(2,16) = 113.433, p < .001$) and no significant Gender differences ($F(1,8) = 4.189, p = .087$) or the Consonant * Gender interaction. A Bonferroni post-hoc test showed significant differences for all pairs of consonants: labial vs. retroflex and dental (both $p < .001$), and dental vs. retroflex ($p = .013$).
Table 2: Mean RMS difference values for 3 consonants by speaker.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Gender</th>
<th>labial</th>
<th>dental</th>
<th>retroflex</th>
</tr>
</thead>
<tbody>
<tr>
<td>KF1</td>
<td>f</td>
<td>0.69</td>
<td>0.25</td>
<td>0.18</td>
</tr>
<tr>
<td>KF2</td>
<td>f</td>
<td>0.57</td>
<td>0.24</td>
<td>0.16</td>
</tr>
<tr>
<td>KF3</td>
<td>f</td>
<td>0.47</td>
<td>0.16</td>
<td>0.14</td>
</tr>
<tr>
<td>KF4</td>
<td>f</td>
<td>0.64</td>
<td>0.07</td>
<td>0.12</td>
</tr>
<tr>
<td>KF5</td>
<td>f</td>
<td>0.49</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>KM1</td>
<td>m</td>
<td>0.65</td>
<td>0.26</td>
<td>0.19</td>
</tr>
<tr>
<td>KM2</td>
<td>m</td>
<td>0.51</td>
<td>0.33</td>
<td>0.23</td>
</tr>
<tr>
<td>KM3</td>
<td>m</td>
<td>0.51</td>
<td>0.30</td>
<td>0.28</td>
</tr>
<tr>
<td>KM4</td>
<td>m</td>
<td>0.62</td>
<td>0.40</td>
<td>0.23</td>
</tr>
<tr>
<td>KM5</td>
<td>m</td>
<td>0.57</td>
<td>0.32</td>
<td>0.23</td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

The results of the study showed that the tongue body for all three consonants was more anterior and somewhat higher next to the high front vowel /i/ and more posterior and lower next to the high back vowel /u/. These results thus exemplified strong coarticulatory effects of adjacent vowels on consonant articulations. Importantly, however, the magnitude of these coarticulatory differences was significantly greater for the labial /p/ than for the two lingual consonants – the dental /t/ and the retroflex /ʈ/. Among the latter two consonants, coarticulatory differences were greater for the dental stop than the retroflex /ʈ/ stop, for all but two speakers. Male speakers showed overall a greater degree of coarticulation with lingual consonants, and more consistent differences between dentals and retroflexes.

Overall, the results of the study support the predictions of the DAC model of coarticulation (Recasens et al. 1997), as the involvement of the tongue body for Kannada consonants is inversely correlated with the degree of their coarticulation to high front and back vowels. Specifically, the lack of the lingual constriction renders the labial /p/ considerably more sensitive to vowel coarticulatory effects, compared to the consonants with the active involvement of the tongue tip, /t/ and /ʈ/ (cf. Recasens 1999; Farnetani and Recasens 2010, among others). While the front part of the tongue for the dental /t/ is constrained during the tongue tip closure at the upper teeth/alveolar ridge, the tongue body is relatively free and is subject to moderate vowel coarticulation. The situation is somewhat different for the retroflex /ʈ/: the tongue tip closure behind the alveolar ridge has to be accompanied by some lowering of the anterior tongue body and possibly some bracing of the tongue sides against the upper teeth (Narayanan, Byrd, and Kaun 1999 on Tamil /ʈ/; on Kochetov, Sreedevi, Kasim, and Manjula, submitted, on Kannada geminate /ʈ/). This would render retroflexes even less sensitive to vowel coarticulation than dentals (cf. Scobbie, Punnoose, and Khattab, in press on Malayalam /ʈ/). Our results largely confirm these predictions.

At the same time, the results of the current study seem to be in contrast with previous (electro-)palatography studies, which reported retroflexes to be more sensitive to vowel coarticulatory effects than dentals (Dave 1977; Dixit and Flege 1991; Krull and Lindblom 1996; Khatriwada 2007). The results obtained by two different methods, however, may not be incompatible. It is possible that that even a small displacement of the tongue body, as observed using ultrasound, can have a tangible effect on the location of the retroflex constriction, resulting in greater linguopalatal contact location differences observed using (electro-)palatography. This question should be further investigated using a combination of ultrasound and electropalatography.

5. Acknowledgements

The authors would like to thank Midula Kasim for invaluable assistance with the data collection and preprocessing, and Natalia Lapinskaya and Avery Ozburn for their help with the annotation and tracing of ultrasound videos. This work was supported by grants from All India Institute of Speech and Hearing and Social Sciences and Humanities Council of Canada.

6. References


Exploring nonlinear relationships between speech face motion and tongue movements using Mutual Information

Christian Kroos¹, Rikke L. Bundgaard-Nielsen²,³, Catherine T. Best³,⁴

¹Curtin University, Australia
²University of Western Sydney, Australia
³La Trobe University, Australia
⁴Haskins Laboratories, USA

cristian.kroos@curtin.edu.au, rikkelou@gmail.com, c.best@uws.edu.au

Abstract

In the current study, we used Mutual Information (MI) to determine the amount of shared information between face motion and tongue movements during speech. We tracked face motion using a passive marker-based motion capture system and measured tongue motion employing Electromagnetic Articulography. The results show widespread associations between the two types of motion, albeit with predominantly relatively low MI values. More importantly with regard to practical applications, a pronounced speaker variability was observed. We further investigated the temporal frequency distribution of the shared information using a wavelet decomposition yielding one-octave band-limited subbands. It was found that the lowest frequency subband contained substantially higher amounts of shared information.

Keywords: mutual information, auditory-visual speech, face motion, electromagnetic articulography, wavelets

1. Introduction

We have previously reported relatively low values of recovered variance when estimating tongue movements from face motion during speech using linear methods (Kroos, Bundgaard-Nielsen, and Best 2012, see also for a brief review of previous research). While the results were in line with human visual speech reading performance, they did not address the open question of whether the relationship between speech face motion and tongue movements might be, to a large degree, nonlinear. In the current study, we used Mutual Information (MI; developed within the framework of information theory) to determine the amount of shared information in the motion signals and investigated how it is distributed across one-octave band-limited frequency subbands. MI is able to detect and measure nonlinear relationships between input variables and is suitable for a wide range of signal types.

2. Method

2.1. Data acquisition

The tongue and jaw motion of three female speakers of American English (aged 22-28 years) were recorded using three-dimensional Electromagnetic Articulography (Carstens AG500), EMA hereafter. Three sensors were attached to the tongue and one to the gums at the front of the mandible. Two sensors at the mastoid processes of the left and right ears and two each at each side. Twenty-one half-spherical 3-mm markers were used to capture facial movements, predominantly on the left hand side since the face area on the right hand side was partially occupied by the wires of the EMA sensors (mouth corner and lower cheek). Seven face markers were placed around the vermilion border of the lips, 3 on the chin, 5 on the cheeks, 4 on the nose (wings, tip and bridge), and 2 at the right eye brow. Finally, three 9-mm spherical markers were sewn to a head band the speaker wore, allowing head motion tracking with the OPT system, too. Figure 1 shows a schematic with the target locations of the markers and the abbreviations for the marker names we will use throughout this paper.

2.2. Post-processing

After temporal synchronisation using the trigger signal from the AG500 Sybox, both types of measurement data were spatially aligned. Special trials in which four EMA sensors were wrapped with reflective tape (and thus becoming OPT markers) provided the global offsets between the EMA and the OPT coordinate systems. The speaker’s head was computationally stabilised using the method proposed in Kroos (2009), thus, the

at the maxilla registered head movements. Face motion was measured using the optical motion capture system Vicon (Vicon Industries, Inc), OPT hereafter. Four MX40 cameras were placed at two different height levels in front of the EMA cube (in the direction the speakers faced) and two each at each side. Twenty-one half-spherical 3-mm markers were used to capture facial movements, predominantly on the left hand side since the face area on the right hand side was partially occupied by the wires of the EMA sensors (mouth corner and lower cheek). Seven face markers were placed around the vermilion border of the lips, 3 on the chin, 5 on the cheeks, 4 on the nose (wings, tip and bridge), and 2 at the right eye brow. Finally, three 9-mm spherical markers were sewn to a head band the speaker wore, allowing head motion tracking with the OPT system, too. Figure 1 shows a schematic with the target locations of the markers and the abbreviations for the marker names we will use throughout this paper.
impact of rigid head motion was removed from the sensors and markers. Analyses of the residuals and the smoothness of the resulting head motion indicated that the OPT tracking using the head band markers was more reliable and it was employed to estimate head movements sample-wise. All data were rotated and translated accordingly with the occlusal plane as reference. Finally, the motion signals were downsampling to 50 Hz.

The face motion tracking data were cleaned manually frame by frame: spurious 'ghost' markers due to mistracking were removed and short-lived passages of tracking loss were interpolated using Vicon’s Woltring quintic spline filter.

From the large corpus recorded, we processed only the traditional children’s story 'Chicken Little'. The story was recorded divided into seven passages comprising about 6-9 sentences each, and the participants read them in a very lively manner. Since distribution-based measures like MI require larger number of samples for robustness, we concatenated all sentences of each speaker. Due to repeated episodes of measurement failure, the tongue dorsum sensor signal of speaker 2 had to be entirely excluded. All speakers included in this study were right-handed.

2.3. Analysis

In order to obtain band-limited frequency subbands, a multiresolution analysis (Mallat 1989) was applied using a discrete wavelet transformation (DWT) (Daubechies 1992). Wavelet transformations represent functions (or signals) in terms of 'small' base functions at different scales 

\[ f(t) = \sum_{s=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} c_{s,l} 2^{-\frac{s}{2}} \psi_{s,l}(2^{-s}t - l) \]

where \( c_{s,l} \) are the wavelet coefficients and \( \psi_{s,l}(t) \) the wavelet function.

We used filters corresponding to a biorthogonal scheme with cubic spline wavelets (see Sánchez, Prelic, and Galán 1996, for details about the algorithm) implemented in the Uvi_Wave wavelet toolbox for Matlab (The Mathworks, Inc.). A multiresolution analysis entails as the last step the application of a corresponding set of synthesis filters to transform the wavelet coefficients back into the original signal domain. The result is a set of signals band-limited to one octave (given the usual dyadic DWT) at different scales: the first one containing frequencies between the Nyquist frequency \( f_s \) and its half \( f_s/2 \), the second one between \( f_s/2 \) and \( f_s/4 \), and so on. We deemed four wavelet levels sufficient for our purposes, thus, we obtained four ‘details’ and an ‘approximation’ as follows: 

\( D_1: 12.5 - 25 \text{ Hz}; \quad D_2: 6.25 - 12.5 \text{ Hz}; \quad D_3: 3.13 - 6.25 \text{ Hz}; \quad D_4: 1.56 - 3.13 \text{ Hz}; \quad A: 0 - 1.56 \text{ Hz}. \)

Mutual Information is an information entropy-based measure which determines the reduction in uncertainty about one random variable given knowledge of another. It is defined based on probability density functions as follows:

\[ I(X;Y) = \sum_{y\in Y} \sum_{x\in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]

where \( X \) and \( Y \) are discrete random variables with joint probability distribution function \( p(x,y) \) and marginal probability distribution functions \( p(x) \) and \( p(y) \).

To be able to estimate the empirical probability distributions, the quasi-continuous movement measurement values had to be binned. Twenty bins yielded a sufficiently fine resolution to capture the essential characteristics of our data. We computed the amount of shared information for all combinations of OPT markers with EMA sensors treating their three coordinates as separate variables. There were 57 OPT (19 markers) and 9 EMA (3 sensors) variables and 513 pairings of them. Since serial correlations might lead to overestimation of mutual information, we randomly permuted the samples of successive local neighbourhoods \( (n_b = 8) \) over the entire length of the motion signals. Note that for the subbands the neighbourhood size was kept constant at 8 assuming sub-sampling of the subbands. As we did not apply sub-sampling, however (to obtain the same number of motion samples for each subband), we adjusted the neighbourhood value accordingly using \( n_s = 2^n_b \). Consequently at e.g., the lowest subband the neighbourhood comprised 256 samples. We repeated these randomisations 50 times. In order to obtain a baseline for statistical comparisons, we also computed the MI values for the same pairings with all samples of each signal randomly permuted, destroying all potential relationships between the signal pairs (50 repetitions).

3. Results and Discussion

Figure 2 shows the results for the three speakers as an image matrix with the value of each pixel assigned to a different colour component of an RGB image. Brighter patches indicate higher mutual information. See the figure caption for details. Many pronounced relationships can be found though there is substantial inter-speaker variability. The tongue tip sensor’s longitudinal component is strongly related with the chin marker’s longitudinal component which is, of course, expected, reflecting the congruent use of jaw opening and closing and tongue motion to vary tongue tip height. However, there is an even stronger relationship between the tongue tip sensor’s longitudinal component and the chin markers’ anterior-posterior component.

Table 1: Overall subband MI mean values.

<table>
<thead>
<tr>
<th>Subband</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.022</td>
</tr>
<tr>
<td>D2</td>
<td>0.016</td>
</tr>
<tr>
<td>D3</td>
<td>0.021</td>
</tr>
<tr>
<td>D4</td>
<td>0.026</td>
</tr>
<tr>
<td>A</td>
<td>0.064</td>
</tr>
</tbody>
</table>

The MI values resulting from the different subbands are shown in Figure 4 - 8. Note that for better contrast the subbands were normalised individually; their brightness values cannot be compared across subbands (but see our subband comparison below). The strong colouring reveals a striking speaker variability. For instance, subband D1 is dominated by speaker 2, i.e. only speaker 2 has a high number of relationships with strong MI in the higher frequencies (12.5 - 25 Hz). If the speakers were exhibiting similar MI values, the representing image elements would be on a grey scale between black and white. However, this appears to be limited to the very low range, values close to zero, i.e. no shared information, with only a few exceptions.
Figure 2: MI values coded in brightness for all relationships of the OPT markers (x-axis) and the EMA sensors (y-axis). Speakers are coded in constituent colours of an RGB image (speaker 1 = red, speaker 2 = green, speaker 3 = blue). Gray levels from black to white indicate equal MI values for all speakers. Red, green or blue patches show that an individual speaker stands out. Mixed colours reflect higher values for only two of the speakers, e.g., orange/yellow: speaker 1 and 2; magenta: speaker 1 and 3; cyan: speaker 2 and 3. Note that for speaker 2 the tongue dorsum data had to be entirely disregarded because of repeated episodes of mistracking. Triplets of rows correspond (from top to bottom) to tongue tip (TT), tongue dorsum (TD) and tongue back (TB), split within by spatial coordinates: \( X = \) lateral, \( Y = \) anterior-posterior, \( Z = \) longitudinal. Column triplets correspond from left to right to Chin Right (CHR), Chin Mid (CHM), Chin Left (CHL), Mouth Corner Right (MCR), Lower Lip Right (LLR), Lower Lip Mid (LLM), Lower Lip Left (LLL), Upper Lip Right (ULR), Upper Lip Mid (ULM), Upper Lip Left (ULL), Nose Nostril Right (NNR), Nose Nostril Left (NNL), Nose Tip (NTP), Nose Base Left (NBL), Nose Base Right (NBR), Cheek Right Lower (CRL), Cheek Right Upper (CRU), Eyebrow Right Lateral (ERL), and Eyebrow Right Central (ERC). Again the triplets are split within according to spatial coordinates: left = \( X = \) lateral, centre = \( Y = \) (anterior-posterior), right = \( Z = \) (longitudinal).

Figure 3: Histogram of all MI values (blue) with subband A superimposed (red).

The MI values for almost all relationships on all wavelet decomposition levels and the full signal were statistically significant (independent sample \( t \)-tests with Bonferroni adjusted \( \alpha = 0.05/8208 \)) for all speakers. However, the underlying differences are often small and might have been partially caused by head motion residuals, though head motion residuals were also small and certainly not apparent in visual inspection. Therefore, the abundance of these small differences can be assumed to truly reflect the interconnectivity of the entire orofacial system including the tongue.

Relationships that exhibit higher amounts of shared information might enable estimation of tongue movements from face motion using a non-linear estimator (e.g., nearest neighbour estimator with MI used for variable selection or artificial neural nets) in future work. However, the strong speaker variability - if confirmed with more speakers - would make practical applications difficult, to say the least.

4. References


Figure 4: MI values for all relationships for subband D1. For more details see the caption of Figure 2.

Figure 5: MI values for all relationships for subband D2. For more details see the caption of Figure 2.

Figure 6: MI values for all relationships for subband D3. For more details see the caption of Figure 2.

Figure 7: MI values for all relationships for subband D4. For more details see the caption of Figure 2.

Figure 8: MI values for all relationships for subband A. For more details see the caption of Figure 2.
Tongue-larynx interactions in the production of word initial laryngealization over different prosodic contexts: a repeated speech experiment

Leonardo Lancia1, Sven Grawunder2
1,2 Max Planck Institute for Evolutionary Anthropology
leonardo_lancia@eva.mpg.de, grawunder@eva.mpg.de

Abstract
We report the results of a repeated speech experiment conducted to elicit laryngealization or full glottal stops before vowel-initial pseudo-words. In order to study potential interactions between the larynx and the tongue, we tested the effects of different vowels (high vs. low) and the effect of the presence of stress on the initial vowel, as well as the effect of the place of articulation of adjacent stop consonants. The results obtained are consistent with the idea that a retracted tongue root favors the production of laryngeal constriction. Indeed we found that laryngealization is stronger on unstressed initial vowels, where it co-occurs with a retracted back of the tongue.

Keywords: laryngealization, word initial strengthening, wavelet-based functional mixed models, recurrence analysis, ultrasound tongue contours.

1. Introduction
Many aspects of the functioning of the larynx are still poorly understood. In part this is due to the fact that this is the product of complex interactions between the many heterogeneous anatomical structures composing the larynx (bones, muscles, soft tissues and cartilages). Also, the behavior of the vocal folds interacts with the behavior of the upper articulators through both biomechanical and aerodynamic linkages. In this paper we focus on laryngealized phonation, a phonation type obtained with adducted vocal folds resulting in less regular vocal fold vibrations, steeper spectral tilt, lower phonation type obtained with adducted vocal folds resulting in less regular vocal fold vibrations, steeper spectral tilt, lower

2. Material and methods
Although the presence of prosodically conditioned laryngealization in natural speech (as observable in the acoustic signal) is not systematic, it has been proposed that the underlying gesture may still be systematically present but not observable due to the nonlinearities characterizing the physical systems involved (Pierrehumbert & Talkin, 1992). It has also been observed that, in the repeated production of VC syllables, speakers of American English systematically produce a glottal constriction before the onset of voicing (de Jong, 2002). We adopted therefore this experimental paradigm to elicit the systematic production of the laryngeal constriction gesture and tested its effects on the surrounding vowels. We expected a blocking effect of stress on the incursion of laryngealization into the stressed vowel. We also tested the interactions with the quality of the surrounding vowels (/a/ vs. /i/), to better understand the relations between changes in the position of the tongue and changes in the indexes of laryngealization.

2.1. Procedure and participants
We asked 5 German female speakers to repeatedly utter, without interruptions and during 10 sec logatomes starting and ending either with the vowel /a/ or with /i/: /'iski/, /is'ki/, /a'ska/, /as'ka/, /'isti/, /is'ti/, /asti/, /as'ti/ at sustained speech rate. In addition to vowel quality and stress position we varied the consonantal contexts in order to observe more extreme tongue movements. The presence of the fricative was meant to prevent spontaneous verbal transformations during the repetitions (eg. from /'iti/ to /'ti:/). Sequences of control logatomes were obtained by replicating the internal plosive in final position (so that no glottal stops was expected: /kiski/, /'kiski/, /'kaski/, /'kas'ka/, /'kasti/, /'hasti/, /as'ta/). During the execution of the experimental task we recorded the acoustic signal, ultrasound (US) images of the tongue and the electroglosstographic signal (EGG). We expected less regular vocal fold vibration (as measured from the EGG signal) when a vowel was preceded or followed by a glottal stop or was produced with a constricted glottis.
2.2. Measurements

The following acoustic parameters were analyzed: F0, F1, F2 and intensity. The adopted measure of EGG-cycle regularity was obtained by applying an original version of recurrence analysis (RA). In the behavior of a dynamical system, a recurrence occurs each time the system is found in a previously observed state. RA permits to estimate the regularity of the behavior of the system underlying the behavior of a time series by counting the number of consecutive recurrences in its evolution over time (see Marwan et al. 2007 for an extended review of the method and of its variants). Before being submitted to the analysis, an observed one-dimensional time-series is transformed trough time-delay embedding in a multidimensional trajectory. This reconstructed trajectory shares many properties with the trajectory of the dynamical system underlying the observed time-series. Once the reconstructed trajectory is submitted to analysis, the percentage of detected recurrences with respect to the square of the length of the trajectory is called the percentage of determinism (%DET) and it is considered as a measure of regularity of the behavior of the system studied. Applying this method to successive time windows we can measure the regularity of the dynamical system underlying an observed trajectory and hereby its evolution over time. Unfortunately the original technique can be applied only to EGG signals in which F0 varies very slowly and, even in this case, modulations of F0 would affect the obtained %DET values. Lancia et al. (2013) proposed a variant of the original method insensitive to changes in the rate of change of the signal which does not rely on embedding of the observed time series and which allows to compute an elastic determinism measure (EDET) which is not sensitive to modulations of F0. However this approach is still sensitive to the number of cycles observed in each analysis window. A measure of regularity that is independent from the frequency of oscillations can be obtained by applying this version of recurrence analysis twice on the same signal. In the first application a recurrence is detected only when the observed trajectory takes a value which is very close to a previously observed value (according to the adopted metric). In a second application the similarity criterion to detect a recurrence is relaxed, so that much more recurrences are observed. The ratio between the EDET values computed with the two similarity thresholds is called elastic determinism ratio (EDET ratio).

2.3. Analyses

The measures were conducted by applying a sliding time window of 25 ms with a time step of 3 ms to each portion of signals containing the final vowel of a logatome and the initial vowel of the following logatome. From each obtained trajectory we removed the interval of time corresponding to the glottal closure and to the glottal stop explosion, when these events were distinguishable from the surrounding vowels. Vowel boundaries were deduced from the EGG signal. Ultrasound images submitted to statistical analyses were extracted at the points in time corresponding to left boundary of the final vowels of the logatomes. Pitch, intensity and formants were obtained from the audio signal with the Praat software. Custom MATLAB scripts were used to extract the EDET-ratio values. Before submitting each EGG signal to recurrence analysis this was band-pass filtered (cut off freqs.: 80 and 600Hz). The time varying RMS of the filtered signal was computed on sliding hamming windows (length: 25ms step: 3ms). It amplitude was rescaled so to vary between 0 and 1. Also its length was modified trough linear interpolation to match the length of the original EGG signal. Finally the filtered signal was multiplied point-wise by the scaled intensity. This preprocessing step helped to reduce the variability due to changes of amplitude of the EGG oscillations observed at the edges of the vowels regardless of the presence of glottalization. At the end of the analysis process we obtained for each sequence of vowels a set of trajectories describing the evolution over time of formants, pitch, intensity, and EDET ratio. Statistical analyses were conducted separately on each parameter with a wavelet based mixed model (Morris and Carroll, 2006). In this approach a wavelet transform is applied to the observed trajectories so that each trajectory is represented by a set of orthogonal wavelet coefficients. These coefficients are considered as dependent variables in a multivariate Bayesian Mixed Model. The following parameters were considered: vowel, consonant, stress position, presence vs absence of initial consonant and position of the word in the sequence to repeat. For each model we started by testing all the possible 3-way interactions between the first four factors. A maximal random effects structure was adopted (with the exclusion of the terms for the correlations between random intercepts and slopes). Tongue images were collected by means of the Terson T3000 ultrasound device connected to a PC and synchronized with the audio signal by using the Ultraspeech acquisition software (Hueber et al., 2008). The ultrasound probe was fixed to the head of the participants by means of a head fixation helmet produced by Articulate Instruments. In order to test hypotheses on tongue positions we used the same kind of wavelet-based Bayesian mixed models as used for the trajectory analyses (Morris et al. 2010). However in the case of ultrasound data, statistical modeling was conducted separately on each speaker and images were submitted to 2D wavelet transform. Wavelet transforms and Bayesian mixed models were applied trough the WMFF software (Morris and Carroll, 2006).

3. Results

For lack of space we present here only the main results obtained for %EDET, for F1 and for the tongue shape (only US data from two speakers will be discussed). Also we will not comment on differences observed at the onset of the final vowels and at the offset of initial vowels (i.e. differences observed far away from the portion of signal where laryngealization is expected).

3.1. Regularity: EDET ratio

In both vowel-initial and consonant-initial logatomes, the regularity of voicing decreases near the intervocalic boundary\(^1\). However this decrease is stronger for logatomes initiated by vowels (Fig. 1, panel a, effect 1). The irregularity extends over a larger region of the unstressed vowel in both kinds of logatomes (panel c, effect 1). The unstressed position has a stronger effect on the first syllable, which becomes less regular all along its duration, this is not the case in consonant-initial logatomes (panel c, effect 3). Regardless of the presence or not of an initial consonant, at the very juncture between logatomes /l/ induces higher irregularity, while increasing regularity during the long offset of the first last syllable (panel b, effect 2). In vowel-initial logatomes, the combined effects of the consonant /l/ and of the vowel /l/ lead to an increase of regularity around the intervocalic boundary.

\(^1\)Although we refer to this boundary as intervocalic, in the many cases in which an oral consonant or a full glottal stop occurred at the juncture of consecutive vowels, the boundary became intervocalic only after the removal of the signal portion that separated the two vowels.
especially on the side of the initial vowel (regularity increases faster at the onset of /isti/ than at the onset of /aska/ panel b, effect 2). With stress on the second syllable vowel /i/ has the effect to reduce the regularity on the most stable portions of the vowels (panel c, effect 3).

3.2. F1

Around the intervocalic boundary F1 is higher in the repetition of vowel-initial logatomes (Fig.2, panel a, effect 1). This difference is reduced when the vowel is an /i/ (panel a, effect 2). Note that in vowel-initial logatomes F1 is always higher at the onset of the initial vowel, while in consonant-initial logatomes this is only true for logatomes containing /i/ (panel a, effect 3). Initially stressed logatomes containing /i/ have a higher F1 in the final vowel (panel b, effect 1). And in general, logatomes starting with a consonant show also higher F1 values at the offset of the final vowel (panel b, effect 2). However this effect of the consonant /t/ on F1 is reduced if the sequence contains /i/ (panel b, effect 3). The effect of word-initial stress on F1 in the second syllable of vowel-initial logatomes is reduced when the consonant is a /t/ (panel d, effect 1).

3.3. Tongue shape

In Figure 3 we display the predictions of the statistical models run on data from two speakers (one speaker per column). For the sake of space we show here only the predictions for vowel /a/ in logatomes containing the consonant /t/, with stress on the first or on the second syllable (top panels vs. bottom panels) and starting with (or without) an oral consonant (top vs. middle panels). Red contours indicate those regions that are significantly different in logatomes starting with an oral consonant vs. a vowel (middle vs. top panels) or with stress on the first vs. the second syllable (bottom vs. top panels). Significant regions not relevant for the identification of the tongue contour (e.g. due to differences in tissue density) were removed for the sake of clarity. A contour indicates that different intensities were observed between conditions in the corresponding region of the images, because the tongue moved in or out of the region. Regions I and III show that before an utterance with initial consonant the tongue is slightly more advanced than before a utterance initial vowel. This difference was not observed for this speaker when the consonant was a /k/ or when the vowel was an /i/ (not shown here). Regions II and IV show that before utterances starting with /t/ the front part of the tongue is more elevated. Regions V and VI indicate the movement of the shadow casted by the hyoid bone, which for this speaker is higher before unstressed initial syllables. This is also true before initial syllables starting with an oral
consonant (not shown here). Region VII indicates additional retraction of the tongue back before vowel initial logatomes with the stress on the second syllable. Regions VIII and IX indicate that, before stress final utterances, the back of the tongue is more retracted. This speaker does not show retraction before consonant initial utterances.

Figure 3: Tongue images at the very offset of the final vowel /a/ in logatomes containing /t/ as estimated trough statistical modeling. Top: vowel-initial logatomes with stress on the first syllable (baseline context). Middle: consonant initial logatomes with stress on the first syllable. Bottom: vowel initial logatomes with stress on the final syllables. The red contours indicate regions where the two contexts result in significant differences from the baseline.

4. Discussion

This pattern of results indicates that the effect of any consonant on adjacent vowels is to reduce regularity, but that this effect is much stronger for a glottal stop. Also, the most fronted tongue configuration (observed when /l/ is produced in the context of /t/) limits the degree of irregularity. The behavior of F1 mimics in some respect the behavior of EDET ratio. However not surprisingly F1 is more sensitive to the nature of the vowels. The dip of F1 that is observed at the juncture between the utterances with consonant-initial logatomes is due to the coarticulatory effect on the width of the oral constriction. This interpretation is consistent with the fact that the decrease is smaller in /a/ than in /l/ and with the effect of stress, which raises F1 more in /a/ than in /l/ (Fig. 2, panel c, effect 3). Indeed the oral constriction in /l/ is wider than in the surrounding consonants, whose presence produces therefore a stronger lowering of F1 near the intervocalic boundary. Higher F1 near the intervocalic boundary in stressed /l/ vs. stressed /l/ can therefore easily be explained as due to constraints on coarticulatory effects that are imposed by stress. This F1 dip is completely absent in logatomes with initial vowels, which can be interpreted as evidence of a retracted tongue and of a narrowing of the pharynx during glottal stop production.

5. Conclusion

In summary we found effects of laryngealization before vowel initial logatomes and that these effects are modulated by the presence of stress which naturally blocks the spread of laryngealization inside the vowel. Also we found evidence that the effect of laryngealization is weaker when the tongue is more fronted (e.g. in /isti/) and that laryngealization is produced with a retraction of the root of the tongue, a position which is more compatible with vowel /l/ than with vowel /l/. These two findings suggest that the preference for low/back vowel contexts is due to biomechanical linkages between tongue root and larynx.

6. References


Studying MRI acquisition protocols of sustained sounds with a multimodal acquisition system

Yves Laprie¹, Michaël Aron², Marie-Odile Berger¹, Brigitte Wrobel-Dautcourt¹

¹INRIA/CNRS/UL, LORIA, Nancy, France
²ISEN, Brest, France
Yves.Laprie@loria.fr

Abstract

The acquisition of dynamic articulatory data is crucial to improve our understanding of speech production. Ultrasound (US) imaging presents the interest of offering a good temporal resolution without any health hazard and at a reasonable price. However, it cannot be used alone because there is no reference coordinate system and no spatial calibration. We describe a multimodal acquisition system which uses electromagnetography sensors to locate the US probe, and the method used to calibrate the US modality.

We experimented this system to investigate the most appropriate acquisition protocol for Magnetic Resonance Imaging. Three strategies were explored for one speaker: (i) stopping phonation just before acquisition, (ii) reiterated acquisitions, and (iii) silent acquisition. The measurements of the minimal tongue to palate distance show that silent acquisition generally offers the minimal articulatory drift while guaranteeing a small tongue to palate distance, i.e. a clear articulation.

Keywords: articulatory data, MRI, ultrasound imaging, multimodality

1. Introduction

Technical advances in acquiring articulatory data have often conditioned scientific breakthroughs in speech production modeling. Very substantial advances have been achieved in the acquisition of static articulatory data. By offering a millimetric spatial resolution of images, 3D MRI images of the vocal tract have enabled more accurate evaluations of vocal tract acoustic modeling (Valdés Vargas, Badin, and Lamalle 2012).

On the other hand, the acquisition of dynamic geometric articulatory data still represents a challenge. Despite its recent emergence, real time MRI (Narayanan et al. 2004; Bresh et al. 2008) is not widely available because of its cost. Additionally, spatial resolution of images is still rather poor and the recording conditions (noise of the machine and supine position) alter speech production. Beside its innocuousness, the advantage of electromagnetography (EMA) is to offer a sufficiently high sampling frequency (400 Hz for the most recent machines) to analyze all speech articulatory gestures. The main weaknesses are the small number of sensors that can be tracked simultaneously, the minimal distance to respect between two sensors to avoid aberrant measures due to magnetic interferences, and the risk of perturbing articulation with wires connecting sensors to the articulograph. On the other hand ultrasound imaging presents some interesting advantages. It is a widely available technique, cheap, offering a good temporal sampling (between 50 and 100 Hz when imaging the vocal tract), and producing an acceptable level of acoustic noise. Data are recorded in the coordinate system attached to the probe.

In order to register tongue contours in a reference coordinate system it is necessary to track the position of the ultrasound probe to get its position during the acquisition. This solution has been chosen by the designers of the HOCUS system (Whalen et al. 2005). Infrared sensors are fixed onto the ultrasound probe, and on glasses attached with an elastic band so that they cannot move relative to the subject’s head. These sensors are tracked via the Nothern Digital Optotrack system. Actually, fleshpoints behind the ears are probably less mobile but the nature of the Optotrack system which utilizes infrared emitting diodes (IREDS) requires that the sensors to be visible from the cameras. The designers of the HOCUS system preferred not to add probe immobilization device so as to avoid the apparition of spurious articulatory compensation gestures. There is thus no guarantee that the probe remains in the mediosagittal plane. This setup gives the position and the orientation of the probe and head in the optical coordinate system. However, the plane of the ultrasound image cannot be known and designers thus added three sensors onto the US probe so that the plane defined by these points approximately coincides with that of the ultrasound image.

There is thus no geometric calibration since the location of a point in the ultrasound image cannot be calculated in the optical coordinate system. The geometrical transformation between sensors glued on the probe and the ultrasound image is only estimated by hand with the involved inaccuracy. Beyond this inaccuracy, it is not possible to know the position of one point of the ultrasound image in this plane since the distance between the probe and this point is unknown even if the resolution of the ultrasound machine is provided by the manufacturer.

It is thus impossible to merge US images acquired with the HOCUS system with images of the vocal tract acquired with another acquisition modality, MRI for instance. We thus developed a multimodal system (Aron, Berger, and Kerrien 2010), which combines ultrasound imaging and electromagnetography. Ultrasound imaging allows tongue tracking while electromagnetography is used to track the position of the ultrasound probe and enables the registration of ultrasound with other modalities, here MRI for instance. One of the main contributions is the spatial calibration of the imaging modalities, especially ultrasound, with respect to the modality used to merge all data, here electromagnetography.

The calibration of an imaging modality with respect to another is generally obtained by considering points visible in both modalities. In our case, the direct calibration of both modalities is impossible because electromagnetic sensors are invisible in ultrasound images. It is thus necessary to design an object,
called phantom, whose geometrical properties are known and which is easily detectable in both modalities. Different techniques were tested in the literature for the US/EM spatial calibration (Mercier et al. 2005), using different kinds of phantoms: cross-wire with a single or multiple point targets, three-wire phantoms, Z-fiducials, wall phantoms. Each design has advantages and disadvantages in terms of easiness of use, accuracy, and precision. There is no agreement about the best phantom design. The phantom we designed is inspired from that of (Khamene and Sauer 2005). Two 5DOF sensor coils were fixed at both extremities $P_1$ and $P_2$ of a rigid wood stick approximately 25 cm long and 3 mm on diameter (as seen on Figure 1).

![Experimental setup used for ultrasound calibration.](image)

This line segment whose equation is known from the electromagnetic sensors fixed at extremities is easily detectable in the US images. The line pointer and the transducer were immersed into water at the room temperature, and thirty images corresponding to different positions and orientations of the transducer were acquired. In each US image, the pointer appeared as an ellipse whose center was manually selected. Such an ellipse was often larger than 10 pixels and noisy. The uncertainty due to noise was overcome by taking a large number of calibration images. Because experiments were made with water at ambient temperature ($20 \degree C$), the speed of sound in this medium is different from the one in human tissue ($\approx 1540 \text{m/s}$). Bilaniuk (Bilaniuk and Wong 1993) showed that the speed of sound in water at $20 \degree C$ is $1485 \text{m/s}$. Therefore every point in US images was corrected by shortening its depth (distance to the US focal point) by a ratio of $1540/1485 \approx 1.04$.

Since the first developments (Aron, Berger, and Kerrien 2010) the system has been substantially modified: (i) it utilizes the new NDI Wave system which offers a better sampling frequency, 100 Hz instead of 50 Hz in the previous system, (ii) it also offers a better calibration of the ultrasound modality and (iii) provides a simpler and, therefore more robust, synchronization system which requires only one PC used to supervise the acquisition system. One of the advantages provided by the calibration of the ultrasound modality is to enabling the fusion with MRI data since the 3D position of points in the ultrasound images is expressed in the electromagnetic coordinate system. MRI images are recorded beforehand and are fused with ultrasound data. This enables the calculation of the minimal distance between the tongue and the palate.

2. Experiment to assess MRI acquisition protocols

We report here first acquisition experiments aiming at studying protocols used to acquire static 3D MRI images of the vocal tract. We were particularly interested in the consequences of stopping phonation during the acquisition. Indeed, subjects are traditionally asked to maintain the same articulation, i.e. keeping all articulators as motionless as possible even if they are obliged to stop phonation during the acquisition. Other solutions consist of stopping phonation just before the acquisition starts, repeating several short acquisitions synchronized with acquisition by identifying voice activity or even using silent speech.

A good acquisition protocol should guarantee a steady position of the tongue to get good quality MRI images together with relevant positions of the tongue. We evaluated the tongue position by measuring the minimal distance between the tongue and the palate. Indeed, we suspected that the tongue lowers as soon as phonation stops, and this distance gives interesting information about the vowel quality because too big a distance corresponds to a centralization.

The tongue contour is visible in ultrasound images and palate surface in MRI images. Both modalities have been merged as explained above. Concerning the impact of acquisition strategies our system presents the strong advantage of not requiring any sensor to be glued onto the tongue like traditional EMA which is likely to alter articulation (Katz, Bharadwaj, and Stettler 2006). Here EMA is only used to track the US probe and to cancel head movements. This guarantees minimal perturbation of the tongue movements.

We investigated three conditions:

1. The speaker starts phonating the sound and then stops after 5 seconds while maintaining the same articulation for approximately 10 seconds,
2. The speakers phonates the vowel for 3 seconds approximately, stops phonation and starts again phonation. He reiterates this scheme during the 15 seconds of the acquisition and is allowed to take breath.
3. The speaker silently articulates the vowel for 15 seconds.

The four vowels /u, i, a, y/ have been recorded with this protocol. The sampling frequency of ultrasound was set to 65Hz and that of the wave system to 100 Hz. Each acquisition produced an ultrasound film of 975 images. The tongue contour was delineated by hand every 5 images, and every 2 images when the tongue movement is fast. Due to the lack of space, we only inserted figures for /u, i, a/. Figures 2, 3, 4 show the three articulation strategies for each vowel. They enable the comparison of the strategies from the stability point of view. The three vowels exhibit the same tendency but with marked differences in terms of relative amplitudes. The strategy of stopping phonation always gives rise to an articularatory drift, i.e. an abduction movement of the tongue, with a marked fast transition for /a/ and almost a linear transition for /u/. For these two vowels the distance to palate increases between 3 and 4 mm. On the other hand, the drift is smaller for /a/, about 1 mm, but the initial distance was also bigger at 8mm and the main articulatory characteristic is the narrow tube in the pharyngeal region.

At the other extremity, the strategy of silent speech gives rise to a limited drift within approximately a millimeter, and slightly more for /a/. Indeed, unlike /u/ and /a/ for which the tongue can be maintained in the same posture because it can contact molars (/u/) or the mouth floor (/a/) /a/ requires a more
substantial effort without the contribution of teeth or mouth floor to adjust the posture. This is likely to explain the slight abduction movement observed for silent /u/. The reiterating phonation gives rise to strong oscillations of the tongue in both directions, i.e. very low minimal distances between tongue and palate, but also large values of this distance when phonation stops. This tendency is not as marked for /a/ probably because mouth opening necessary to produce /a/ corresponds less to the movement of the tongue than that of the mandible, which has a greater inertia. Depending on the vowel the amplitudes of oscillations of the reiterating strategy are well between these two extremities for /a/ reach the values of the stopping and silent strategy for /u/ and exceed them in the case of /i/. In the latter case it should be noted that the articulatory drift accompanying the stopping strategy is not complete at the end of the recording. It is likely that the maximal value of the oscillations is the limit of this drift.

We also evaluated the horizontal movement of the constriction during acquisition. A remark should be made about this measure. The horizontal position of the highest point of the tongue contour drawn on ultrasound images is not very precise because the constriction of a vowel is not very strong and therefore extends over a fairly large part of the palate and also because the highest point is generally located in a flat part of the tongue contour. We thus only give this measure for the stopping strategy. As shown by Figures 5, 6, 7 the constriction location moves backwards approximately 5 mm for /u/. This movement coincides with the abduction tongue movement.

This experiment was intended to choose the best strategy for acquiring MRI image for one specific subject. Before this experiment we expected that the reiterating strategy could favor the realization of articulatory targets close to those of sustained phonation. To some extent this strategy avoids the abduction movement of the tongue, with the counterpart that it is difficult to realize the same target several times without the help of an imposed phonetic context. Besides, the amplitude of the articulatory effort necessary to restart phonation is another explanation and could trigger a strong movement of the tongue upwards which blurs the results.

The silent speech strategy turns out to give a good stability of the tongue even if there is no guarantee that the articulatory configuration corresponds to that of the target vowel. However, it can be seen that the minimal distance between the tongue and palate is close to that of phonation. In addition, the horizontal movement of the maximal constriction seems to be sufficiently small to ensure that the tongue shape is reasonably close to that of the phonated vowel. This led us to keep the silent strategy with prior phonetic training to guarantee the relevancy of vocal tract shapes recorded with MR imaging.

3. Conclusion

This first experiment with the new version of our multimodal system shows that acquisition protocols used in MRI should not involve changes in phonation, i.e. stopping phonation when the acquisition lasts too long, or reiterating articulatory gestures. We will now test other speakers in order to compare different strategies that can be used for MRI acquisitions of static vocal tract shapes. Three-dimensional data require an acquisition duration of approximately 15 seconds. By taking into account the additional time required for the subject to reach the target before the noise of the MRI machine starts it is difficult to sustain phonation up to the end of the recording. It seems that creaky voice, as proposed by...
4. References


Geometric articulatory model adapted to the production of consonants

Yves Laprie¹, Béatrice Vaxelaire², Martine Cadot¹

¹ INRIA/CNRS/UL, LORIA, Nancy, France
² IPS, Strasbourg, France

Yves.Laprie@loria.fr

Abstract

This work deals with the construction of articulatory models which can be easily adapted to a new speaker and enable a better approximation of tongue contours corresponding to consonants. Data used are three corpora of X-ray films. The first corpus was used to construct an articulatory model and design an adaptation procedure. The evaluation carried out on the third corpus shows that this adaptation performs well. Geometric fitting provided by the first model was often insufficient in the region of the consonantal places of articulation of the second corpus. Tongue contours delineated from X-ray images were thus corrected by considering virtual articulatory targets and the weight of consonants was increased in the Principal Component Analysis (PCA). Furthermore, the coefficients of the linear components are not calculated by projecting contours onto the PCA base vectors, but with an optimization procedure so as to guarantee a good approximation in the constriction region of consonants.

Keywords: articulatory models, articulatory synthesis, consonants, adaptation, X-ray images

1. Introduction

Articulatory synthesis is a valuable means of analyzing speech production because it provides a link between the vocal tract shape and the speech signal, and thus enables the origin of acoustic cues and the impact of articulatory gestures to be investigated in parallel. This first requires the shape of the vocal tract to be approximated. Many works have been dedicated to this topic and gave rise to geometrical models controlled by a small number of parameters, for instance that of Maeda (1990) which uses seven linear components derived from a corpus of X-ray images via guided principal component analysis. Despite its interest, this model was mainly designed to analyze vocalic structures VCV and VCCV are within an identical vocalic context /aCa/ or /aCCa/. These data were recorded by a French female speaker at a frame rate of 50 fps. 15 sentences have been proposed up to now. This represents a total number of 1050 images (374x466 pixels because only the rectangle corresponding to speech, i.e. 672 images, were considered.

The second corpus, called C2 and recorded in the eighties, comprises 58 short sentences formed of 4 to 6 syllables. Unlike the first corpus each sentence has a meaning. The corpus proposes a number of VCV or VCCV sequences. Both syllabic structures VCV and VCCV are within an identical vocalic context /aCa/ or /aCCa/. These data were recorded by a French male speaker at a frame rate of 50 fps. 15 sentences have been processed (256x256 pixels) but only images corresponding to speech, i.e. 672 images, were considered.

The third corpus, called C3, is the historical corpus used by S. (1979). It comprises 10 short sentences formed of 4 to 6 syllables and contains all the phonemes of French. These data were recorded by a French female speaker at a frame rate of 50 fps. This represents a total number of 520 images, or more precisely of articulatory contours delineated images since original images are not available.

As it can be noticed, the three corpora were designed with very specific objectives. Furthermore, their size is very limited. Hence, these corpora are not phonetically balanced, and more critically with respect to our objective about the construction of models covering the articulatory variability of French some phonemes are absent or in a very limited number. Despite these weaknesses, the size, the coverage of the entire vocal tract, the quality of images, the three speakers, and its dynamic character compared to MRI images make these corpus a very valuable articulatory resource.

2. Description of data

X-ray data used in this work comprise three corpora. The first, called C1 and recorded in the nineties, was initially designed to study coarticulation in French. It comprises four films. The first two are a series of six short sentences ranging from /se dø si ylt/ to /se dø sikst ltrR/ (each sentence contains one non-labial consonant between /l/ and /r/ than the previous one) at normal and fast speech rates. The last two are a series of /VCV/ /aku iku aku atu itu utu/ at normal and fast speech rates. These data were recorded by a French male speaker at a frame rate of 25 fps. In total, this corpus comprises 946 images (256x256 pixels) but only images corresponding to speech, i.e. 672 images, were considered.

In parallel, we recently conducted experiments on articulatory copy synthesis from X-ray films (Laprie, Loosvelt, et al. 2013) of the DOCVACIM corpus. The input data consisted of the vocal tract shape (from the larynx to the lips) given by the contours delineated from X-ray images and the temporal segmentation of the speech signal in speech sounds. We exploited the time coordination plans between the source and the vocal tract proposed by Maeda (1996). Speech resynthesized from the X-ray images and covering both vowels and consonants sounds very correct even if there is only one image every 40 ms. However, this approach presents two weaknesses: (i) full contours have to be provided. (ii) the transition from one contour to the next is not possible easily, and thus area functions were used instead and interpolated from one image to the next. A better solution would be to specify the vocal tract shape via parameters of an articulatory model. This work deals with the construction of articulatory models and focuses on the aspect of model adaptation and construction strategies so as to obtain a model which approximates vocal tract shapes of consonants successfully.
3. Model adaptation

The articulatory model corresponding to the corpus C1 was constructed following the strategy presented by Laprie and Busset (2011). This model uses the jaw opening as a main mode of control which gives rise to one or two linear components according to the expected quality of fitting with original data. Linear components are obtained via Principal Component Analysis (PCA). It should be noted that the rotation and the translation of the mandible are taken into account, and not only the translation of the lower central incisor as often in other articulatory models. The movement of the mandible is subtracted from articulators linked to the mandible, i.e. the tongue and the lower lip. We chose to subtract the geometrical movement of the mandible rather than to remove the correlation between the mandible and tongue because we wanted to make as few assumptions as possible on the link between the mandible and the tongue. PCA is then applied to the curvilinear contour of the tongue to obtain between 4 and 6 linear components.

Similarly, lip deformations are represented by two linear components. Finally, the larynx and epiglottis are represented by one linear component. The epiglottis, which is essentially a passive articulator since this is a cartilage, is submitted to the movements of the tongue particularly when the tongue moves backwards. We thus added a collision algorithm which detects contacts between the tongue and the epiglottis and pushes it if need be. The reconstruction error of the tongue with 6 linear components is 0.51 mm.

It is well known (Vorperian et al. 2009; Lammert et al. 2011) that the origin of anatomical variability are the length of the mouth and pharynx cavities, together with their relative orientation. Our adaptation thus takes into account the rotation of the mouth cavity around the upper incisor, the scale factors in the directions of the mouth and pharynx cavities, and the angle formed by the mouth and the pharynx.

The first transformation consists in an non isotropic homothety and a rotation intended to adjust the mouth angle and the scale factors of the mouth and pharynx. The coordinates \( x' \) and \( y' \) of the point transformed are given by:

\[
\begin{align*}
    \begin{bmatrix}
        x' \\
        y'
    \end{bmatrix} &= 
    \begin{bmatrix}
        \alpha_x \cos \theta & -\alpha_y \sin \theta \\
        \alpha_x \sin \theta & \alpha_y \cos \theta
    \end{bmatrix}
    \begin{bmatrix}
        x - x_{UI} \\
        y - y_{UI}
    \end{bmatrix}
    + 
    \begin{bmatrix}
        x_{UI} \\
        y_{UI}
    \end{bmatrix}
\end{align*}
\]

where \( \alpha_x \) is the scale factor of the mouth cavity, \( \alpha_y \) that of the pharynx, and \( x_{UI} \) and \( y_{UI} \) the coordinates of the upper incisor.

The second transformation consists in rotating the pharyngeal cavity. In order to focus the pharynx and not to affect the rest of the vocal tract the rotation decreases when moving from the pharyngeal wall to the front of the mouth cavity, equals zero above a line (the red line (C, UI) in Fig.2) formed by the upper incisor (point UI in Fig.2) and a point located at the top of pharynx (point C) and increases from this line to the pharynx. The angle \( \theta \) of the rotation applied to an original point \( P \) is thus a function of the projections of this point onto the line (C, UI) and its perpendicular through C.

This adaptation is thus purely geometrical and introduces some incorrect warpings in the tongue shape since according to the tongue articulatory parameters one fleshpoint may be affected or not by the rotation applied to the paryngeal cavity. However, it enables a good fitting with all the MRI and X-ray images we have at our disposal. A more anatomical based adaptation procedure would require anatomical data (provided by MRI or X-ray images) for many speakers to derive adaptation strategies.

The evaluation of this adaptation procedure has been carried out on the corpus C3. The model has been adapted to the speaker the via the procedure described above.

Fig. 1 shows an example of fitting with 8 linear components for a /u/. It can be noticed that the front part of the tongue recovered by the model is probably more realistic than the contour drawn from the X-ray image. Unfortunately, the exact resolution of the original images is not known. The pixel size has been estimated to 0.5 mm by considering that the vocal tract length for this female speaker was close to 16 cm. The reconstruction error is approximately 0.560 mm.

Table 1 gives the average reconstruction error as a function of the number of linear components used to approximate the tongue contour.

It can be seen that the model approximates the shapes of the tongue very well since the reconstruction error is only slightly higher than for the original speaker. However, two remarks should be done. First, the contours of the corpus C3 used for evaluation do not cover the sublingual cavity. The precision would not probably have been as good if the sublingual cavity has been considered. Secondly, and it is probably more important, the number of images corresponding to consonants is small and the contours are sometimes not complete. The evaluation is thus probably more optimistic than it should be.

4. Building models adapted to consonants

Since we were interested in resynthesizing speech from X-ray films we initially tried to used the vocal tract shapes approxi-

![Figure 1: Fitting of the tongue model.](image-url)
mated by the articulatory model. This would have allowed coarticulation models to be designed and evaluated. However, this model cannot be used for copy synthesis since the geometrical precision in the region of the constriction is not good enough, especially when there is contact between the tongue and the palate. The construction of articulatory models adapted to consonants raises several issues. The first is related to the nature of contours used to derive linear components. When dealing with vowels there is no contact between the tongue and other fixed articulators (palate, teeth). Factor analysis used to determine linear modes of deformation of the tongue only takes into account the influence of the tongue muscles. This is no longer the case with consonants, since a contact is realized between the tongue and the palate for stops /k, g, t, d/ and the sonorants /l, l/ in French. The deformation factors thus incorporate the “clipping” effect of the palate. When approximating a shape presenting a contact with the palate, the articulatory model undergoes difficulties to render this contact, and rather generates a smooth shape with only a punctual contact with the palate as illustrated by Figure 3.

4.1. New strategy of model construction adapted to vowels

Following the idea of using virtual articulatory targets (for instance, Birkholz, Kröger, and Neuschaeter-Rube (2011)) that lie beyond the positions that can be reached, here the palate, we edited delineated tongue contours presenting a contact with the palate. We chose a conservative solution which consists of keeping the tongue contour up to the contact point and extending it while guaranteeing a “natural shape”. These new contours do not cross the palate for more than 10 mm.

As such, this first modification alone is not sufficient, because the number of images corresponding to consonants is small even if the corpus used in this work is phonetically balanced. 1015 images have been annotated carefully and used to construct the articulatory model by applying the strategy presented by Laprie and Busset (2011). Preliminary investigations showed that the contribution of tongue contours of /l/ is essential, because they exhibit a very marked tongue tip. We thus duplicated a number of /l/ X-ray images in order to increase the weight of deformation factors corresponding to the tongue tip. Factor analysis (Principal Component Analysis) constructs a base of vectors used as linear deformation modes. Tongue contours are approximated by projecting the vector of tongue points onto the base vectors. This gives the best possible fitting. This approach is no longer possible since the tongue shape may cross the tongue palate, and is thus clipped by the palate contour when analyzing real tongue contours delineated from X-ray images. It is therefore not possible to project the tongue contour to get the best fitting. The contribution of each linear deformation vector was thus obtained via optimization. Since the objective is to achieve a good geometrical fitting at the place of articulation, the contribution of points near the constriction with the palate is increased by incorporating appropriate weights in the numerical criterion to be minimized. Additionally, the fitting criterion takes into account the whole tongue contour including the sublingual cavity, which plays an important acoustic role in the acoustics of consonants.

This approach has been tested on the X-ray images of the corpus C2. It requires more components than models build on the corpus C1 but provides a very good fitting with original tongue contours, i.e. 0.830 mm in average with 6 components over the whole tongue contour and only 0.567 mm in the region of the main place of articulation.

5. conclusion

The first part of this work shows that a model constructed for one speaker can be easily adapted to fit the vocal tract shapes produced by a second speaker. Even if this evaluation was carried out for two speakers (one male and one female) it is likely that it could be applied successfully on other speakers since it is fairly simple and general. Hence, the adapted model can be used to synthesize speech for the new speaker. On the other
hand, articulatory models are often constructed and evaluated on corpora which contain a small number of consonants only. The deformation modes derived from these corpora are thus unable to approximate the vocal tract shapes for consonants.

The evaluation results show that more linear components are required to reach a good fitting in the case of consonants. This seems quite normal since the tongue has to realize a contact at a very precise place on the palate. The current version of the fitting procedure gives equal importance to both sides of the constriction, i.e. the front and the back cavities. It would probably be possible to increase the precision at the front cavity by decreasing the precision imposed at the back cavity to further improve the acoustic properties of synthetic speech.

Even if the adaptation method proposed can easily be applied to the last model, the articulatory coordinates, i.e. the weight of each linear component, are speaker dependent because they are also related to the palate shape. We will now connect the acoustic simulation with the articulatory model instead of the vocal tract shapes derived directly from X-ray images (Laprie, Loosvelt, et al. 2013). Beyond articulatory synthesis, this will enable the investigation of the compensatory effects linked with the palate shape (Brunner et al. 2006).

Figure 4: Example of tongue approximation of an /l/. Yellow contours are original contours. The approximated VT contour is represented by the red line.

6. References


Information density of speech: Languages differ in time per degree-of-freedom in picture description tasks

Mark L. Latash1, Elizabeth Carlson1, and Irina L. Mikaelian2

1 Department of Kinesiology, The Pennsylvania State University, University Park, PA 16802, USA
2 Department of Germanic and Slavic Languages and Literatures, The Pennsylvania State University, University Park, PA 16802, USA

ml111@psu.edu, eac5224@gmail.com, ixml2@psu.edu

Abstract

Using simple two-object picture description tasks, we studied the differences in the speed of information transmission across five languages. All speakers showed strong linear dependences between the number of degrees-of-freedom (DOF) and speech time. The amount of time per DOF was much shorter for Mandarin Chinese as compared to the other four languages (English, Russian, Vietnamese, and Korean). The Korean speakers spent more time per DOF compared to other languages. We discuss the results as possible consequences of the language morphology. In a pilot study, speech time was analyzed during descriptions of cartoons from “The New Yorker”. DOFs were estimated by two experimenters independently. Speech time increased linear with DOFs but at about half the rate as compared to the two-object tasks. We conclude that, while the general increase in ST with DOF is valid over a broad range of picture description tasks, less constraining tasks may show faster speech.

Keywords: picture description, information transmission, morphology, speech time, Fitts’ Law

1. Introduction

How much time is needed to transmit a certain amount of information using natural spoken language? While information transmission is obviously one of the most important functions of speech, and time is one of the most important characteristics of any human action, this question has not been addressed until recently. The first studies with simple picture description tasks produced a few non-trivial results (Latash and Mikaelian 2011; Latash et al. 2011). First, speech time (ST) has been shown to scale linearly with the log-transformed “index of difficulty” (ID):

\[ ST = a + k \cdot \ln(ID), \]

where \( a \) and \( k \) are constants. This dependence is similar to the famous Fitts’ Law, which describes movement time as a function of an index of difficulty (Fitts 1954; reviewed in Meyer et al. 1982). Within Fitts’ law, ID is proportional to the ratio of movement distance to target size.

Second, there was a significant difference between native speakers of English and of Chinese. While both groups showed ST values proportional to \( \ln(ID) \) with the correlation coefficients typically over 0.9, the Chinese speakers saved close to 40% of time while describing simple pictures with two familiar objects located and oriented differently with respect to each other.

Several hypotheses were tested in the mentioned studies. In particular, it was shown that the quality of information transmission (quantified as the number of errors per utterance) did not differ between the English and Chinese groups. It was also shown that the difference in ST was not due to easier perception of the pictures, which were comprised of only two objects each and could be viewed as resembling Chinese characters. This possibility was tested in control trials when the subjects had to reproduce the same pictures using actual objects without describing them. In those trials, the time from picture presentation to task completion was similar between the two groups.

The differences between the two languages were also not due to faster/slower speech between the groups: The primarily reason for the group differences was the different number of syllables used by the speakers while ST per syllable was similar. A hypothesis was offered that the difference in ST per index of difficulty was primarily due to morphological features of the two languages, isolating for Chinese and inflectional for English.

The purpose of this study has been to test this hypothesis by expanding the range of languages to include one more isolating language (Vietnamese), one more inflectional (Russian), and one agglutinative (Korean). We also started a study with the purpose to explore whether the linear relation between ST and \( \ln(ID) \) discovered in the highly-constrained and structured simple picture description task can be extended for more natural, less constrained speech tasks. Here we present pilot data from that study.

2. Methods

Native speakers of Mandarin Chinese (n = 8), American English (n = 8), Korean (n = 7), Russian (n = 8), and Vietnamese (n = 8) produced “quick and accurate” verbal descriptions of pictures presented on a computer screen in a self-paced manner. All the participants were about the same age and level of education (undergraduate and graduate students); the number of male and female subjects per group was also about the same.

The pictures always involved two objects: a plate and one of three other objects (a stick, a fork, or a knife), which were located and oriented differently with respect to the plate in different trials. The stick was a symmetric object with the two ends and two sides identical to each other. In contrast, the fork had two ends (the prongs and the handle), while the two sides were symmetrical. The knife had two ends (the tip and the handle) and also the two sides (the dull side and the side with the blade). The second object could be located in any one of eight positions: under, above, to the left of, or to the right, or upper- or lower-left, or upper- or lower-right corners of the plate. The objects could be oriented vertically, horizontally, at a 45° angle from the vertical, or slightly tilted off the vertical or horizontal axes. A full
description of the picture had to include the name of the object, its location and orientation with respect to the plate, and also, for the non-symmetrical objects, the disposition of their ends and sides. Each series consisted of four sets, 18 pictures in each set. The compositions of the 72 pictures were selected at random from all possible compositions to have a balanced representation of different difficulty levels.

Each task was assigned an index of difficulty (ID) according to the following equation:

\[ ID = ID_{OB} \times ID_{LO} \times ID_{OR}. \]  

(2)

where \( ID_{OB} \) is index of difficulty of the object, \( ID_{LO} \) is index of difficulty of the location, and \( ID_{OR} \) is index of difficulty of the orientation. \( ID_{OB} \) has been assigned three values, 1 (stick), 2 (fork), and 4 (knife). \( ID_{LO} \) has been assigned two values, 1 (above, below, to the left, and to the right of the plate) and 2 (at one of the corners). \( ID_{OR} \) has been assigned two values, 1 (vertical or horizontal) and 4 (tilted 45° and slightly tilted). Note that ID (ranging from 1 to 32) corresponded to the number of degrees-of-freedom (DOF) the participants had to describe (ranging from 1 to 6). In this report we analyze and present the relations between speech time and DOF.

Within the main task, the participants were tested in pairs. One subject (the Speaker) described the picture, and the other one (the Performer) had to re-create the picture using the actual objects. The Speakers were free to select their preferred strategy of describing the objects, but they were always reminded to try to do this as quickly as possible. They were also reminded to obey the rules of grammar. Speech time (ST) of the Speaker, movement time of the Performer, and other timing indices were quantified. The number of errors was also quantified. In control trials, no speech was involved: Each subject re-created the pictures using the objects (the Performer’s task) while looking directly at the pictures presented on the computer screen.

For each subject, ST (and other timing indices) values for the trials with identical DOF were averaged before further processing. Linear regression analysis of ST as a function of DOF was performed for each subject. In addition, the data were averaged across subjects within each language group for each DOF separately, and linear analysis was run on the averaged data. In addition, mixed-effects ANOVAs were used with two factors: Language (five levels) and DOF (six levels).

We also performed a pilot study involving two tasks. First, the subjects (so far, four subjects were tested, all native speakers of the American English) described the same two-object pictures with the plate and a stick, a fork, or a knife as in the first experiment. The subjects were tested individually, so they performed the Speaker task without a Performer. Next, the subjects were presented with a set of cartoons from “New Yorker”. In that task, the subjects were shown 32 cartoons, one by one, and were instructed to describe quickly and completely what they saw. The subject’s voice was recorded and speech time was measured. Further, the descriptions were transcribed, and an index reflecting the number of DOFs was assigned to each of the descriptions. DOFs reflected the number of objects, attributes, actions, and relations present in the transcribed descriptions. Two investigators assigned DOF values to each of the transcribed utterings independently of each other; the investigators did not know the corresponding time indices.

3. Results

In this report, we focus in this report on speech time (ST) as the main outcome variable. Reaction time (RT, from picture presentation to the initiation of speech) showed relatively minor differences between the groups. RT increased with DOF, but this dependence was relatively weak leading to an increase in RT by 10-15% over the whole range of DOF (from 1 to 6). Movement time by the Performer and total Task Completion Time showed dependences on DOF similar to those of ST. So, all the effects seen in ST(DOF) dependences were also seen in those timing indices.

All the subjects scaled ST with the task difficulty (quantified as DOF). Linear regressions of the ST(DOF) dependence showed significant results for each of the subjects. There was a significant difference among the languages. ST was the shortest among the Chinese speakers who spent, on average, about 0.65 s per DOF. The English speaking subjects were the next fastest (0.93 s/DOF), followed by the Vietnamese (1.19 s/DOF), Russian (1.3 s/DOF) and Korean (1.44 s/DOF) speakers. There were only minor differences among the intercepts of the linear regression equations; the intercepts ranged between 0 and 1 s. These findings are illustrated in Figure 1.

There were no significant differences among the languages in the number of errors (on average, one error per 12 pictures, typically in the description/performance of the object orientation). Only minor differences across the five language groups were seen in performance time in the control trials (performance without speech production). In those trials, performance time scaled linearly with DOF, but the regression coefficients did not differ across the five groups.

The main results illustrated in Figure 1 were supported by a two-way ANOVA that showed main effects of both factors,
DOF and Language ($p < 0.01$). There were significant pairwise differences between all the pairs of DOFs. Performance of the Chinese speakers (filled blue circles in Fig. 1) was faster than that by any other language group. The Korean speakers (filled green squares) were also significantly slower than the English (red open circles) and Vietnamese (orange filled rhombus) speakers.

In the pilot experiment, when the subjects described the two-object pictures with the plate and one more object (stick, fork, or knife), all four subjects showed scaling of ST with the number of DOF similar to the ones illustrated in Figure 1. The correlation coefficients computed over the six data points were very high ($>0.99$ for each subject) while the speech rate was slightly higher than the average value for the English-speaking subjects (corresponding to about 1.05 s/DOF). The descriptions of the cartoons took much longer, ranging between 7 and 24 s. While the estimates of DOFs differed between the two experimenters (see the blue and black data points in Figure 2), the overall linear increase in ST with DOF was evident in both sets of DOF estimates, and the correlation coefficients were over 0.7 ($p < 0.01$). Figure 2 illustrates the data for one of the subjects. Note the about two-fold increase in the speech rate during the performance of the cartoon task, as compared to the two-object task, evident in both sets of data points. Note also the non-zero intercept in both regression equations. The latter observation suggests that the overall dependence of speech time on the number of DOFs is likely non-linear although it can be well approximated by linear functions within certain ranges.

![Graph](image)

Figure 2: The dependence of speech time on the number of degrees-of-freedom (DOF) for one of the subjects. The filled-red dots show the performance of this subject in the two-object picture description task. The filled-blue and open-black clouds of data points show the performance in the cartoon description task as estimated by two experimenters (Exp-1 and Exp-2). Linear regression lines and regression equations are presented.

### 4. Discussion and conclusions

Speech time scaled linearly in the two-object picture description tasks across all subjects and languages. This scaling resembles strongly the classical Fitts' law, which links movement time to an index of difficulty computed as the ratio of distance to target size (Fitts 1954; Fitts and Petersen 1964; reviewed in Meyer et al. 1982; Plamondon and Alimi 1997). Fitts discussed this scaling based on the information theory, while later studies linked the scaling to cognitive processes at the level of movement planning (Bradi et al. 2009; Juras et al. 2009).

Equation (1) explained a large amount of variance in the speech time data, commonly over 95%. There were differences among individual participants in the coefficients $a$ and $k$. These differences were modest across speakers of the same language, but they could be rather dramatic between the languages. Indeed, as one can see from Figure 1, the regression coefficient $k$ ranged from 0.65 in the fastest group (Chinese) to 1.44 in the slowest group (Korean). One of the main results of this study is the demonstration that natural languages differ significantly in the amount of time needed to transmit a small amount of information defined by the task design.

The control series involving only the Performer has provided indirect support for the method of DOF estimation in the two-object tasks. Indeed, when required to perform only the motor part of the task, the subjects showed a classical Fitts-like linear scaling of movement time with DOF (equivalent to $\ln$(ID), see for detail Latash et al. 2011). So, with respect to performance in this series, the method of DOF estimation led to qualitatively the same results as could be expected from the classical method based on the ratio of target distance to target width.

Quick speech production was studied by the group of Sternberg (Sternberg et al. 1978, 1980, 1988). In those studies, the question was: How long does it take a person to pronounce an utterance consisting of ‘n’ elements (words, numerals, or non-words) with each element consisting of ‘m’ syllables? The subjects saw the text in advance and did not need to construct a phrase but rather to utter the elements as quickly as possible. Sternberg described a close to quadratic increase in the speech time with the number of elements and the number of syllables per element.

In our study, the speech times were 3–4 times longer than in the mentioned studies by Sternberg for utterings with comparable numbers of syllables (quantified in Latash et al. 2011). In contrast to Sternberg’s studies, speech rate did not change with speech time leading to a linear increase in speech time with DOF. So, one may conclude that the mechanisms that define speech production under the “speak as fast as possible” instruction are different from those that act under the instruction “describe the picture as fast as possible”.

The difference in speech time seems to correlate with the morphological type of language: Isolating for Chinese, inflectional for Russian and English (English being more analytical than Russian), and agglutinative for Korean. However, the results for the Vietnamese were quite different from those for the Chinese group despite the fact that both languages are isolating and both use tones as means of information transmission. Note that using tones for information transmission has been suggested in several studies (Valaki et al. 2004; Lee 2007).

We have been interested in regularities of the relationships between ST and task difficulty that would reflect the natural variability of speech. This feature distinguishes our study from the recent study of information density during reading (Pellegriino et al. 2011). In addition, we define DOFs as features of the task, not as features of the language such as the number of words or syllables.
The pilot results presented in Figure 2 are intriguing although one cannot draw conclusion from such a limited data set. The results suggest that there is a linear relation between speech time and DOF in less constrained cartoon description tasks that took much more time and involved description of many more features of the pictures. While our definition of DOF in these tasks is subjective and may even be called arbitrary, the data for both independent assessors of DOFs showed qualitatively similar dependences between speech time and DOF. Two features of these data make them different from the data in the two-object description tasks. First, the regression coefficient \((k)\) corresponded to a much less time per DOF in the cartoon description task. Second, there was a large intercept of the regression line (over 4 s) in contrast to the very low intercept values for the two-object task (under 1 s). Taken together, the observations suggest that the dependence between speech time and DOF is non-linear, but it can be approximated by a linear function over certain ranges of DOFs. This pilot study is the first attempt to explore the advantage some languages have over others in speed of information transmission, and in particular, to investigate whether that advantage persists or disappears with an increase in picture complexity and the introduction of other subtleties.

5. References


Do people converge to the linguistic patterns of non-reliable speakers?
Perceptual learning from non-native speakers

Shiri Lev-Ari¹, Sharon Peperkamp²

¹Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
²Laboratoire de Sciences Cognitives et Psycholinguistique (DEC, EHESS, CNRS), Paris, France

shiri.lev-ari@mpi.nl, sharon.peperkamp@ens.fr

Abstract

People’s language is shaped by the input from the environment. The environment, however, offers a range of linguistic inputs that differ in their reliability. We test whether listeners accordingly weigh input from sources that differ in reliability differently. Using a perceptual learning paradigm, we show that listeners adjust their representations according to linguistic input provided by native but not by non-native speakers. This is despite the fact that listeners are able to learn the characteristics of the speech of both speakers. These results provide evidence for a disassociation between adaptation to the characteristic of specific speakers and adjustment of linguistic representations in general based on these learned characteristics. This study also has implications for theories of language change. In particular, it cast doubts on the hypothesis that a large proportion of non-native speakers in a community can bring about linguistic changes.

Keywords: phonological representations, perceptual learning, language change, language contact, non-native speakers

1. Introduction

Speakers’ phonological representations are based on their accumulated experience, and specifically, the input they have received. For example, because speakers of different languages are exposed to different distributions of sounds, they will draw the boundary between similar sounds differently, and therefore interpret the same sound as belonging to a different category in accordance with their experience. For example, a French speaker will categorize as /p/ what an English speaker would interpret to be /b/ (Abramson & Lisker, 1970).

Furthermore, speakers’ representations are malleable; they are constantly shaped by further input (e.g., Flege, 1995; Goldinger, 1998). Any novel production listeners are exposed to influences the way they will interpret future speech. For instance, listening to a speaker who produces stops with deviant Voice Onset Times influences the way listeners will interpret stops by the same speaker as well as by other speakers in the future (Kraljic & Samuel, 2007).

Several accounts for the manner in which input influences future perception exist. Some theories of speech perception postulate that listeners retain all tokens they have encountered, together with contextual information, such as speaker identity. When new tokens are encountered, the stored tokens are activated according to their perceptual and contextual similarity, and thus guide the new tokens’ interpretation (e.g., Goldinger, 1998).

Other theories posit that listeners strip all the context-specific characteristic of the speech and only maintain its abstract representation. The properties of the abstract representation though depend on the common features of the encountered input (e.g., McClelland & Elman, 1986).

Importantly, listeners are not passive vessels that simply store all input, giving all tokens equal weight in all contexts. For example, listeners interpret the same token differently depending on who they believe the speaker to be. Thus, a certain token is more likely to be interpreted as /b/ than /v/ if the listeners believe the speaker is a woman rather than a man, because differences in vocal tract size influence the location of the boundary between these vowels (Johnson, Strand & D’Imperio, 1999). Similarly, speakers’ perceived age or regional background can influence listeners’ perception and interpretation of the sounds they hear (Koops, Gentry & Pantos, 2008; Niedzielski, 1999). These findings suggest that the weight given to stored tokens depends on their relevance for the situation.

An even more sophisticated aspect of the speech perception mechanism is its ability to modify existing representations according to input that is perceived to be reliable, but not according to input that is perceived to be unreliable or unrepresentative. For instance, listeners do not change their phonological representations in accordance with input provided by a speaker holding a pen in her mouth even though they change their representations after exposure to the exact same tokens when the speaker does not seem to have any obstructions in her mouth (Kraljic, Brennan & Samuel, 2008).

In this paper we investigate whether listeners similarly adjust their representation in accordance with input provided by reliable, but not by unreliable speakers, and in particular, whether listeners’ representations are therefore influenced by the speech of native, but not by the speech of non-native speakers. Answering this question will help us better understand not only the mechanisms of perceptual learning, but also the relation between adapting to the characteristics of specific speakers and adjusting general representations. Relatedly, the results of this study can indicate whether a large proportion of non-native speakers in a community can lead to language change, as has been proposed.

1.1. Non-native speakers and language change

Non-native speakers’ speech is influenced by the properties of their first language, and is therefore unrepresentative of the common native productions (Flege, 1987). The deviations from common values are what makes non-native speech accented, and usually harder to understand. With experience, listeners adapt to these deviations, and their understanding of novel accented speech improves (Baese-Berk & Bradlow, 2013). From the point of communication efficiency though, listeners should rely on these learned properties only when listening to speakers with the same characteristics. That is, input from non-native speakers should not influence the way the speech of native speakers is processed and interpreted, as its deviations are atypical for native speakers’ speech.
Nevertheless, several language contact accounts postulate a substratum language change, that is, language change that is brought about by having many non-native speakers in the language community. They propose either that those non-native speakers introduce into the language features from their first language or that they fail to acquire existing features and thus lead to their demise. It is argued that via repeated interactions between native and non-native speakers, and therefore repeated accommodation of native speakers to these non-native speakers, the novel patterns spread in the community (Niedzvelski & Giles, 1996; Trudgill, 1986; but see Hinskens & Auers, 2005).

In the case of interaction between speakers of different languages or dialects, such accommodation is argued to be a trigger of dialect leveling, as the latter often follows regular interaction between speakers of different dialects (Trudgill, 1986). Others have similarly suggested that languages with a large number of non-native speakers become simplified with time, due to the learners’ simplification of the language (McWhorter, 2007). Yet for non-native speakers to leave a mark on a language, native speakers must learn and propagate the foreign features that are used by the non-native speakers. As the literature on perceptual learning indicates, such a process might not be likely.

In this study, we test whether non-native speech can influence the representations of native listeners. Importantly, the feature that we manipulate and whose influence we examine is not explicitly perceived as deviant. Therefore, we do not test whether listeners will adopt features that they explicitly perceive as erroneous, as we know that listeners are reluctant to do that even when the speaker is native (Ivanova, Pickering, McLean, Costa & Branigan, 2012).

2. Study

We use a perceptual learning paradigm (Norris, McQueen & Cutler, 2003). In this type of paradigm, listeners are exposed to an ambiguous segment, such as a stop with a Voice Onset Time that is intermediate between that of /b/ and /p/, in a context that disambiguates it (e.g., les six /hp?fatanes ‘the six bananas’ vs. le long /plont ‘the long bridge’). They are later tested on their perception of the two phonemes that the ambiguous segment fell in between with a phoneme categorization task. Generally, listeners exhibit learning of the speaker’s productions by shifting their boundary between the two phonemes in accordance with the interpretation of the speaker’s earlier productions. When the ambiguous feature is not one that simply reflects individual physiological characteristics, the listeners also generalize their learning, as indicated by exhibiting a boundary shift even when categorizing the speech of novel speakers (Kraljic & Samuel, 2007). We hypothesize that this generalization is blocked when during exposure, listeners are presented with the speech of a non-native speaker. To recapitulate, participants in our experiment listened to either a native or a non-native speaker whose either /pl/ or /b/ tokens were manipulated. Later, they were tested with the speech of either the same speaker or a novel native speaker. We hypothesized that all listeners should be able to learn the characteristic of the speech, and thus show an effect of perceptual learning when tested with the speaker again. In contrast only those who listen to the native speaker should generalize the learned altered VOTs to new speakers, indicating a general influence on their representations, and therefore only they should show an effect of perceptual learning with a novel speaker.

2.1. Method

Participants. One-hundred-fifty-nine native French speakers were randomly assigned to one of eight conditions. Five participants were excluded because analysis of their performance indicated that they do not rely on VOTs when categorizing stops, at least in the tested VOT range.

Stimuli. Exposure stage. Twenty French words containing /bl/ and 20 French words containing /pl/ were selected. All were common words depicting concrete objects (e.g., bananes ‘bananas’). One male native French speaker and one male native Dutch speaker were recorded reading these words in short noun phrases that contained a determiner, the noun and an adjective (including numerals), e.g., les six bananes. None of the adjectives contained a stop. The VOTs of the two speakers in the target words did not differ (18.6 vs. 21.5 ms, p>0.05). We then created another version of each target word by cutting out the pre-voicing in the voiced stops and the positive voicing in the voiceless stops. Thus, all altered words had stops with a VOT of 0. An earlier pilot study confirmed that all productions sounded natural and that the modification was not detectable by naïve participants. The speakers also read 119 filler noun phrases that did not contain any stops.

Each target and filler word was then matched with two pictures. Both pictures depicted the object in the NP, but only one of each pair of pictures matched the adjective as well.

Phoneme categorization. The same native French and native Dutch speakers, as well as an additional female French speaker were recorded saying the words bulle ‘bulle’ and pull ‘pul’. For the pull end of the continuum, we used a token of 29ms – the VOT value in the speech of the male native speaker. The VOT of the Dutch speaker was particularly short for this word (16ms), and would not have been sufficiently long to allow creation of a VOT continuum. We therefore artificially lengthened it by copying the part of the existing VOT and pasting it at the end to match the VOT duration of the other training speaker. The bulle end of each continuum was created by cutting out the positive VOT from the pull token and pasting pre-voicing from that speaker’s production of bulle. We then created 10 additional steps in between the continuum’s ends by cutting out medial sections of positive or negative voicing. The values of all tokens in the continua of the two training speakers (in ms) were: -90, -50, -15, -10, -5, 0, 5, 10, 15, 20, 25 and 29.

Pre-tests indicated that the perceptual transition point from voiced to voiceless stop was earlier for the two training speakers than for the generalization speaker, potentially due to the gender difference (Swartz, 1992). Therefore, the range of values in the continuum of the generalization speaker extended to higher VOT values and was: -90, -60, -10, -5, 0, 5, 10, 15, 20, 25, 30 and 59. Other than that, the continuum was prepared in the same manner as the continua of the training speakers.

Note that in all continua, the two extreme tokens on each end of the continuum served as filler reference points. They were repeated fewer times and responses to them were not analyzed (see Procedure and Results).

Procedure. Participants came for two sessions, at least two days apart. In the first session, participants performed both the exposure task and the phoneme categorization task. In the second session, participants performed only the phoneme categorization task thus providing a baseline measure of participants’ categorization. We obtained the baseline measure at the end rather than the beginning to avoid drawing participants’ attention to the manipulated phonemes, and thus potentially to our manipulation of the audio files.

In each trial of the exposure stage, participants saw two different pictures of an object (e.g., a modern port and an old...
and listened to an auditory phrase that matched only one of the pictures (e.g., le vieux port ‘the old port’). Their task was to select the picture that matched the phrase. In this stage, half of the participants listened to the native speaker and half listened to the non-native speaker. Half of the participants in each Speaker condition heard the natural tokens of /p/ and the altered tokens of /b/. The other half heard the natural tokens of /b/ and the altered tokens of /p/. There were 159 items in total: 20 with /b/, 20 with /p/, and 119 filler items. The first two items were filler items. The rest appeared in random order.

Following the exposure stage, participants performed the Phoneme Categorization task. Half of the participants in each speaker condition listened to the same speaker as in the exposure stage. The other half listened to the novel female French speaker. Participants were explicitly told whether the speaker they were listening to was the same one as before or a different one. The phoneme categorization task contained two blocks. In each block, the two shortest and two longest tokens were repeated four times each, and the other eight tokens repeated ten times each, totaling 96 trials per block. One trial with the extreme token on the pull end of the scale preceded by a the first block to provide a reference point and practice. In the second session, participants performed the exact same Phoneme Categorization task, but without the exposure phase.

2.2. Results

To examine whether non-native productions can influence listeners’ linguistic representations, we needed to first identify the cases in which native productions influenced listeners’ representations. We therefore first analyzed the responses of participants in the generalization condition who were exposed to manipulated native speech. A mixed model analysis with Participant as a random variable and Baseline Performance, Manipulated Phoneme (voiced, voiceless), Item and the interaction of the latter two as fixed effects revealed a main effect of Baseline ($\beta=3.38$, $p<0.0001$), indicating that participants’ responses in the baseline session predicted their responses in the experimental session. The analysis also revealed effects of Manipulated Phoneme ($\beta=1.36$, $p<0.03$) and Item ($\beta=-0.41$, $p<0.0001$), but these were modulated by a the Item interaction of Manipulated Phoneme x Item interaction ($\beta=0.29$, $p<0.001$), indicating that the effect of perceptual learning was not present or of equal magnitude for the different items. We therefore carried out separate analyses on each item to determine in which items the effect of perceptual learning was manifest. Analyses for items 3, 4, 5, 6, 9 and 10 did not reveal any effect of Manipulated Phoneme (all $p>0.1$), but only effects of Baseline Performance (all $p<0.04$, except for item 9, $p=0.1$). For Item 7 there was a marginal effect of Manipulated Phoneme ($\beta=0.71$, $p<0.06$), as well as an effect of Baseline Performance ($\beta=2.23$, $p<0.01$), and for Item 8, there was only an effect of Manipulated Phoneme ($\beta=1.02$, $p<0.05$). These results indicate that the perceptual manipulation has most clearly influenced the perception of Item 8, where it was even strong enough to completely eliminate the effect of participants’ baseline perception of this item.

We next examined whether the performance of participants who were exposed to the non-native speech also showed an influence of the manipulated phonemes on their representations. Analyses of performance on each item separately, as well as on all items together, showed that participants were not influenced by the manipulated phoneme for any of the items (all $p>0.1$). Furthermore, we ran a joint mixed model analysis of participants in both the Native and Non-native speaker conditions over responses for Item 8 to see if participants’ responses in the two Speaker condition are indeed different. The analysis revealed an effect of Baseline Performance ($\beta=3.2$, $p<0.01$) and the predicted Speaker x Manipulated Phoneme interaction ($\beta=-1.14$, $p<0.05$). To conclude, the results show that only the speech input that was provided by the native speaker influenced listeners’ phonological representations.

One may wonder though whether non-native speech failed to influence listeners’ representations because it is harder to learn rather than because listeners block the influence of less representative speech. To examine whether that is the case, we tested whether participants who were tested with the same speaker succeeded in learning the speech of the non-native speaker. Separate analyses for each item indicated that listeners’ performance showed an effect of Manipulated Phoneme for item 5 ($\beta=-0.49$, $p<0.04$), and marginal effects of Manipulated Phoneme for items 7 ($\beta=-0.41$, $p<0.06$) and 8 ($\beta=-0.58$, $p<0.06$). Importantly, jointly analyzing the responses of participants who listened to the native and non-native speakers in the Same Speaker condition did not reveal an interaction of Speaker and Manipulated Phoneme for any of the items separately nor for all items together. This indicates that the magnitude of the perceptual learning effect was similar in the Native and Non-native speaker conditions. Lastly, to ensure that the performance in the Same Speaker and Generalization condition are indeed different, we ran an analysis of responses to Item 8 by participants in all conditions. This analysis revealed an effect of Baseline Performance ($\beta=2.15$, $p<0.0001$), an interaction of Speaker with Manipulated Phoneme ($\beta=1.02$, $p<0.04$), a marginal interaction of Generalization and Manipulated Phoneme ($\beta=0.77$, $p<0.09$), and the predicted 3-way interaction of Speaker, Manipulated Phoneme and Generalization ($\beta=1.61$, $p<0.02$). Together, the results show that even though listeners are able to learn the speech of both native and non-native speakers, they do not adjust their representation in accordance with it, and therefore only generalize native input.

3. Discussion

Individuals learn language from their environment: They learn the characteristics of the language of the speakers they hear and adjust their representations accordingly. Yet not all input that listeners encounter is equally reliable. For example, individuals should rely less on speech input provided by someone who has just been to the dentist or who is holding a pen in her mouth. Similarly, individuals should rely less on the speech of speakers who deviate much from the majority of their community, such as non-native speakers, people with speech impediments, or children. Our results indicate that listeners indeed take the reliability of the speaker into account when learning their speech.

Using a perceptual learning paradigm, we manipulated the VOT of bilabial stops, a feature that differs across languages, and that non-native speakers therefore sometimes produce with a foreign accent. The manipulation was applied to every token on the pull end of the scale, and due to the semantic context, imperceptible to listeners. Previous research has shown that exposure to such altered speech influences listeners’ later interpretation of speech by that speaker, as well as by other speakers (Kraljic & Samuel, 2007; Norris et al., 2003). In other words, such exposure can alter listeners’ representations of voiced and voiceless stops in general. Indeed, the results of our study replicated this effect. Crucially, the results of our study show that such influence on representation only occurs when the input is provided by native speakers. On the one hand, all listeners were able to learn the characteristics of the speaker’s speech, in line with previous research showing quick accommodation to foreign
accent (Bradlow & Bent, 2003). Yet despite this ability to learn the speaker’s speech characteristics, only those who listened to the native speaker adjusted their representation in general according to the newly learned characteristic, and used it when interpreting the speech of a new speaker.

One may wonder why listeners showed generalized perceptual learning on a single item only, that of 15 ms of VOT. One potential reason might be that even though listeners generalize the characteristics they learn to other speakers, the influence of these characteristics might be smaller than when listening to the same speaker, especially when there is a need to map the characteristics to speech with different properties due to, for example, a change in gender. It is also worth noting that even though the effect of generalized perceptual learning was only significant for one item, the participants’ performance in three additional items (i.e., items 5, 6 & 7, with VOT values of 0, 5 & 10ms) was in the predicted direction, but failed to reach significance.

Our study, then, shows that listeners block input of non-native speakers form influencing their representations. Future research should examine whether there are native speakers that are also considered to be a less fit model to learn from, and whose input is therefore not weighted as heavily as others’. For example, children, or even adults of a different generation might be considered to be less good models. Another question that remains open is how listeners decide how to classify speakers. Do listeners rely on top-down information, such as knowledge that they listen to a non-native speaker, or are they influenced by bottom-up cues, such as the general deviations of the speech input from other speakers? In other words, it could be that participants in our study implicitly blocked any influence of the speech of the non-native speaker because of their knowledge that he was a non-native speaker. Alternatively, it could be that the participants encoded the deviation of the non-native speaker from average speakers in the community on multiple features other than VOT, and thus implicitly tagged him as an outlier and blocked any influence of his speech. The latter account suggests that even native speakers might not all be learned from to the same degree. In particular, listeners may learn less from native speakers the further their speech is from the average speaker or from the speakers that are perceived to be good models.

The results of this study also have implications for language change. Many communities in today’s world include a significant proportion of non-native speakers. According to some theories, such composition should lead to language change. Yet for substratum language change to be a likely result, non-native speech must be able to influence native representations. The results of our study suggest that this is not the case.

One objection might be that substratum language change could also come about not by accommodation of native speakers to non-native speakers but by transfer of the features from non-native parents to their children. While our study does not speak to that, some research suggests that this might not be the case either. Children of non-native speaking parents are usually exposed to speech by other community members as well, and research on dialect variation has shown that the ultimate speech patterns of children of other-dialect speakers often end up resembling those of the community rather than the parents (e.g. Flocia, Delle Luche, Durrant, Butler & Goslin, 2012). More research is needed, though, to examine whether children of non-native parents have representations that differ from those of children of native speakers.

At the same time, there is ample evidence that language contact can lead to language change. Note that we do not argue that language contact cannot trigger language change. Rather, we argue that such contact-driven language change is not brought about by accommodation to non-native speakers. One alternative route for language change might be the influence of a second language on the first one when native speakers learn the languages of other language communities.

The main conclusion of this study, though, is that while listeners’ representations are influenced by the speech they hear, not all speech input is equally able to modify listeners’ representation, and in particular, non-native speech does not influence native representations.

4. References


Intra-cluster Timing in Romanian Stop-initial Onsets
Stefania Marin

Institute of Phonetics and Speech Processing, Ludwig-Maximilian University Munich, Germany

marin@phonetik.uni-muenchen.de

Abstract

This study investigates intra-cluster timing of Romanian onset clusters with varying segmental composition. Results show that the consonants in rhotic onsets PR, KR were timed significantly further apart than in onsets whose second consonant was a lateral (PL, KL), sibilant (PS, KS), stop (KT) or nasal (KN). Unlike in German (Bombien et al. 2013), the consonants in clusters KL and KN were timed similarly. A consonant-initial effect was only observed for the comparison PS-KS but was not easily interpretable as a place order effect (Chitoran et al. 2002). The results, while different from the pattern reported for German (Bombien et al. 2013), support the influence of aerodynamic and perceptual requirements on intra-cluster timing, while at the same time highlighting the role that language-specific timing patterns may have in shaping differences between particular clusters.

Keywords: intra-cluster timing, complex onsets, Romanian

1. Introduction

Segmental composition has been shown to significantly affect the timing relationship between the consonants within a cluster. For example, the consonants in German onset cluster KL have been shown to be timed closer together than those in onset KN, and likewise the consonants in onsets PL and KL compared to PS and KS (Bombien et al. 2013). The difference between KL and KN onsets has been attributed to perceptual constraints: aerodynamic simulations indicated that a great degree of overlap between the nasal and the velar would attenuate the velar burst characteristics due to nasal leakage and would thus compromise perception (Hoole et al. 2013). Diachronically, this has been hypothesized to lead to instability of KN, compared to KL, explaining for example the loss of onset KN in English. Likewise, the longer temporal lags in PS/KS compared to PL/KL have been attributed to perceptual and articulatory constraints (Bombien et al. 2013): a longer lag prevents the stop from being perceptually masked by the salient sibilant frication, and it also allows for the precise formation of the specific articulatory posture required for sibilant production. In the Bombien et al. (2013) study, identity of the first consonant in the cluster (/pl, /kl/) did not have a robust effect.

In addition, rhotic onset clusters such as PR have been shown to be less overlapped than lateral PL clusters in German, French, Portuguese and Romanian, likely due to aerodynamic factors required in producing the uvular approximant in German and French, or the alveolar trill in Portuguese in Romanian (Cunha, 2012; Hoole et al., 2013; Marin & Pouplier, 2014).

Finally, another cluster timing effect attributed to perceptual requirements is the so-called place order effect, whereby a front-to-back stop cluster such as BG would be more overlapped than a back-to-front cluster such as GB, as the former but not the latter would allow more overlap without fully masking the first consonant in the cluster (Chitoran et al. 2002). This effect has originally been observed for Georgian stop-stop clusters, and while various studies have attempted to generalize it to other cluster types, it has become evident that order effects are truly pertinent only to stop-stop clusters where recoverability issues play a greater role than in the case of other cluster types (cf. for example Galatos et al. 2010). In addition, Georgian clusters also exhibited a place of articulation effect, whereby clusters composed of coronal and dorsal consonants in either order were more overlapped than clusters composed of labial and coronal consonants; this effect was explained as likely due to the language-specific grammatical status of the respective clusters, where labial-coronal clusters, unlike labial-velar or coronal-velar are non-harmonic clusters (Chitoran et al. 2002).

If the previously reported differences in intra-cluster timing are due to perceptual/aerodynamic requirements, it is expected that they should hold cross-linguistically. In the present study, we investigate onset /kl/ and /pl/-initial clusters in Romanian, to further test how type of the second consonant in the cluster (lateral, rhotic, nasal, fricative, stop) affects intra-cluster timing. Potential order and place of articulation effects are also addressed, although as will be discussed, the Romanian data are not entirely appropriate for this purpose.

2. Methods

Articulatory (EMA) data from five native Romanian speakers were recorded and analyzed. The stimuli were real words, embedded in carrier phrases. All clusters were monomorphic. Six repetitions were targeted for each word.

Using the Matlab-based Mview software developed by Mark Tiede at Haskins Laboratories, kinematic events defining onset of movement, target achievement and release of consonants were determined on the basis of changes in the velocity profiles of the relevant articulatory movements (cf. for example Marin & Pouplier 2014 for further methodological details).

One measure of intra-cluster timing was defined as the temporal lag between release of the first consonant in a cluster and achievement of target of the second consonant in a cluster: 

\[ \text{Lag} = \text{Target}_{C2} - \text{Release}_{C1} \]

(Figure 1). This measure captures the temporal latency between release of the first consonant and target achievement of the second consonant, and replicates one of the measures used by Bombien et al. (2013). A larger value on this measure indicates a greater lag between the two consonants. Beside absolute lag values, normalized lag values were also computed in relation to duration of the constriction interval of the cluster, defined from achievement of target of the first consonant to release of the second consonant: 

\[ \text{Normalized Lag} = \frac{\text{Lag}}{\text{Release}_{C2} - \text{Target}_{C1}} \]

This normalization indicates how much of the constriction interval is taken by the lag between release of first consonant and target achievement of the second consonant.
is relevant in terms of perceptual recoverability in that it captures whether the achievement of target of the second consonant potentially masks the release of the first one. It likely also reflects articulatory/aerodynamic constraints on how closely two constrictions may follow each other (cf. Bomblen et al. 2013).

A second intra-cluster timing measure, plateau overlap, was used following the analysis of Chitoran et al. (2002). This measure indicates when movement for the second consonant begins relative to the constriction interval (plateau) of the first consonant: \( \text{Plateau Overlap} = \frac{(\text{MovementOnsetC}_2 - \text{TargetC}_2)}{(\text{ReleaseC}_2 - \text{TargetC}_2)} \) (Figure 1). A negative value indicates that movement of the second consonant precedes target achievement of the first consonant, i.e. movement onset for the second consonant fully overlaps constriction interval of the first consonant. A value between 0 and 1 indicates at what point within the first consonant’s plateau movement for the second consonant begins, i.e. what percentage of the constriction interval is overlapped, with a value of 0.1 for example indicating that the second consonant begins at 10% within C1’s plateau, and therefore that 90% of it is overlapped. A value over 1 indicates no overlap between movement onset of second consonant and constriction interval of the first consonant, i.e. movement for the second consonant begins after the first consonant has been released. Overall, smaller values indicate increased overlap, and larger values decreasing/no overlap.

For statistical analyses, mixed linear models were computed using the lme4 package for R, with \( p \)-values being determined by comparing a model including the factor/interaction of interest with a model with no fixed factor/no interaction (cf. Bates 2010). This method circumvents the difficulty in estimating denominator degrees of freedom for mixed linear models. For post hoc comparisons, the \( p \)-values were determined using the Tukey adjusted contrast in the multcomp package for R (Hothorn et al. 2008). The data were analyzed with fixed factors First Consonant (/p/, /k/) and Second Consonant (/l/, /l/, /s/, /l/, /h/), and random factor Speaker. On the basis of previous research, we predict a difference between lateral and rhotic clusters, lateral and sibilant clusters as well as between lateral and nasal, with no effect of first consonant in the cluster.

### 3. Results

Lag means as a function of cluster are plotted in Figure 2, and the normalized lag means in Figure 3. The results were qualitatively the same if absolute or normalized lag values were used as the dependent variable, so they are presented simultaneously. As shown in Table 1, the fixed factors and the interaction between them were all significant. The lags were overall greater for /p/-initial than /k/-initial clusters. Post hoc analyses for factor Second Consonant showed that the main effect was due to rhotic clusters being significantly different from all other clusters (\( p<.001 \)), with no other types being significantly different from each other. Pairwise comparisons between clusters confirmed that the Second Consonant effect was due in both /p/-initial and /k/-initial clusters to the rhotic clusters (PR, KR) having significantly larger lags than any other clusters (\( p<.001 \)). The interaction between factors was due to PS having significantly larger lags than KS (\( p<.05 \)), while PL/KL and PR/KR did not differ from each other. In other words, the First Consonant effect was carried out by the PS/KS contrast, while the Second Consonant effect by the rhotic vs. lateral/nasal/fricative/stop contrast.

**Table 1. Statistical results of mixed linear models for dependent variables Lag and Normalized Lag, with fixed factors First Consonant, Second Consonant, and random factor Speaker.**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Lag</th>
<th>Normalized Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Consonant</td>
<td>( F=10.88, p=.001 )</td>
<td>( F=9.34, p=.003 )</td>
</tr>
<tr>
<td>Second Consonant</td>
<td>( F=58.85, p&lt;.001 )</td>
<td>( F=32.94, p&lt;.001 )</td>
</tr>
<tr>
<td>First Consonant * Second Consonant</td>
<td>( F=26.58, p&lt;.001 )</td>
<td>( F=20.25, p&lt;.001 )</td>
</tr>
</tbody>
</table>

![Figure 1: Example measurement for one PL production. Dotted lines indicate the kinematic events for C1 /p/, measured on the basis of lip aperture (LA). Continuous lines indicate the kinematic events for C2 /l/, measured on the basis of tongue tip (TT) vertical movement. Shaded boxes show the constriction interval (plateau) for each consonant.](image)

![Figure 2: Mean (+/- 1SE) lag values (ms) between release of the first consonant and achievement of target of the second consonant in a cluster.](image)

![Figure 3: Mean (+/- 1SE) lag values (ms) between release of the first consonant and achievement of target of the second consonant in a cluster normalized by constriction interval.](image)
Table 2. Statistical results of mixed linear models for dependent variable Plateau Overlap, with fixed factors First Consonant, Second Consonant, and random factor Speaker.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Plateau Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Consonant</td>
<td>F=1, p&lt;.05</td>
</tr>
<tr>
<td>Second Consonant</td>
<td>F=44.53, p&lt;.001</td>
</tr>
<tr>
<td>First Consonant x Second Consonant</td>
<td>F=27.32, p&lt;.001</td>
</tr>
</tbody>
</table>

Figure 4: Mean (+/- 1SE) plateau overlap values (ms) indicating when movement for the second consonant begins relative to the constriction interval of the first consonant.

The results for the plateau overlap measure were qualitatively similar to those of the lag measures, with one exception being sibilant clusters. Plateau overlap means are shown in Figure 4, and the statistical results are summarized in Table 2. No overall effect as a function of first consonant identity was observed, while factor Second Consonant and its interaction with First Consonant were significant. Post hoc analyses showed that for factor Second Consonant, the main effect was due to rhotic clusters being less overlapped (p<.001) and sibilant clusters being more overlapped (p<.01) than all other cluster types. Pairwise comparisons between clusters confirmed that the Second Consonant effect was due in both /p/-initial and /k/-initial clusters to the rhotic clusters (PR, KR) being significantly less overlapped than any other clusters (p<.001), and to sibilant clusters (PS, KS) being significantly more overlapped (p<.05) than lateral clusters (PL, KL). For matched comparisons (PL-KL, PR-KR, PS-KS), no effect of First Consonant was observed (p>.05).

4. Discussion

The results confirmed a very robust rhotic effect, corroborating the patterns previously reported (Cunha 2012, Hoole et al. 2013, Marin & Pouplier 2014), but not a lateral vs. nasal difference (KL vs. KN), or a lateral vs. sibilant difference (PL/KL vs. PS/KS), contrary to the pattern reported for German (Bombien et al. 2013). At first sight, the lack of a KL-KN difference, as well as the lack of a difference between lateral and sibilant clusters, seems to speak against a perceptual basis for the asymmetry in German. However, comparing the Romanian target-release lags with those reported for German, it becomes evident that Romanian KL and KN lag values are in the range for German KN (around 30ms) rather than for German KL (around 10ms). Indeed, Bombien et al. (2013) highlight the extremely short lags of the lateral clusters (especially KL) in comparison to the other clusters they examined. This suggests that the KN lag in Romanian may be large enough so that nasal leakage would not mask the velar burst. Likewise, PS/KS lags in Romanian, comparable to those in German (20-25ms), are large enough to meet both perceptual and articulatory requirements. Rather, the current results suggest that Romanian lateral clusters are less overlapped than German lateral clusters, pointing to the possibility that Romanian clusters may overall be less overlapped than German ones.

An overall initial-consonant effect was observed on the target-release lag measure with larger lags for /p/-initial than /k/-initial clusters, but matched comparisons indicated that this effect was carried out by clusters PS and KS alone. From a perceptual point of view, one would expect, if anything, that PS would allow shorter lags than KS, since in front-to-back clusters such as PS, the first consonant with a more anterior constriction is less likely to be masked by the more posterior constriction of the second consonant. Also from a production perspective, one would expect the same pattern: since /p/ has been shown to be less resistant to co-articulation than /k/ (Recasens et al. 1997), the sibilant would be expected, if anything, to encroach the labial more than the velar and not vice versa. It is not entirely clear therefore what factor(s) determine the Romanian PS/KS pattern. Chitoran et al. (2002) have observed for Georgian that labial-coronal stop clusters were less overlapped than velar-coronal ones, but this difference was likely due to the language-specific grammatical status of labial-coronal (non-homorganic) vs. velar-coronal (homorganic) Georgian clusters, so it is not clear to what extent this would apply to a cluster in a different language.

An order effect could not be systematically tested using the current data. Firstly, the available onset clusters conflate order with place of articulation: thus, if all effects reported for Georgian (Chitoran et al. 2002) are generalizable, then /p/-initial clusters are expected to be more overlapped than /k/-initial clusters as an order effect, but less overlapped as a place of articulation effect. Second, and perhaps more importantly, as mentioned in the introduction, order effects are hypothesized to play a role in shaping stop-stop intra-cluster timing, rather than apply to any cluster types. Since only one cluster (KT) in the current data meets this description, no comparisons could be carried out. Given that KT is a back-to-front stop-stop cluster, it may seem at first surprising that it exhibits a similar timing pattern to KL (or PL), but it must be emphasized that the second consonant in the cluster achieved its target at least 20 ms after release of the first consonant, which for perceptual purposes may be enough regardless of cluster composition.

The consistently much larger target-release lag in the case of rhotic clusters is determined not by perception, but by the aerodynamic requirements for producing a trill, whereby the trilling tongue tip articulation (which is the articulation measured here for the trill) must be synergistically supported by a preceding tongue dorsum retraction (cf. Solé 2002, for a discussion of the aerodynamic parameters required to initiate a trill).

Finally, regarding the plateau overlap measure, the one exception to the target-release lag measure pattern pertained to sibilant clusters. Thus, by this measure PS and KS showed significantly more plateau overlap than PL/KL, and did not differ in overlap degree from each other. The results thus indicate that the sibilant in the PS/KS clusters starts earlier than the lateral in PL/KL, relative to the constriction interval of /p/-/k/. Nonetheless, in relation to release of /p/-/k/, the sibilant reaches its target at a time comparable to the lateral in PL/KL. Also, although the sibilant starts at the same time relative to
the constriction interval of either /p/ or /k/, it reaches its target later when following /p/ than when following /k/. This further suggests that the PS/KS lag asymmetry may be due to particular (perhaps word/language-specific) constraints or perhaps measuring artifacts, an issue that remains to be explored in future research.

In conclusion, the current data do not contradict previous perceptual/aerodynamic accounts of differing overlap patterns as a function of cluster composition. Particular intra-cluster lag differences are however not generalizable in the absence of knowledge of overall timing patterns of a language. Thus, larger intra-cluster timing lags for KN compared to KL, or for PS/KS compared to PL/KL cannot be automatically predicted on the basis of the German pattern, without first knowing how the consonants in onset PL/KL are timed in the respective language. The diachronic asymmetry between KL and KN may therefore not hold for languages that exhibit overall greater intra-cluster lags. The results overall highlight the importance of direct comparisons between clusters in a variety of languages so that cross-linguistic patterns could be separated from language-specific ones.

5. Acknowledgements

Work supported by the European Research Council under the European Union’s Seventh Framework Programme/ERC Grant Agreement no. 283349-SCSPL.

6. References


Vowel Compensatory Shortening in Romanian
Stefania Marin¹, Lia Saki Bučar Shigemori¹

¹ Institute of Phonetics and Speech Processing, Ludwig-Maximilian University Munich, Germany
marin@phonetik.uni-muenchen.de, lia@phonetik.uni-muenchen.de

Abstract
This study systematically investigates the change in relative vowel and consonant duration as a function of increased complexity of onsets and codas (vowel/consonant compensatory shortening), and its interaction with type of onset/coda in Romanian. Results show that vowels shorten when they follow complex stop and liquid onsets, but not nasal or fricative ones, and when they precede liquid complex codas, but not nasal or obstruent ones. Compared to their timing as singletons, only liquid (but not nasal or obstruent) consonants shorten in both onset and coda position. This pattern contradicts the general predictions of either a compression or a gestural account.

Keywords: compression effects, compensatory shortening, syllable temporal organization, clusters

1. Introduction
Vowel compensatory shortening (i.e. vowel shortening due to increasing syllable complexity, henceforth VCS) has been reported in several studies (Katz 2012, Marin & Pouplier 2010, Munhall et al. 1992, Shaiman 2001). The conditioning environment, as well as the theoretical interpretation of VCS, have however been a matter of debate. On the one hand, VCS has been viewed as one instance of compression effects, whereby segments in syllables of greater complexity tend to shorten relative to their duration in simpler syllables, so that the overall syllable duration remains fairly regular (cf. Munhall et al. 1992). Under this hypothesis, vowels are expected to be shorter in syllables with complex onsets or codas, compared to vowels with simple margins.

On the other hand, under a gestural approach to syllable organization (Brownman & Goldstein 1998), VCS has been hypothesized to be a consequence of a temporal organization specific to onsets, but not to codas, and hence VCS is predicted when onset but not when coda complexity increases. Specifically, onsets are hypothesized to be timed globally to the following vowel, i.e. the onset’s mid-point (so called “c-center”) is assumed to maintain a stable relationship to the vowel. Consequently, when consonants are added to the onset, the rightmost and leftmost consonants are assumed to shift rightwards and leftwards respectively, relative to their timing as a singleton, so that onsets of varying complexity line up along their c-centers. This shift would result in an increasing overlap of the rightmost consonant and the following vowel, resulting in an apparent vowel shortening in the complex vs. simplex onset condition (i.e. an apparent VCS). No such increasing overlap is expected for codas, since they are hypothesized to be timed locally, i.e. only the vowel-adjacent consonant is assumed to be coordinated to the preceding vowel, and this timing should not be affected by increasing coda complexity.

The empirical evidence on VCS is partly contradictory even for a single language, English. For onsets, VCS has been reported independently of onset composition (Katz 2012, Marin & Pouplier, 2010), a pattern compatible with the predictions made by either a compression or a gestural account. The evidence for coda VCS is however mixed across studies and coda types. Thus, VCS in the context of obstruent codas such as /t-p/, /l-p/ has been observed by Munhall et al. (1992) and Shaiman (2001), but not by Byrd (1995). More recent studies (Katz 2012, Marin & Pouplier 2010) have corroborated the result of Byrd (1995) for obstruent codas, while at the same time reporting VCS for liquid codas (such as /l-p/, /l-rp/), a pattern not predicted by either a compression account (which would predict VCS independently of coda type) or a gestural approach (which would predict no VCS as a function of coda complexity increase). Under either approach, additional mechanisms would have to be postulated to account for the empirical English coda pattern (see e.g. Katz 2012). In the present study, we test to what extent the English VCS results are extensible to a new language, Romanian, and we specifically investigate the role of onset/coda composition on VCS.

Additionally, we also investigate consonant duration as a function of syllable complexity. Under a compression approach, consonant compensatory shortening (CCS) is also expected to be a mechanism available in maintaining comparable durations between simpler and more complex syllables (in addition to VCS, or in a trade-off relation with VCS). Previous research on English has suggested an overall tendency for consonants in clusters to be shorter than their singleton counterparts (Haggard 1973), although the effect was dependent on both consonant type involved and position in the syllable and cluster: thus, for example /l/ shortened in both onset and coda, but /m/ only shortened in coda if it was the first consonant in the cluster (e.g. /m/ shortened in coda /-mdl/, but not in coda /-lm/, or in onsets /sm-/) - likewise stops such as /p/ shortened if they were the second consonant in a coda cluster (/l-p/), but not the first (/l-ps/). A different study suggested an overall shortening of consonants in onset clusters, but a lengthening of consonants in coda clusters (O’Shaughnessy, 1974), but with different degrees of shortening in onset as a function of consonant type (minimal for stops, greater for liquids). Note that in this study no statistical analyses were carried out as only three speakers were recorded in the onset condition and two speakers in the coda condition. The current data allow us to investigate the relationship between VCS and CCS and their interplay with onset/coda composition.

2. Methods
Articulatory (EMA) data from five speakers were used to compare vowel duration in cluster words with vowel duration in corresponding singletons (e.g. /glæs/ vs. /læs/ forming Set GL-, with /l/ being henceforth referred to as the vowel-adjacent consonant). Syllable complexity was increased in either onset (e.g. Set GL-), or coda (e.g. Set –LG: /mʌɡ/ vs. /mʌlɡ/, with /l/
being the vowel-adjacent consonant in this set). Sets were further grouped on the basis of the vowel-adjacent consonant into four series: liquid (/l/, /r/), nasal (/n/, /m/ in onset, /n/ in coda position), fricative (/s/) and stop (/p/, /k/, /t/ in onset, /p/ and /k/ in coda). All available sets are listed in Table 1. All clusters were mono-morphemic and the stimuli were real words, embedded in carrier phrases. Six repetitions were targeted for each stimulus word (cf. Marin 2013, Marin & Pouplier 2014 for methodological details).

Table 1: Experimental sets.

<table>
<thead>
<tr>
<th>Series</th>
<th>Onset</th>
<th>Coda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>SM, SHM, KN</td>
<td>MN</td>
</tr>
<tr>
<td>Fricative</td>
<td>PS, KS</td>
<td>SK, SM</td>
</tr>
<tr>
<td>Stop</td>
<td>SP, SHP, SK, SHK, KT</td>
<td>PS, PT, KS, KT</td>
</tr>
</tbody>
</table>

Vowel (and consonant) acoustic segmentation is known to be problematic for certain segments and in certain contexts (e.g. vowels in the context of liquids, and especially the alveolar trill) and often different criteria have to be adopted on a case-by-case basis (cf. Haggard 1973, Katz 2012). In the current study, the availability of articulatory data allowed us to circumvent this problem by defining vowel and consonant durations on the basis of consistent kinematic events across contexts. Thus, using the Matlab-based Mview software algorithm developed by Mark Tiede at Haskins Laboratories, kinematic events defining target achievement and release of consonants were determined on the basis of changes in the velocity profiles of relevant articulator movements (cf. Marin 2013, Marin & Pouplier 2014 for details). The semi-automatic labeling procedure consisted in first detecting in the region of interest two velocity peaks corresponding to the articulator moving towards and away from target (Peak 1, Peak 2 in Figure 1). Target achievement (Target) was defined as the point in time at which velocity fell below 20% of Peak 1, and target release (Release) was defined as the point at which velocity exceeded 20% of Peak 2. All landmarks were determined automatically after manually selecting a point of analysis.

Figure 1: Example articulatory segmentation showing vertical movement and velocity trajectories for the tongue tip (TT), and kinematic events for consonants /l/ and /s/ in word “glas”.

Vowel duration was defined as the interval between release of /l/ and target of /s/. Vowel-adjacent consonant duration (in this instance duration of consonant /l/) was defined as the interval between target and release of the consonant of interest (Figure 1). For evaluating CCS, only the vowel-adjacent consonant duration was compared across singletons and clusters since only this was available in both conditions.

For statistical analyses, mixed linear models were computed using the lme4 package for R. To determine p-values for the main effect and the interactions between factors, a model including the fixed factor/interaction of interest was compared to the same model with no fixed factor/no interaction (cf. Bates 2010). This method circumvents the difficulty in estimating denominator degrees of freedom for mixed linear models. The p-values thus obtained is reported along with the F-value of the mixed linear model; because denominator degrees of freedom are difficult to estimate, they cannot be reported (moreover, they no longer play a role in computing p-values, given the method used here). For post hoc comparisons, the p-values were determined using the single-step adjusted method in the multcomp package for R (Hothorn, Bretz, & Westfall 2008). The data were analyzed with fixed factors Complexity (singleton, cluster), Position (onset, coda) and Series (liquid, nasal, fricative, stop), and Speaker and Set were included as random factors. Fixed factors Position and Series were of interest to the extent that they interacted with Complexity, hence we focus on reporting the results for Complexity as a main effect and its interaction with Position and Series.

The prediction under a compression account is that VCS and CCS should be observed regardless of where (onset or coda) syllable complexity increases, and regardless of composition. Possibly, trade-off relations may be at play between VCS and CCS. A gestural approach would predict vowel shortening as a function of onset but not coda increased complexity (with no clear predictions being made for consonant duration beyond the general prediction that compensatory shortening is the result of increasing overlap rather than gestural shortening and that it should be expected to the extent that consonants in clusters overlap each other more). Taking into account the English results, the predictions on vowel duration are more nuanced, with vowel shortening being expected for all onsets regardless of their segmental composition, but only for the liquid coda sets. The predictions on consonant duration are that shortening should be observed mainly in onset condition, and more robustly for liquids than other consonant types.

3. Results

3.1. Vowel duration

Vowel durations in the singleton and cluster conditions as a function of series are plotted in Figure 2. A significant effect for factor Complexity (F = 57, p<.001), for its interaction with Position (F = 8.24, p=.004) and Series (F = 9.6, p=.001), as well as for the three-way interaction (F = 5.43, p<.001) was observed. Post-hoc tests showed that in onset position, liquid and stop sets had significantly shorter vowel durations in the cluster compared to the singleton condition (p<.007). No significant difference in vowel length between conditions was observed for the nasal onsets (p>.05), and a significant vowel lengthening in the cluster condition was observed for the fricative onsets (p=.015). In coda, the vowel was shorter in the cluster condition only for the liquid sets (p=.001), with no significant difference being observed for the other series (p>.05; the comparison for the fricative series was at trend level, p=.057). The results therefore confirmed VCS for liquid and stop onset sets, and for liquid (and marginally fricative) coda sets.
3.2. Consonant duration

Vowel-adjacent consonant durations in the singleton and cluster conditions as a function of series are plotted in Figure 3. For consonant duration, factor Complexity (F = 29, p<.001), and the two-way interactions with Position (F = 5.48, p=.019) and Series (F = 2.76, p=.041) were significant, but not the three-way interaction (F = 0.35, p>.05). Post-hoc tests exploring the two-way significant interactions showed that across all series, the consonants were overall shorter in complex onsets than in simple onsets (p<.001), with no complexity effect in coda position (p>.05); across syllable positions, liquids in clusters were shorter than singleton liquids (p<.001), with no other series effects (p>.05). The absence of a three-way interaction suggests that CCS for liquids should be observed independently of position, and indeed a post hoc analysis confirmed that in both onset and coda position, liquids in the cluster condition were significantly shorter than singleton liquids (Liquid onsets: p<.001; Liquid codas: p=.037).

4. Discussion

In Romanian, both VCS and CCS were observed for selected syllable types: VCS was robustly observed for the liquid and stop onsets and for the liquid codas, while CCS was overall observed for liquids, and in onset position; possible trading relations between VCS and CCS could be observed for the nasal and fricative onsets, where the consonant but not the vowel shortened as a function of syllable complexity. Interestingly, when syllable complexity increased in coda, CCS was only observed in contexts where VCS was also present, i.e. within the liquid series. This pattern contradicts the general predictions of a compression account, and overall more conditions were found under which segmental compression was blocked than observed. This suggests that segmental compression to the extent that it is present, should be viewed as an epiphenomenon rather than as a driving mechanism in speech production.

At the same time, a gestural account to syllable organization also does not entirely predict the observed vowel duration pattern. Specifically, the vowel shortening pattern for liquid codas, while mirroring the pattern reported for English, does not follow from the gestural organization hypothesized for codas. In addition, fricative and nasal onsets do not conform to the predictions made for onset temporal organization: for the fricative sets, the vowel lengthens rather than shortens with increased syllable complexity, while no difference in vowel duration is observed for the nasal series. In addition to the principled distinctions between the organization of onsets and codas, additional factors would have to be considered to account for the unexpected patterns in the fricative and nasal onset conditions, as well as in the liquid coda condition.

Onset clusters where the vowel-adjacent consonant is a fricative (/p/-l, /k/-l) are not attested in English, so it is not clear whether the onset fricative pattern is language- or cluster-specific, although recent data from Polish (Pastätter & Pouplier 2014) suggests that it may be cluster-specific. Given the known coarticulatory resistance of the sibilant (Recasens 2012), it could be envisioned that the sibilant and the vowel cannot overlap more in the cluster condition than they already do in the singleton condition, and hence the sibilant cannot shift rightwards relative to its timing as a singleton, as it would be expected for complex onsets (cf. Pastätter & Pouplier 2014, for a similar suggestion for Polish sibilant onset clusters). What is however surprising in the Romanian result is that the vowel is significantly longer in the cluster condition, suggesting that the sibilant and vowel overlap even less in the cluster compared to the singleton condition. At this point it is not clear why that is the case. In Romanian, these fricative clusters are extremely rare, and it may be that they are produced differently than regular onsets, or produced entirely differently than their singleton controls (cf. Marin 2013).

The nasal onset pattern is surprising since it not only diverges from the pattern predicted for onsets, but also from the pattern observed for English nasal onsets. As far as can be determined, nasals in the two languages are not different in any crucial way, so it is not straightforward to propose an analysis that would predict the two patterns in the two languages. When examining the sets in more detail, it can be seen that there is on average a vowel shortening for set SM- (which is similar to SM-, SN- in the previous English studies), but not for sets SHM- and KN-. Like in the case of /p/-l, /k/-l, clusters /shm/- and /kn/- are rare in Romanian, so again they may not be treated like regular onsets (indeed, in a different type of analysis, there were suggestions that sets PS-, KS-, KN- in Romanian pattern together, cf. Marin 2013).

Finally, the liquid coda pattern mirrors that of English, and for English additional perceptual factors were postulated to explain why the vowel shortened in this context (Katz 2012, Marin & Pouplier 2010). Alternatively, production particularities characterizing liquids but not nasals or obstruents may result in an apparent vowel shortening given the measure employed here (as well as acoustic vowel segmentation conventions). Thus for example, the tongue dorsum retraction gesture needed for producing the Romanian trill, which precedes the tongue tip constriction, is part of the measured “vowel” (either in an articulatory measure as we use here, or in an acoustic measurement). If this tongue dorsum retraction gesture shortens, just like the tongue tip constriction has been shown to shorten, in the cluster vs. singleton condition, then the measured “vowel” would also automatically shorten. If so, the observed VCS for liquid codas is actually an artifact of the imperfect articulatory and acoustic measures available to us (cf. Marin & Pouplier 2014 for a more comprehensive discussion in the context of various cross-linguistic liquid clusters).
The results overall highlight the need for empirical data from a variety of languages and onset/coda types to evaluate competing theoretical assumptions regarding vowel and consonant compensatory shortening.

5. Acknowledgements

Work supported by the European Research Council under the European Union’s Seventh Framework Programme/ERC Grant Agreement no. 283349-SCSPL.

6. References

Frequency of Glottalization in Hungarian Read and Spontaneous Speech
Alexandra Markó
Department of Phonetics, Eötvös Loránd University, Budapest, Hungary
marko.alexandra@btk.elte.hu

Abstract
The present study analyzed the frequency and distribution of glottalization in 30 Hungarian speakers’ spontaneous and read speech. The aim of the study was to find out whether either the mode of speech (read or spontaneous) or the speaker’s gender, age, average f0 had an effect on the frequency of glottalization. Glottalization was more frequent in spontaneous than in read speech, and female speakers used glottalization more frequently. Glottalization was overrepresented in middle-aged women’s spontaneous speech, while it was the least frequent in the read speech of young men. A medium correlation was shown between the ratio of glottalization and the speakers’ average f0 in read speech. However, no correlation was detected either between frequency of glottalization and age, or between the ratios of glottalized syllables and glottalized filled pauses. Of the factors examined, gender had the strongest effect on frequency of glottalization.

Keywords: glottalization, gender, age, spontaneous and read speech, f0

1. Introduction
The vibration of the vocal folds is usually ‘quasi-periodic’. This means that the glottal impulses are almost evenly spaced and their amplitude is roughly equal. The periodicity is perturbed by small-scale cycle-to-cycle random variation both in the spacing and in the amplitudes of the impulses (measured by jitter and shimmer). Larger deviations from this modal voice are usually referred to as phonation types or voice qualities. Irregular phonation is a phonation type characterized by the irregular vibration of the vocal folds. It corresponds to regions of voiced speech with substantial, abrupt, cycle-to-cycle changes in either the spacing of the glottal impulses, or their amplitudes, or both. In such cases, the deviation from periodicity exceeds the usual jitter and shimmer values (present in regular phonation) (see Surana and Slifka 2006). This deviation is clearly audible for people with normal hearing.

Irregularity can show up in a number of forms. For instance, Batliner et al. (1993) distinguished six types of irregular phonation (laryngealization) in approx. 30 minutes of spontaneous and read speech by four German speakers. Dilley et al. (1996) studied texts read out by five American English speakers in radio news programs with respect to irregular voice quality occurring in word initial vowels. They defined four types of realizations.

‘Glottalization’ is an umbrella term for several forms of irregularity (see e.g., Dilley et al. 1996), with its use less than fully consistent in several authors’ work. In any case, ‘glottalization’ seems to have a ‘common sense’ interpretation in the literature, which will also be adopted in this paper.

Glottalization is a multifunctional phenomenon. In some languages, it expresses a phonological contrast (see e.g., Gordon and Ladefoged 2001). Several researchers have investigated the role of glottalization in expressing emotions and/or tried to use it in automatic emotion recognition (e.g., Batliner et al. 2007; Gobl and Ní Chasaide 2003). The boundary marking role of phrase/utterance final glottalization has been confirmed by a number of studies (e.g., Henton and Bladon 1988; Fant and Kruckenberg 1989; Slifka 2006). In Hungarian, too, glottalization often occurs sentence finally, both in read and in spontaneous speech (Böhm and Ujváry 2008; Markó 2009). Glottalization can also signal the end of the turn (Redi and Shattuck-Hufnagel 2001). In English, glottalization also occurs between adjacent vowels flanking a word boundary (Gimson 1980); and word initially it often occurs before an initial vowel in English (Dilley et al. 1996), in German (Kohler 1994) as well as in Hungarian (Markó 2012). The frequency of occurrence of glottalization is speaker dependent to a large extent: some speakers hardly produce any irregular voicing, while some produce it fairly frequently (e.g., Henton and Bladon 1988; Redi and Shattuck-Hufnagel 2001). Therefore, voice quality has an eminent role in human speaker recognition (Böhm and Shattuck-Hufnagel 2007).

The present study analyzed the frequency and distribution of glottalization in 30 Hungarian speakers’ speech. The aim of the study was to find out whether either the mode of speech (read or spontaneous) or the speaker’s gender, age, average f0 had an effect on glottalization frequency. The hypothesis was that the ratio of glottalized syllables is larger (i) in read speech (vs. spontaneous), (ii) in women’s speech (vs. men’s), in the case of (iii) higher age and (iv) deeper f0.

In spontaneous speech, filled pauses can be realized in various ways, some more frequent than the others. In Hungarian speech, the most common realization of filled pauses is a schwa-like voicing, while in some cases nasal sounds can also be detected (similar to [m]) as well as the combination of these two types (Horváth 2010). Thus, filled pauses can also be glottalized, a phenomenon specifically addressed in our analysis.

2. Subjects, material, method
The material was selected from the Hungarian spoken language database BEA (Gósy 2012). Female and male speakers were equal in number; they represented three age groups with 10 young, 10 middle-aged, and 10 elderly speakers (Table 1).

Table 1: Distribution of subjects by age and gender.

<table>
<thead>
<tr>
<th></th>
<th>Females</th>
<th></th>
<th>Males</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interval (yrs)</td>
<td>Mean (yrs)</td>
<td>Interval (yrs)</td>
<td>Mean (yrs)</td>
</tr>
<tr>
<td>Young</td>
<td>21–24</td>
<td>22.2</td>
<td>21–24</td>
<td>22.0</td>
</tr>
<tr>
<td>Middle-aged</td>
<td>39–46</td>
<td>43.6</td>
<td>39–41</td>
<td>40.0</td>
</tr>
<tr>
<td>Elderly</td>
<td>54–60</td>
<td>57.4</td>
<td>57–66</td>
<td>61.0</td>
</tr>
</tbody>
</table>
The subjects were native speakers of standard Hungarian, with typical articulation and normal hearing, and their voice production did not show any pathological trait, either. The subjects’ read texts total up to 65'399" (with each recording taking approximately 2 minutes, and the material consisting of 18,779 syllables in total). The spontaneous speech material altogether takes 1 hour and 46 minutes (comprising 3 to 4 minutes long, more or less continuous monological parts of life-interviews, with a total of 26,372 syllables).

The syllable was chosen as the item of labeling based on several considerations. First, there is no clear-cut boundary between regular and irregular phonation, therefore measuring the overall duration of irregularity would be unreliable. Labeling the speech sounds in terms of irregularity would be uncertain in continuous speech, and comparison of the speakers would be difficult because of the diverse articulatory strategies. So instead, counting the syllables that contain irregular phonation seemed to be a more appropriate measure, following Henton and Bladon (1988) and Böhm and Ujváry (2008). A syllable counted as glottalized if at least one irregular fundamental period belonged to that syllable. The filled pauses were labeled the same way.

The analysis of glottalization was performed in accordance with the methodology of previous studies (e.g., Dilley et al. 1996), combining visual and auditive information. Acoustically, glottalization was identified if (i) the duration or amplitude of the basic periods suddenly changed to a significant extent (including the occurrence of a glottal stop, see Dilley et al. 1996); or if (ii) the fundamental frequency suddenly fell below the speaker’s usual pitch range. In addition, as a perceptual criterion, cases in which the timbre was audibly hoarse or creaky were taken into consideration. Following the general practice in the literature on phonation types, no particular quantitative criterion was set up for irregular phonation. Praat 5.2 (Boersma and Weenink 2010) was used for labeling.

In both read and spontaneous speech, we analysed the ratio of glottalized syllables (compared to the total number of syllables occurring in the given speech production, in %), the length of glottalized syllable sequences (number of consecutive glottalized syllables), and the position of glottalized syllables within the inter-pause stretches of speech. In addition, turn-final glottalization, as well as glottalization in filled pauses were analyzed in spontaneous speech. Fundamental frequency in non-glottalized syllables was measured by the voice report function of Praat. The factors of ‘speech mode’, ‘age’, and ‘gender’ were taken into consideration. Paired and independent sample t-tests, two-way repeated measures ANOVA, and Pearson’s correlation were carried out using SPSS 15.0.

3. Results

Glottalization was more frequent in spontaneous than in read speech (Fig. 1), and this difference is significant in terms of speakers [t(29) = −3.203; p = 0.004]. Female speakers used glottalization more frequently than males, but this difference was significant only in the case of read speech [t(29) = −2.739, p = 0.012]. Frequency of glottalization was proved to be speaker-dependent. Speakers who rarely glottalized in read speech produced similar results in their spontaneous speech and vice versa (Pearson’s r = 0.767, p = 0.001). Frequency of glottalization did not correlate with the speakers’ age, however.

Glottalization was the most frequent in interviews with middle-aged female speakers; the average was 34.2% (SD 7.6%). The ratio of glottalized syllables was the smallest in the reading performance of young males, with a mean value of 14.7% (with SD 2.8%). The person who glottalized the least was an elderly man, with 4.9% of the syllables glottalized in read speech, and 6.0% in spontaneous speech. The highest ratio was measured in a middle-aged woman’s spontaneous speech: here, 47.4% of the syllables were glottalized. The ratio of glottalized syllables exceeded 40% only in the case of women, but it did so in each age group.

The length of glottalized syllable sequences was also analyzed. Such sequences were found to be of different lengths in both speech modes, with an average of 1,793 syllables in read and 2,751 syllables in spontaneous speech. 66.7% and 66.2% (respectively) of the occurrences consisted of only one syllable. 16.8% and 17.4% of the sequences consisted of two syllables, and 7.0% and 6.4% of them were three syllables long. In the reading task, syllable number in the remaining 9.5% of sequences ranged between 4 and 13, while in spontaneous speech, the remaining 10.0% ranged between 4 and 19 syllables. Trisyllabic or even longer glottalized sequences were found most frequently in middle-aged and elderly women’s speech (Fig. 2 and 3).

![Figure 1: Ratio of glottalized syllables as a function of speakers’ gender, age, and the mode of speech.](image-url)
The boundary marking role of utterance final glottalization has been confirmed by a number of studies for Hungarian. Some studies suggest that stretches of speech at which vocal cords start and finish voicing are motivated for glottalization (see e.g. Dilley et al. 1996; Gordon and Ladefoged 2001; Slifka 2006). Accordingly, the position of glottalized syllables was analyzed within inter-pause intervals of speech in both speech modes, and at the end of turns in spontaneous speech.

Read and spontaneous speech did not differ in terms of the position of glottalized syllables within inter-pause intervals: 20.4% of first syllables(s) were found to be glottalized in read and 17.7% of them in spontaneous speech. Similarly, the final syllables were glottalized in 19.0% and 17.3% of the cases, respectively. The entire inter-pause interval was glottalized in 0.1% and 1.4% of occurrences. The most typical position of glottalization was internal: 60.5% in read and 63.7% in spontaneous speech (Fig. 4). Similar ratios were measured in all gender and age groups.

Turn-final glottalization, however, was frequent. The average number of turns per interviews was 3.1 (between 1 and 9 per speaker). 47.5% of the turns ended with (often several syllables long) glottalized syllables. The question arose if speakers who produced more glottalization than others did so as a means of signaling the end of their turns. The result of the correlation test was not significant.

The average fundamental frequency was defined for each speaker’s spontaneous and read speech production (the data in terms of gender and age groups can be seen in Figure 5). The average $f_0$ is usually lower in the case of spontaneous speech than in reading aloud, and this difference was significant [$t(29) = 3.196; p = 0.007$]. However, in gender-specific comparison, only the male speakers showed significant difference of average $f_0$ in terms of speech mode [$t(14) = 8.227; p = 0.001$]. It was tested if there was a link between the ratio of glottalized syllables and the average fundamental frequency. A correlation was found only in read speech (Pearson’s $r = 0.535; p = 0.033$).

Combining the various analyzed factors in the two-way repeated measures ANOVA, only gender had a significant effect on frequency of glottalization [$F(1, 24) = 136.473, p < 0.001$, Partial Eta Squared = 0.850].

With regard to glottalization in filled pauses, the first question was how often the speakers had produced filled pauses. The total amount of filled pauses was 478 in the whole material of spontaneous speech, the average frequency of filled pauses being 5.0 per minute (SD 2.6). The frequency of filled pauses was strongly speaker-dependent, varying between 1.0 (an elderly woman) and 12.1 (a middle-aged man) filled pauses per minute. The frequency of filled pauses did not show any correlation with the speakers’ age or gender.

The ratio of glottalized filled pauses (with the total number of filled pauses for a given speaker considered as 100%) is also very divergent (Fig. 6). The lowest ratio was 7.7% (a young woman, with only 1 filled pause glottalized out of 13), whereas the highest value was 94.1% (an elderly woman, 16 filled pauses glottalized out of 17).
On average, 60.0% of the filled pauses were partly or totally glottalized in spontaneous speech (SD 21.4%). No correlation was detected between frequency of glottalized syllables and frequency of glottalized filled pauses.

Figure 6: Ratio of glottalized filled pauses in spontaneous speech as a function of speakers’ gender and age.

### 4. Discussion and conclusion

The present study analyzed the frequency, positional and functional characteristics of glottalization in 30 Hungarian speakers’ spontaneous and read speech. The hypothesis was that the ratio of glottalized syllables would be larger in read speech, in women’s speech, in elderly age and with a lower f0.

Frequency of glottalization differed significantly between reading aloud and spontaneous speech, however, it was also highly speaker-dependent. Contrary to the hypothesis, glottalization was more frequent in spontaneous than in read speech, which can be explained by the planning differences between these two speech modes.

In the analyzed corpus the frequency of glottalization exceeded the values measured in earlier studies on Hungarian. While in the literature 10 to 14% of the syllables were found to be glottalized (in reading aloud, see e.g., Böhm and Ujváry 2008; Markó 2011), in the present material the average ratio is above 20% in both spontaneous and read speech. A possible reason behind this is the notable difference between the earlier and the present research in the number of subjects (10–12 vs. 30 speakers).

Results showed that glottalization was more pronounced in female speech and (concomitantly) more frequent with a higher f0 (in reading aloud). Only gender had a significant effect on frequency of glottalization. Presumably, social factors may account for this result. No link was detected either between the frequency of glottalization and age, or between the ratios of glottalized syllables and glottalized filled pauses.

While utterance and turn final position may motivate the occurrence of glottalization, in inter-pause intervals no such tendency could be detected. It thus seems likely that glottalization serves a boundary marking function for some speakers.

### 5. Acknowledgements

The present research was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

### 6. References


ParaLing ’07, pp. 17–22.


Acoustic correlates of masculinity in voices of Lebanese women with Reinke’s edema.

Nayla Matar¹,², Cristel Portes², Leonardo Lancia³, Thierry Legou², Fabienne Baider⁴

¹Faculty of Medicine, Saint-Joseph University, Department of Otolaryngology Head and Neck Surgery, Hôtel-Dieu de France Hospital, Beirut, Lebanon; ²Aix-Marseille Université, CNRS, LPL, UMR 7309, Aix-en-Provence, France; ³Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; ⁴University of Cyprus, Nicosia, Cyprus

nayla.matar@hdj.unj.edu.lb, cristel.portes@lpl-aix.fr, leonardo_lancia@eva.mpg.de, thierry.legou@lpl-aix.fr, helenafab@yahoo.fr

Abstract

Women with Reinke’s edema (WR) complain of a masculine sounding voice. Their voice is interesting to study the gender stereotypes of the human voice. Our objective is to understand the role played by various phonetic cues in the gender ambiguity of the voices of WR. We compared their acoustic parameters to those of normal voices of men (MN) and women (WN). To relate our findings to the complaints of WR, we verified that the differences observed were related to self-evaluation of gender and to perception of gender by naïve listeners. The studied language is Lebanese.

In the production experiment, we asked 56 speakers (28 WN, 18 MN and 10 WR) to produce 3 vowels /a/, /i/ and /u/. The acoustic parameters that distinguish the vocal quality of WR are: low F0, low H1-H2, high entropy of the return times and high subharmonic to harmonic ratio. The vowel space seems reduced in WR. In the self-evaluation task, participants judged their voice and personality separately on a masculinity/femininity scale. WR judge their voice as being masculine and their personality as being feminine. Preliminary results from the perceptual experiment confirm that gender is more ambiguously perceived in WR voices.

Keywords: Voice, Gender, Acoustic parameters, Reinke’s edema, Masculinity, Femininity, Self-evaluation, Perception.

1. Introduction

Gender- defined by Oakley (1972) as the social classification into masculine and feminine - is one of the important information transmitted by voice. As soon as a listener hears a voice, the physical features of that voice interact with the listener’s own schema constructed through sociocultural stereotypes, and a judgment is made about the speaker’s gender. This judgment helps the listener to adapt his attention to the speaker’s personality even at preliminary levels of the perception process and in a mostly unconscious way. Johnson et al. (1999) have demonstrated that listeners integrate abstract gender information with phonetic information in speech perception. When listeners are primed to identify a talker as either female or male, they access gender expectations for what the talker should sound like, and employ these expectations in speech perception: this induces a shift in phoneme boundaries between hood and hud, for example. Such findings show how strongly gender identification is organized in voice perception and how crucial it is in interlocutors’ co-adaptation in language exchange.

Women who smoke may develop a vocal fold pathology known as Reinke’s edema (RE) (Fig.1). These women complain of being called “Sir”, especially in situations where other gender cues (i.e. visual cues) are not available as, for instance, over the phone. A well-documented explanation is that there voices have lower fundamental frequencies (f0) than non-smoking women’s voices (Zeitels, 1997). However other studies show that f0 is not the only clue for the gender of the voice, especially that many women with low f0 are still recognized as women (Andrews, 1997).

So where does the anecdotally reported complaint of women with RE come from? What acoustic parameters, other than f0, make them sound like men?

Men and Women with normal voices represent the two bounds of the gender continuum. Many studies show that the mean value and variations of f0 are the main parameters that distinguish feminine/masculine voices (Andrews, 1997; Avery, 1996; Munson, 2007). Other studies attribute a major role to formant frequencies (FF) (Andrews, 1997; Avery, 1996; Bachorowski, 1999). For example, changing the FF and/or bandwidth while maintaining a constant f0 changes the perception of the voice gender (Assmann 2006). Some studies also link jitter, shimmer and Harmonic to Noise Ratio to gender, a rough/irregular voice being considered more masculine and a breathy voice more feminine (Klatt,1990; Bellandese, 2009). WR as well as men with RE do have higher jitter, shimmer and HNR than normal controls (Zeitels, 1997), however this is also the case with other hoarse patients who do not complain from gender ambiguity of their voice.

Because WR voices are aperiodic and because the features captured by measures like jitter, shimmer and HNR are not appropriate to describe strongly aperiodic voices (Little, 2007), we looked for other acoustic parameters related to amplitude and frequency variations in periodic and aperiodic voices like the entropy of the distribution of the return times of the signal and the scaling exponent obtained through detrended fluctuation analysis. These measures require stationarity of the signals; this is why we analyze here only sustained vowels.

This work aims at improving our knowledge of the role of voice quality features in gender distinction by evaluating their impact on an understudied population. The studied language is Lebanese, for which no studies of gender and voice have been proposed yet.

2. Method

2.1. Speakers and Recording session

Three groups of participants included in the study: 28 WN, 18 MN and 10 WR for a total of 56 participants, with a mean age (+/-standard deviation) of 41.6 (+/-8.8), 40.2 (+/-8.2) and 53.2 (+/-4.6) years respectively. The study was approved by the Ethics Committee at Saint-Joseph University, Beirut, Lebanon (CEHDF 390). Participants signed an informed consent before the experiment started. They were told that the study was on the acoustic parameters of Lebanese. They were not instructed on the gender aspect in the beginning to avoid introducing a bias. However, at the end, they were informed about the real aim of
the study and were given the opportunity to withdraw their consent and have their data destroyed. None opted to do so.

As for the WR population, the diagnosis of Reinke’s edema was done after endoscopic examination of the vocal folds by a trained laryngologist using a video endoscope (Storz, Germany). The difference with the appearance of normal vocal folds is shown in Fig. 1.

![Image of vocal folds with Reinke’s edema and normal vocal folds](image)

Figure 1: (Top) Vocal Folds with Reinke’s edema, (Bottom) Normal vocal folds.

Instructions were presented to speakers on a laptop screen using the E-prime software. The task was performed in a quiet room with minimal ambient noise. We used a H2 Handy Recorder (Zoom Corporation, China) with a LEM-headset microphone (LEM P2416, LEM Industries). Recorded signals were sequenced and annotated using the PRAAT software (Boersma, 2001). After a test trial, subjects were instructed to produce 3 stable vowels /ɪ/, /ʌ/ and /ʊ/ continuously for 3 seconds, 6 words, and 12 short sentences using habitual vocal pitch and loudness. There were 3 repetitions for each stimulus. Only the vowel parameters are reported here, words and sentences will be studied later.

### 2.2. Production study

Most voice quality measures were collected using the Voice Sauce program for Matlab (Shue, Chen and Alwan, 2011): these were F0, the first 4 formants and relative bandwidth, the RMS energy, the cepstral peak prominence (CPP) (Hillenbrand, 1987), the harmonic to noise ratio (HNR) in 4 different frequency bands (0-500 Hz, 0-1.5 kHz, 0-2.5 kHz, 0-3.5 kHz), the subharmonic to harmonic ratio (SHR) related to the complexity of the subharmonic structure of the voicing signal, and 5 measures of spectral tilt (H1-H2, H2-H4, H1-A2, H1-A3, H1-A4, corrected following Hanson, 1997). We also measured the Entropy of the distribution of the return times of the voicing signal estimated trough recurrence analysis and the scaling coefficient from detrended fluctuation analysis. These nonlinear measures capture temporal and amplitude variability in the voicing signal better than the classical jitter and shimmer measures especially in aperiodic sounds (Little, 2007).

We run a separate linear mixed model for each measure and handled pseudo replication trough Bonferroni correction (applied to the p values obtained through Monte Carlo Markov Chain sampling). Significance threshold was set to 0.05. In each model the predictors where the class of the speaker (MN, WN and WR with WR as reference level), the vowel quality (/ɪ/, /ʌ/ /ʊ/ using the average across the vowels as reference level) and the interactions between these two predictors when its inclusion increased the model fit to the data (according to Chi square tests). The random terms were the speakers’ identity and a speaker-specific effect of the vowel quality.

### 2.3. Self-evaluation study

The self-evaluation of voice gender can be performed by using different types of scales including one scale with a masculine and feminine ends or two different scales, one for femininity and one for masculinity. We used a unique scale with 1= maximal masculinity and 9= maximal femininity. In doing so, we adopted the postmodern theory (Butler, 2008; Perry, 2001; Owen, 2011) that states that gender perception is made on a continuum and is different from the suggestion that women’s and men’s voices are treated as different entities (Pepiot, 2011).

The self-evaluation scores of the gender of voice and personality were compared between the groups using the non-parametric Kruskal-Wallis and Mann-Whitney tests. The Wilcoxon signed rank test was used to compare the self-evaluation of voice gender score with the self-evaluation of personality gender score within each group. The significance threshold was set to 0.05.

### 2.4. Pilot perception experiment

We present preliminary results of a perception study that we are conducting in order to verify if the gender ambiguity of WR is perceived by naive listeners. At present, the experimental material includes the vowels of 5 speakers of each of our 3 groups. Vowels produced by each of the speakers, were presented randomly to a panel of naive listeners (4 women, mean age=23 [SD: +/-1] years; 5 men, mean age=24 [SD: +/-1] years). They were asked to judge each vowel using a button box with 1=surely feminine, 2=most probably feminine, 3=cannot decide, 4=most probably masculine, 5=surely masculine. For women, the ratings 1 and 2 were considered correct and the ratings 3, 4, 5 were considered incorrect. For men, the ratings 4 and 5 were considered correct and the ratings 1, 2, 3 were considered incorrect.

### 3. Results

#### 3.1. Production study

The F1 and F2 of the vowels /ɪ/, /ʌ/ and /ʊ/ of the three groups are plotted in Fig. 2. The area of the vocal space (VSA) spanned between F1 and F2 of the 3 vowels was calculated using the following formula (Weismer, 2001):

\[ VSA = 0.5 \cdot \text{ABS} (F1 + F2a - F2a) + F1a(F2a - F2i) + F1a(F2i - F2a) \]
In comparison to the area of the vocal space of WN (136196 Hz²) and MN (147225 Hz²), the area of the vocal space of WR (60707 Hz²) seems reduced and flattened along the direction of covariation of the two formants. Indicating that /u/ is somewhat centralized in these speakers.

The total number of observations included in the statistical analysis is 349. The following acoustic parameters were significantly affected by the speakers' class across vowels: the entropy of the return times is higher for WR than for MN and WN (tWN,WR=−2.942, p=0.0435; tMN,WR=−4.306, p<0.001). F0 is higher for WR than for MN (tWN,WR=−3.354, p=0.026) but it is lower for WR than for WN (tWN,WR=4.484, p<0.001). The SHR is higher in WR speaker than in WN speakers (tWN,WR=−3.588, p=0.009). The difference between the amplitude of the first and second harmonics (H1-H2) is lower for WR than for WN (tWN,WR=3.262, p=0.04).

![Formant F2](image)

**3.2. Self-evaluation study**

The self-evaluation scores of voice were: 6.9 (+/-1.3) for WN, 3.2 (+/-2.0) for MN and 3.4 (+/-2.2) for WR. It was significantly different between the three groups (Kruskal-Wallis test, p<0.001) with WR ranking their voice as masculine as MN (Mann-Whitney test, p=0.8). The self-evaluation of personality was: 7.0 (+/-1.8) for WN, 2.9 (+/-2.0) for MN and 6.3 (+/-1.8) for WR. WR ranked their personality as feminine as WN (Mann-Whitney test, p=0.05). For WR, there was a negative correlation between the self-reported voice and personality (r=−0.75) (Wilcoxon signed rank test, p=0.03). This self-evaluation study demonstrated that our group of WR describes their voice as masculine and this is in contradiction with the way they describe their personality. A diagram of the self-evaluation of voice and personality is shown in Fig.3.

![Figure 2: F1 and F2 of the vowels /i/, /a/ and /u/ of the 3 groups.](image)

**3.3. Pilot perception experiment**

Naive listeners rated MN correctly in 97% of the time; they rated WN correctly in 89% of the time, and WR in 58% of the time. WR were more often rated as masculine compared to WN. It is also important to note that the rating of women in general was more often incorrect that the rating of men. A diagram of the ratings is shown in Fig.4.

![Figure 4: Perception study: Correct and Incorrect rating of voices of WR, MN, WN. Error bars represent two standard deviations of measurement.](image)

**4. Discussion and Conclusion**

This is the first study on gender in Lebanese Arabic voices. We have demonstrated, through the self-evaluation study, that WR describe their voice as masculine whereas they describe their personality as feminine. This observation confirms what has been previously reported anecdotally in the literature about the complaint of WR concerning the gender ambiguity of their voices. Our results in WR can be related to those on transgender persons who have a conflict between their self-perception of a feminine personality with a masculine voice (Owen, 2011).

We found that f0 is lower in WR than in WN as previously mentioned in the literature (Zeitels, 1997, Lim 2006). Conversely, no significant differences were found in formant frequencies between our 3 groups although many authors (Andrews 1997, Avery 1996, Bachorowski 1999, Fant 1975) previously demonstrated that higher formants, especially in low vowels, are related to feminine voices. This finding might be explained by the fact that vocal tract dimensions of WR are close to those of WN. However, we do not have enough data to compare the vocal tract dimensions of WR, MN, WN. By looking at the area of the vocal space it seems that WR have a smaller vocal space than WN. This finding might contribute to the masculine perception of WR since it is known that females have larger vocal space than males (Weirich, 2013) in American English (Hillenbrand, 1995), in British English (Whiteside, 2001) in German (Simpson, 2007), and that masculine sounding women have a more retracted vowel space (Munson 2006). Note however that the vowel space of WN and MN are similar in our data. We are currently exploring this difference in vowels produced in more natural contexts (words and sentences).

Three measures related to the voice quality have been found to distinguish the voices of WR: H1-H2, entropy of the return times and SHR. H1-H2 is related to the relative length of the open glottal phase, its lower value in WR indicates a short open phase and less breathy voice. The entropy of the return times captures temporal variability in the voice signal. It is higher for WR than for MN and WN. It reflects the
aperiodicity of the voice that may be linked to the roughness of the voices of WR. Roughness has already been linked to masculine sounding voices (Klatt, 1990; Bellandese, 2009). The SHR is higher in WR speaker than in WN speakers. It indicates the complexity of the sub harmonic structure of the voicing signal that has been correlated not only to voice quality but also to pitch perception. When the ratio is smaller than 0.2, the subharmonics do not have effects on pitch perception. As the ratio increases approximately above 0.4, the pitch is mostly perceived as one octave lower that corresponds to the lowest subharmonic frequency. When SHR is between 0.2 and 0.4, the pitch seems to be ambiguous (Sun, 2002). In our study the SHR is 0.7 in WR which could contribute to the perception of WR voices as one octave lower than it is actually, closer to the pitch of the voices of MN. It also reveals a rough vocal quality that is more correlated to masculine voice (Bellandese, 2009) as previously stated.

Finally, the preliminary results from the perception study confirm the ambiguity of WR's voices with respect to gender information. Overall our results open several perspectives for future developments. In order to establish more links between our findings and the gender ambiguity of WR's voices, we will conduct, in the future, correlation studies to verify if there is a correlation between the gendered perception of voice and these acoustic parameters; if there is a correlation between the self-evaluation of voice and personality of the speakers and the perception of voice by naïve listeners; if there is a correlation between the self-evaluation of voice and these acoustic parameters. We will also be including a group of hoarse women without a history of smoking to see if the complaint of WR is unique or shared by other hoarse women. We will also study the acoustic parameters in vowels extracted from connected speech because we are aware that sustained connected speech because we are aware that sustained

correlation and individual talker identity are present in a short vowel segment produced in running speech”. J Acoust Soc Am; 106 (2): 1054-1063.


Reaching a Goal with Limited Means:

A Study of Contrastive Focus in Children and Adults

Lucie Ménard1, Amélie Prémont1, Paméla Trudeau-Fisette1, Christine Turgeon1, Dominique Côté1, and Mark Tiede2

1Laboratoire de phonétique, Département de linguistique, UQAM and Center for Research on Brain, Language, and Music, Montréal, Canada
2Haskins Laboratories, New Haven, USA

menard.lucie@uqam.ca, amelie.premont@gmail.com, ptrudeauifisette@gmail.com, turgeon.christine.2@courrier.uqam.ca, cote.dominique.12@courrier.uqam.ca, tiede@haskins.yale.edu

Abstract

This paper deals with articulatory and acoustic correlates of contrastively accented syllables in French, in children and adults. Ten 4-year-old children and ten adults were recorded while producing repetitions of /bVb/, /bVb/, /bVb/, /bVb/ sequences in two prosodic contexts: in a neutral condition and under contrastive focus. Synchronous recordings of tongue movements in the midsagittal plane; lip and jaw positions; and speech signals were made using an ultrasonic device (Sonosite 180 Plus), a Northern Digital Optotrak system, and a directional microphone. Lip positions and tongue shapes were analyzed at vowel midpoints in both prosodic contexts. Formant frequencies, RMS amplitude, F0, and duration were also extracted from the acoustic signals. Results showed that focused vowels were higher, louder, and longer than their neutral counterparts for all participants. However, the effect of focus on lip position was smaller for children than for adults. Regarding tongue shape and position, no effect of prosody was found in children, unlike adults. Results are discussed in terms of motor control development.

Keywords: speech development, prosody, focus

1. Introduction

During the course of speech development, sequential mastery of larynx mechanisms, jaw position, and tongue shapes are involved in developing the vocal repertoire of young speakers (Vihman, 1996). In typically-developing children, control of glottal and supraglottal mechanisms is achieved at different stages of development. Laryngeal control, examined through produced fundamental frequency (F0) contours and final lengthening patterns, is acquired quite early during infancy, whereas the inter-articulatory dynamics of the lips, jaw, and tongue are not adult-like until the end of the first decade of life. It is striking to observe that despite the different endpoints of these developmental trajectories, children produce intelligible sound sequences by 4 years of age. What are the strategies they use to reach intelligibility?

To explore the links between intelligibility and speech development, we conducted an experiment using prosodic contrasts as a means of manipulating distinctiveness. It has been reported that in contrastively accented constituents, phonemic contrasts are enhanced, thus increasing the constituent’s saliency for the listener (Cho & McQueen, 2005; de Jong, 1995; Hay, Sato, Coren, Moran, & Diehl, 2006). In such prosodically strong conditions, speakers often hyperarticulate the canonical speech features associated with vowels and consonants (Cho, Lee, & Kim, 2006). As suggested by Hay et al. (2006), the different strategies used to increase perceptual saliency can be explored by investigating distinctiveness-enhancing contexts such as focus (Ménard et al., 2013). In 4-year-old children, it has already been reported that lip gestures are less affected by prosodic focus than they are in adults (Grigos and Patel, 2010; Ménard et al., 2006). We hypothesized that during the course of speech development, weighting related to a lip or tongue gesture would depend on the degree of maturity of control for that gesture. Different gestures would thus be enhanced to signal focus in children versus adults. The objective of this paper was thus to explore the strategies used to enhance perceptual saliency in 4-year-old children and in adults, by studying the articulatory and acoustic correlates of contrastive focus in French.

2. Method

2.1. Participants and corpus

Ten typically developing 4-year-old children and ten adults were recruited. In this paper, we report data from 6 children and 8 adults. All subjects were native speakers of Canadian French who lived in the Montreal area. All participants passed a 20-decibel-hearing-level (dB HL) pure-tone screening procedure at 250, 500, 1000, 2000, and 4000Hz. The corpus consisted of the four French vowels /i y u a/ in /bVb/ syllables. The target syllables were embedded in carrier sentences of the type “Je suis bVb, le musicien” (“I am bVb, the musician”). The sentences were part of a story. Ten repetitions of each sentence were obtained in each of the following prosodic conditions: neutral manner (without any particular accent, for instance: “My friend Bab plays music.”) and contrastive focus (for instance, the word BAB in: “My friend BAB plays music, not my friend Ken.”). Ten repetitions of each sentence, in each of the two prosodic conditions, were performed.

2.2. Experimental procedure

Synchronous recordings of tongue movement in the midsagittal plane (at NTSC 29.97 Hz) and of the speech signal (at 44.1 kHz) were made using an ultrasound device (Sonosite 180 Plus) and a multidirectional microphone. A Northern Digital Optotrak system was used to concurrently record audio and the positions of infrared emitting diodes (iREDs) placed on the lips and chin. iREDs were also positioned on the ultrasound probe and on the forehead of participants, to provide a representation of the data in a movement-corrected head-centric frame of reference (Whalen et al., 2005). The experimental setup is shown in Figure 1(a). After head-movement correction and remapping to a coordinate system
centered on the upper incisors and aligned with the occlusal plane, the data were projected onto a 3D view in which the position of the iREDs and the tongue imaging plane were visible. The acoustic onset and offset of the target vowel were labelled on the acoustic signal and the iRED coordinates were extracted at the vowel midpoint. Two parameters were used to characterize lip geometry: lip opening (distance, in the vertical dimension, between the upper lip iRED and the lower lip iRED) and upper-lip protrusion (distance, in the horizontal dimension, between the upper lip iRED and the reference position).

Tongue surface contours were measured using EdgeTrak (Li et al., 2003). Prior to the analysis, a pixel-to-cm ratio was calculated and used to convert all (x, y) coordinates from pixels to cm. Tongue shape and position were analyzed using the method described in Ménard et al. (2013). Only two parameters will be discussed in this paper. First, tongue front-back position corresponded to the x coordinate of the highest point of the contour (point C in Figure 1b). Second, tongue curvature degree was measured by the ratio of segments CD over AB, as shown in Figure 1(b). Regarding acoustic measures, the first three formant frequencies were extracted at the vowel midpoint, using the LPC algorithm implemented in the Praat speech analysis software (Boersma & Weenink, 2014). F0 values, duration, and RMS were also measured at the vowel midpoint. Formant and F0 values were transformed into mel units.

![Figure 1](image1.png)

**Figure 1:** (a) Position of the IREDs. (b) Parameters used to quantify tongue position and shape. The solid line represents the tongue contour and the dashed lines represent a triangle that fits the contour. Tongue front-back position corresponds to point C. Tongue curvature corresponds to the ratio CD/AB.

3. Results

3.1. Acoustic results

Differences between F0 values, in mels, between vowels produced in the focus condition versus vowels produced in the neutral condition are presented in Figure 2. Data were averaged across vowels and speakers. As suggested by this figure, repeated-measure ANOVAs conducted on F0, duration, and RMS values with speaker group (children vs. adults) as the between-subject factor and prosodic condition (focus vs. neutral) as the within-subject factor revealed a significant effect of prosodic condition ($F(1,12)=52.25; p<0.001$). In focused contexts, vowels were consistently higher, longer, and louder. A significant effect of speaker group was only found for F0, for which values were higher in children than in adults ($F(1,12)=37.10; p<0.001$). More importantly, no interaction with speaker group was found, suggesting that children and adults use those parameters to the same extent when going from the neutral to focused conditions.

![Figure 2](image2.png)

**Figure 2:** Average values of F0 (mels) (upper left graph), duration (msec) (upper right graph), and RMS (lower graph) in focused and neutral vowels.

Formant values for vowels are presented in Figure 3, for adults (left graph) and children (right graph). Dispersion ellipses enclose ± 1.5 standard-error-of-the-mean values.
Repeated-measure ANOVAs conducted separately on F1 and F2 values with speaker group (children vs. adults) as the between-subject variable, prosodic condition (focused vs. neutral) and vowel (/i/, /y/, /a/, /u/) as the within-subject variables revealed a significant main effect of speaker group, on F1 (F(1,12)=204.05; p<0.001) and on F2 (F(1,12)=125.90; p<0.001). As expected from anatomical differences between children and adults (Vorperian et al., 2009; Ménard et al., 2007), children’s F1 and F2 values were higher than the F1 and F2 values for adults. Furthermore, the vowels were significantly different in terms of F1 (/a/ being higher than /i/, /y/, and /u/ for children and for adults) (F(3,36)=574.22; p<0.001), and in terms of F2 (/u/ having higher F2 than /y/ and /a/, which in turn had higher F2 than /u/) (F(3,36)=478.01; p<0.001). This expected difference reflects the French phonological implementation of vowels in the acoustic space.

More importantly, a significant interaction between speaker group and prosodic condition was found for F1 (F(1,12)=7.07; p<0.05). Participants significantly increased F1 values when going from the neutral to the focused condition, the more so for adults than for children. No interaction effect was observed on F2, at the group level.

3.2. Articulatory results

Regarding the articulatory strategies used to implement these differences, lip opening and upper lip protrusion values are presented for the four vowels in both prosodic conditions in Figure 4 (left graph: adults; right graph: children).

4. Discussion and conclusion

The results in this study showed that the articulatory and acoustic effects of contrastive focus differ for 4-year-olds and adults. First, it was shown that both speaker groups used increased loudness, pitch, and duration to produce focused syllables compared with neutral syllables. However, lip and tongue gestures did not significantly differ in the focused and neutral contexts for children’s vowels, in contrast to adults’ vowels. This articulatory difference was reflected in the formant patterns. This result agrees with previous studies showing that at 4 years of age children have not yet mastered...
the full scale of articulatory contrast related to prosodic prominence (Grigos and Patel, 2010; Smith and Goffman, 1998, Ménard et al., 2006). In the current study, adult speakers enhanced their articulatory gestures, together with F0, to signal prominence, and they also hypoarticulated the neutral syllable, as seen by the reduction of their lip openings. Such a pattern was not observed in children, suggesting that they do not exploit the lower range of articulatory contrasts. Thus, the observed undifferentiated kinematic patterns in the different conditions could be due to the inability to reduce unfocused syllables (Allen and Hawkins, 1980). The greater values of lip opening, as depicted in Figure 4, suggest that, in line with Smith and Goffman’s observations (1998), children produce relatively large articulatory movements considering the size of their oral structures, compared to adults. Children have generally learned to exploit the higher range of the hyper-hypoarticulation scale, but the ability to reduce the magnitude of articulatory movements (i.e., exploiting the lower range of the scale) comes later in the acquisition of motor control, presumably because it may require fine-tuning of articulation. However, the fact that some effect of focus on F1 patterns was found in children (although limited compared to adults), despite any evidence of lip or tongue differentiation related to prosodic context, deserves further analysis. Individual analyses relating articulatory parameters and acoustic effects are currently underway.

Overall, the results observed in a distinctiveness-enhancing context such as contrastive focus suggest that speech representations in 4-year-old children are shaped by motor constraints.

5. Acknowledgements

This work was supported by grants from the Social Sciences and Humanities Research Council of Canada and from the Natural Sciences and Engineering Research Council of Canada. Thanks to Marlene Busko for copy-editing the paper.

6. References


Acoustic analysis and automatic detection of laughter in Hungarian spontaneous speech

Tilda Neuberger, András Beke, Mária Gósy

Research Institute for Linguistics of the Hungarian Academy of Sciences
neuberger.tilda@nytud.mta.hu, beke.andras@nytud.mta.hu, gosy.maria@nytud.mta.hu

Abstract

The accuracy of speech recognizers may decrease in the case of spontaneous speech because of non-verbal vocalizations such as laughter. Previous studies showed that laughter resembles to speech sounds in terms of their acoustic characteristics. The aim of the present research is to perform for the first time for Hungarian language an acoustic analysis of the differences between laughter and speech, and to develop an efficient method in order to differentiate laughter from speech segments. The classification was carried out in the BEA Hungarian spoken language database, using different classifiers (GMM, SVM, ANN). The results show that the GMM–SVM system trained with short-term features and acoustic parameters seems to be the best method for solving this problem.

Keywords: laughter, machine learning, spontaneous speech

1. Introduction

The non-verbal communication plays an important role in human speech comprehension. Spontaneous speech frequently contains non-verbal vocalizations; one of the most frequent phenomena is laughter. Laughter is an inborn, species-specific indicator of affection that provides information about the emotional state of the speaker. It is a socially constituted and easily decodable phenomenon that has various functions in everyday conversations. In meetings, it serves to regulate the flow of the interactions (back-channel sound sequences), mitigates the semantic context of the preceding utterance, or serves as a stress-reducing strategy (Rothgäns et al. 1998). It can also provide cues to semantic content of the utterances in conversations, such as jokes or topic changes, and can indicate the boundary of the turns; hence laughter and other non-verbal sounds are of increasing interest in discourse analysis.

Laughter was investigated by researchers coming from various fields. It was observed in relation to the psychology of humor (e.g. Goldstein and McGhee 1972; Holland 1982; Martin 2007), on the one hand, and its acoustic properties were studied (Bachorowski et al. 2001; Bickley and Hunnicutt 1992; Rothgäns et al. 1998), on the other hand. The perceptible sound sequence(s) of typical laughter is usually like those of breathy CV syllables (e.g., /hV/ syllable). This sound sequence might acoustically resemble to speech sounds due to the fact that they are produced by the same human vocal tract. Speech and laughter were found to be quite similar to each other in their ‘syllable’ durations, formant structure, and Root-Mean-Square (RMS) energy (Bickley and Hunnicutt 1992). Previous studies found that the average duration of laughter appeared between 395 ms and 915 ms, and the mean fundamental frequency of laughter was between 160 Hz and 502 Hz in female and between 126 Hz and 424 Hz in male speakers (Bachorowski et al. 2001).

Speech recognition research focused on non-lexical sounds, because these events may cause an important decrease in the accuracy of speech recognizer in spontaneous speech compared to read speech (Butzberger et al. 1992; Furui 2005). In addition, this decrease in accuracy is caused by the phonetic variations of spontaneous speech together with occurrences of disfluencies. Various types of features (spectral, cepstral, prosodic, perceptual ones) were investigated for laughter detection using diverse classification techniques. Gaussian Mixture Models (GMMs) were trained with Perceptual Linear Prediction (PLP), pitch and energy, pitch and voicing, and modulation spectrum features to model laughter and speech by Troung and van Leeuwen (2005). Their results showed equal error rates ranging from 7.1% to 20.0% of the cases.

For detection of overlapping laughter, Kennedy and Ellis (2004) used Support Vector Machines (SVMs) trained on four features: MFCCs, delta MFCCs, modulation spectrum, and spatial cues. They achieved a true positive rate of 87% of the cases.

Troung and van Leeuwen (2007) developed a gender-independent laughter detector using different classification techniques and also their fusion (GMM, SVM, MLP) with various types of features. They observed that SVM performs better than GMM in most of the cases, but the fusion of the classifiers improved the performance of the classification (lower equal error rate of around 3% were obtained).

The laughter detector (using Hidden Markov Models) developed by Campbell et al. (2005) can automatically recognize four types of laughter (hearty, amused, satirical, or social laugh) in Japanese (the identification rate is larger than 75% of the cases).

Presegmented laughter and speech segments were classified appropriately in 88% of the test segments using HMMs by Lockerd and Mueller (2002). Cai et al. (2003) also modeled laughter using HMMs with MFCCs and perceptual features (short-time energy, zero crossing rate). These methods achieved average recall and precision percentages of 92.95% and 86.88%, respectively.

Campbell (2007) measured pitch, power, duration, and spectral shape in the analysis of laughter and laughing speech. Neural networks (ANN) were successfully trained to identify the nature of the interlocutor (social or intercultural relationships).

Knox’s and Mirghafori’s (2007) method for non-presegmented frame-by-frame laughter recognition produced an equal error rate of 7.9% of the cases. They used ANN trained on MFCC, AC PEAK, and F0 features.

Gupta et al. (2013) used deep neural network (DNN) classifier and also time series filtering and masking for laughter and filler detection. They obtained area under receiver operating characteristic (ROC) curve of 93.3% for laughers and 89.7% for fillers on the test set.

For the first time for Hungarian language, the present study investigates the acoustic characteristics of laughter and
compares them to those of speech. Our aim is to develop an accurate method in order to recognize laughter events in spontaneous speech by differentiating laughter and speech. In our experiment various classifiers were combined with each other to achieve high-level classification performance.

2. Subjects, material and method

We used the BEA Hungarian Spoken Language Database (Gósy 2012) to train and test the detector. It is the largest speech database in Hungarian, which contains material of 260 hours produced by 280 speakers (aged between 20 and 90 years). For the present study, we used the conversational speech material (of 35 males and 40 females): a total of 75 meetings, whose average duration was 16 minutes. The laughter segment boundaries were identified by human transcribers; the definition of onset and offset of the segments was not the task of the classifier. Laughing can occur while speaking (speech-laughs or laughed speech, see e.g., Bickley and Hunnicutt 1992; Trouvain 2001); in the present study these phenomena were excluded. Our presegmented data contains 653 manually annotated segments: 332 laughter and 321 speech segments (in this case: randomly selected words from the BEA word-level annotation of spontaneous conversations). We used 1/3 of the data in the testing set and 2/3 of them in the training set.

2.1. Features

The acoustic parameters (APs): F0 mean and SD, jitter, shimmer, RMS, HNR, ZCR, spectral slope, LPC-CoG were extracted using Praat voice analysis software (Boersma and Weenink 2011). The short-term features: MFCC (12 coefficients (1–12) + log energy + Δ + ΔΔ) were extracted using MATLAB 7.12 software. All features were extracted in 25 ms Hamming window shifted with 10 ms steps. The statistical analysis was carried out by SPSS 13.0 software.

2.2. Feature selection

For the present research, selection of features was based on Receiver Operating Characteristic (ROC) curves and the Area Under Curve (AUC) measure calculated for the ROC (based on Wang and Tang 2009). We applied this feature selection method to all features used (APs, MFCC) only in the training set.

2.3. Classification method

Classifiers developed with different algorithms or features may be able to complement each other. We used the following types of combined method: GMM–ANN and GMM–SVM for laughter and speech segment classification. For normalization of the scores the classifier UBM was used. Both laughter and speech are modeled by GMM (the number of components was 4 and we used diagonal covariance of GMM) and adapted to Universal Background Model (UBM). UBM is trained on large set of elements from the database. In this experiment only means were adapted to UBM based on MAP (Maximum A-Posteriori Probability) method. After GMM–UBM modeling the so-called supervector can be obtained by concatenating each of the mixture component mean vectors (Campbell et al. 2006). This GMM-supervector was the input feature for the SVM and ANN. In the case of testing procedure the same processing was applied as in the training method. In case of SVM algorithm RBF (Radial Basic Function) kernel was used. The RBF has two parameters which have to be optimized. These parameters are the kernel’s parameter, and the soft margin parameter C. 3-fold cross-validation and grid search method were used on the training set to choose the optimal combination of these parameters. In case of ANN 2-layers Multilayer Perceptron (MLP) was used. Back-propagation function was used for the training in MLP. We calculated the performance of the classifier using the test set in various ways. We measured the accuracy (acc.), precision (prec.), recall (rec.), F-measure (F-m.) of the classifiers. The main performance value was represented by EER and it was illustrated by means of DET (detection error tradeoff) curve (see Martin et al. 1997).

3. Results

The average duration of laughter was 911 ms and the standard deviation was 605 ms (Fig. 1). For comparison, the average laughter duration was 1615 ms with a standard deviation of 1241 ms in the Bmr subset of the ICSI Meeting Recorder Corpus (Knox and Mirghafori 2007).

We measured various features to characterize the laughter and speech segments, and analyzed them statistically (using ANOVA) to identify those features that provide significant differences between laughter and speech segment. The results showed that there were significant differences in mean values between laughter and speech segment in most of the cases (Jitter: F(1, 651)=51.15; p=0.001; Shimmer: F(1, 651)=27.78; p=0.001; HNR: F(1, 651)=38.93; p=0.001; Spectral slope: F(1, 651)=0.02; p=0.894; LPC CoG: F(1, 651)=10.37; p=0.0013; F0 mean: F(1, 651)=655.88; p=0.001; ZCR: F(1, 651)=46.14; p=0.001; RMS: F(1, 651)=79.08; p=0.001).

The separation ability of the APs was tested using ROC method (Fig. 2).

The F0 had the largest discriminating power out of all tested APs. F0 values were found to be 207±49 Hz in males’, and 247±40 Hz in females’ laughter, while 165±45 Hz in males’ and 198±39 Hz in females’ speech segments. Previous studies showed that laughter was highly variable in its fundamental frequency. The mean F0 values of laughter are different across studies: 138 Hz in men and 266 Hz in women (Bickley and Hunnicutt 1992), 284 in man and 421 Hz in women...
GMM–SVM and GMM–ANN systems were trained with various sets of features. We wanted to learn which of the classifiers and which of the features provides the best results. The results show that classifiers based on MFCC provided the best result, while the poorest result was yielded by the classifier based on APs (Fig. 3). The DET plot shows that GMM combined with SVM gives better result than GMM–ANN in all cases (Table 1).

We trained and tested our classifier systems with different combinations of features. Previous research showed that the classifier method trained with short-term and long-term combined features can improve the performance (Knox and Mirghafori 2007). We also combined the short-term features with the long-term features (MFCC+APs). The results show that in the case of GMM–SVM classifier trained with combined features gave better result than that with single features. The EER value decreased from 3.72% (trained with MFCC) to 2.5% (trained with combined features), while in the case of GMM–ANN system there no decrease was found (Fig. 4). Hence, the best result was obtained by GMM–SVM system using MFCC+APs combined features (EER: 2.5%) (Table 2).

In summary, the best result was obtained by GMM–SVM system using MFCC+APs combined features (EER: 2.5%). For comparison, Knox’s and Mirghafori’s (2007) method (ANN trained on MFCC and F0 features) for non-presegmented frame-by-frame laughter recognition produced an equal error rate of 7.9%. Troung’s and van Leeuwen’s (2007) results revealed that the fusion of techniques (GMM, SVM, MLP) improved the performance of the classification; lower equal error rate was around 3%.

Table 2: The performance of the GMM–SVM classifier trained with various features

<table>
<thead>
<tr>
<th></th>
<th>GMM-SVM</th>
<th>GMM-ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFCC</td>
<td>APs</td>
</tr>
<tr>
<td>Acc. (%)</td>
<td>97.37</td>
<td>95.79</td>
</tr>
<tr>
<td>Prec. (%)</td>
<td>98.95</td>
<td>100</td>
</tr>
<tr>
<td>Rec. (%)</td>
<td>95.92</td>
<td>95.74</td>
</tr>
<tr>
<td>F-m. (%)</td>
<td>97.41</td>
<td>95.83</td>
</tr>
</tbody>
</table>

The results indicated that the classification of laughter and speech can be solved more efficiently using GMM–SVM than using GMM–ANN which is supported by the higher accuracy value and lower EER value. In the training set (Fig. 5) the GMM–SVM gave lower error level than GMM–ANN.

<table>
<thead>
<tr>
<th></th>
<th>GMM-SVM</th>
<th>GMM-ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFCC+APs</td>
<td></td>
</tr>
<tr>
<td>Acc. (%)</td>
<td>86.73</td>
<td>77.89</td>
</tr>
<tr>
<td>Prec. (%)</td>
<td>95.51</td>
<td>72.58</td>
</tr>
<tr>
<td>Rec. (%)</td>
<td>91.05</td>
<td>91.84</td>
</tr>
<tr>
<td>F-m. (%)</td>
<td>90.91</td>
<td>81.08</td>
</tr>
</tbody>
</table>

We trained and tested our classifier systems with different combinations of features. Previous researches showed that the classifier method trained with short-term and long-term combined features can improve the performance (Knox and Mirghafori 2007). We also combined the short-term features with the long-term features (MFCC+APs). The results show that in the case of GMM–SVM classifier trained with combined features gave better result than that with single features. The EER value decreased from 3.72% (trained with MFCC) to 2.5% (trained with combined features), while in the case of GMM–ANN system there no decrease was found (Fig. 4). Hence, the best result was obtained by GMM–SVM system using MFCC+APs combined features (EER: 2.5%) (Table 2).

In summary, the best result was obtained by GMM–SVM system using MFCC+APs combined features (EER: 2.5%). For comparison, Knox’s and Mirghafori’s (2007) method (ANN trained on MFCC and F0 features) for non-presegmented frame-by-frame laughter recognition produced an equal error rate of 7.9%. Troung’s and van Leeuwen’s (2007) results revealed that the fusion of techniques (GMM, SVM, MLP) improved the performance of the classification; lower equal error rate was around 3%.

Table 1: The performance of the GMM–SVM and the GMM–ANN classifiers trained with different features

<table>
<thead>
<tr>
<th></th>
<th>GMM-SVM</th>
<th>GMM-ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFCC</td>
<td>APs</td>
</tr>
<tr>
<td>Acc. (%)</td>
<td>86.73</td>
<td>92.86</td>
</tr>
<tr>
<td>Prec. (%)</td>
<td>95.51</td>
<td>77.89</td>
</tr>
<tr>
<td>Rec. (%)</td>
<td>91.05</td>
<td>72.58</td>
</tr>
<tr>
<td>F-m. (%)</td>
<td>90.91</td>
<td>81.08</td>
</tr>
</tbody>
</table>

The results indicated that the classification of laughter and speech can be solved more efficiently using GMM–SVM than using GMM–ANN which is supported by the higher accuracy value and lower EER value. In the training set (Fig. 5) the GMM–SVM gave lower error level than GMM–ANN.

Figure 3: DET plot of GMM–SVM and GMM–ANN classifier trained with various features

Figure 4: DET plot of GMM–SVM and GMM–ANN classifier trained with combined features

Figure 5: Illustration of two-dimensional plots of the laughter and speech segments based on GMM–ANN (top) and GMM–SVM (bottom)
4. Discussion and conclusion

The present study aimed at classifying laughter in spontaneous speech. We could determine the most efficient parameters and the most accurate method for laughter detection using ROC AUC and EER analyses. We investigated several feature sets whereby MFCC gave the best result. The reason of the poor classification performance of the APs might be explained by the only relatively high discriminating power of F0. The high AUC value and the low EER value confirmed that the GMM–SVM system based on short-term features is suitable for automatic classification of the laughter segments and speech segments in Hungarian spontaneous speech. We tested whether a classifier trained with combined features reduces the EER values. We could observe that GMM–SVM using combined features gives lower EER values than the classifier trained with single features. It means that both short-term and long-term features play an important role in laughter recognition.

In all cases the GMM–SVM gave higher accuracy than GMM–ANN. Previous research has showed that SVM is more effective than ANN (Ding and Dubchak 2001; Makkamala et al. 2002; Tyagi 2008) due to the following reasons: (i) in the case of many ANN the classifier function is influenced by the data set while the SVM algorithm is not (because SVM depends only on support vectors), (ii) the advantage of SVM is the ability to reduce the high number of dimensions of features due to the utilization of kernel functions, (iii) in choosing parameters, the SVM is less complex than ANN, (iv) ANNs use the empirical risk minimization principle which might lead to worse generalization than SVM does (Chen et al. 2003).

The integration of new features and further algorithms concerning acoustic similarities and differences between laughter and speech will be used in our planned future study. We plan to test our model using corpora of other languages because culture specific laughter is supposed to exist (Campbell et al. 2005).

5. Acknowledgements

This work was supported by OTKA 108762.

6. References


Segmental and positional effects in tonal alignment: An articulatory approach
Henrik Niemann, Martine Grice & Doris Mücke

IUL Phonetik, University of Cologne, Germany
{henrik.niemann|martine.grice|doris.muecke}@uni-koeln.de

Abstract
In this production study, we investigate the effects of phrasal position and segmental makeup on the alignment of tonal events corresponding to rising nuclear pitch accents. Conditions included target words in initial, noninitial and final position, monosyllabic and disyllabic target words, and open and closed target syllables. Although cross linguistically the beginning of the rise (corresponding to the L tone) is relatively stable in its alignment with acoustic and articulatory landmarks, the endpoint, or target, of the rise (H) is notoriously variable. This highly variant peak alignment pattern on the acoustic surface corresponds to a stable coordination between the tone and the articulatory vowel gesture.

Keywords: tonal alignment, articulatory phonology, coordination, contextual effects

1. Introduction
In recent years there have been many studies investigating the temporal arrangement of F0 events with respect to the words with which they are coproduced. The first were concerned with landmarks in the acoustic signal, followed by studies exploring landmarks in the articulatory gestures used to produce the words (see Ladd et al. 2008 for an overview). Most of these studies have been concerned with the question of stability – or invariance – of F0 targets or events relative to these landmarks under different prosodic and segmental conditions (see Wichman et al. 2000 for an overview). The main focus of attention has been on pitch accents with a rise up to a high peak (L+H*), and consequently on F0 events corresponding to the low (L) and high (H) tones.

It has been found across many studies in a number of languages that the L is consistently aligned just before or after the acoustic onset of the accented syllable (Arvaniti, Ladd & Menn 1997, Ladd et al. 1999), and at the beginning of the articulatory gesture for the onset consonant (Mücke et al. 2012, Niemann, Mücke, Nam, Goldstein & Grice 2011). However, the alignment of the F0 peak for H shows a high degree of variation. Its timing appears to be sensitive to context, both the proximity of an upcoming phrase boundary and syllable structure. For example, it is aligned earlier in phrase-final position compared to other positions within the intonation phrase (Silverman & Pierrehumbert 1990, Caspers & van Heuven 1993) and is aligned later in closed syllables than in open ones (Mücke et al. 2009, Prieto & Torreira 2007). There is evidence that the effect of syllable structure on peak alignment disappears when investigating the peak alignment with articulatory gestures (Mücke, Grice & Hermes 2008). However, most of these studies were restricted to transvocalic gestures, that is they investigated consonantal gestures (opening of the lips during the vowel), and only indirectly captured the vocalic gesture (the tongue body movement directly related to vowel production). In this study, we systematically investigate the timing of rising nuclear pitch accents with landmarks in the acoustics and in relation to both consonantal and vocalic gestures.

2. Method
2.1. Speakers and recordings
We recorded four native speakers of German (S1-S4), all female and aged between 26 and 32 years. All speakers grew up north of the Benrather isogloss.

Kinematic data were obtained with a 3D Electromagnetic Articulograph (Carsten AG 500) at 200 Hz and smoothed with a 3-step floating mean. The acoustic data were recorded with the built-in, time-synchronized microphone and digitized at 16 kHz. Articulatory movements were tracked by sensors on the upper and lower lips as well on the tongue tip, blade and body. We recorded 288 tokens (4 spkrs x 3 target words x 3 phrasal positions x 8 repetitions).

2.2. Speech material and measurements
The speech material consisted of monosyllabic and trochaic disyllabic target words, the latter having either an open or a closed syllable (CV: /ma/, CV.CV /ma:mi/ and CVC.CV /manzi/). All target words were placed in carrier sentences designed to elicit nuclear rising pitch accents in (a) phrase-initial, (b) phrase-noninitial and (c) phrase-final position. Subjects were instructed to read this carrier sentence displayed on a computer screen in a comfortable and natural way. For each syllable structure, we maintained the sequence of stressed and unstressed syllables in the carrier sentence. The following carrier sentences with the target word /ma/ exemplify our corpus. Brackets indicate prosodic boundaries. Target words are bold and underlined.

(a) [Dann dachte sie:] [Ma mineralisierte das Wasser.]0
(Then she thought: Ma mineralised the water.)
(b) [Die Ma mineralisierte das Wasser.]0
(The Ma mineralised the water.)
(c) [Sie sah dann die Ma.]0
(Then she saw the Ma.)

Acoustic and articulatory data were labelled manually using the EMU speech database system (Harrington 2010). In the F0 trace, we identified local turning points for L (local minimum) and H (local maximum) in the vicinity of the rising nuclear LH pitch accent. Consonantal gestures were labelled via the Lip Aperture Index (Byrd 2000). We labelled local minima and maxima in the interlip distance trace corresponding to the opening and closing gesture of the bilabial consonants. In addition, we determined the peak velocities of the consonantal gestures taking zero-crossings in their respective acceleration traces. The target of the vocalic gesture was identified by means of a local minimum in the tongue body trace taking a zero-crossing in the velocity derivation.
The following variables (lags) were computed in the acoustic and articulatory domains:

(I) Measures relative to acoustic landmarks (fig. 1):
- **L-C1ons**: Beginning of the nuclear rise relative to acoustic onset of the accented syllable
- **H-EndSyll**: End of the rise relative to the acoustic offset of the accented syllable. This landmark refers either to the end of the accented vowel (as in the case of CV: and CV:.CV) or to the end of the coda consonant (in the case of CVC.CV)

![Figure 1: Acoustic landmarks for F0 alignment patterns](image)

(II) Measures relative to articulatory landmarks (fig. 2):
- **L-minC1**: Beginning of the rise relative to the articulatory maximum of the onset’s closing gesture (landmark 1)
- **L-relC1pvel**: Beginning of the rise relative to the peak velocity of the onset consonant’s release (landmark 2)
- **H-maxC1**: End of the rise relative to the maximal lip opening between C1 and C2 (transvocalic opening, landmark 3)
- **H-targV**: End of the rise relative to the articulatory target of the vocalic gesture (landmark 4)

![Figure 2: Articulatory landmarks for F0 alignment patterns](image)

3. Results

We conducted an overall rmANOVA (repeated measures ANOVA) based on cell means with POSITION and SEGMENTAL MAKEUP as independent variables and SPEAKER as random factor on the different acoustic and kinematic alignment measures. In case of significant effects, subsets were created and tested pairwise in Bonferroni-corrected post-hoc tests.

3.1. Alignment of L

Figure 3 displays the acoustic alignment of L relative to the beginning of the accented syllable.

![Figure 3: Alignment lags between L and the acoustic onset of C1](image)

L occurs on average of 59 ms after the beginning of the accented syllable. The rmANOVA revealed a small but significant effect of POSITION on the alignment of L relative to the onset of the accented syllable. [F(2, 6) = 7.92, p < 0.05 *]. However, subsequent Bonferroni-corrected post-hoc tests did not show any significant differences. The factor SEGMENTAL MAKEUP did not reach significance indicating a stable alignment across our target words.

A similar picture arises in the articulatory domain. Figure 4 displays the articulatory alignment patterns in terms of the L relative the maximum of the onset’s closing gesture (L-minC1). L occurs slightly after the consonant’s maximum closure. The rmANOVA showed only a small but significant effect of SEGMENTAL MAKEUP on the alignment of L relative to this articulatory landmark [F(2, 6) = 6.26, p < 0.05 *]. On average, L occurs 27 ms after the maximum constriction of the initial C and 40 ms before the following peak velocity (L-relC1pvel). In terms of articulation, this time window correlates with the release of the consonantal gesture.

![Figure 4: Alignment lags between L and the articulatory maximum of the onset’s closing gesture](image)

Table 1 provides means and standard deviations for the alignment of L relative to acoustic and articulatory landmarks.
Table 1: Means and standard deviations in parenthesis for the alignment lags of L in ms

<table>
<thead>
<tr>
<th>position</th>
<th>segmental makeup</th>
<th>L-C1</th>
<th>L-minC1</th>
<th>L-relC1pvel</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial</td>
<td>CV:</td>
<td>24 (23)</td>
<td>26 (21)</td>
<td>-54 (20)</td>
</tr>
<tr>
<td></td>
<td>CV:CV</td>
<td>29 (21)</td>
<td>39 (22)</td>
<td>-38 (23)</td>
</tr>
<tr>
<td></td>
<td>CVC.CV</td>
<td>32 (22)</td>
<td>43 (18)</td>
<td>-29 (23)</td>
</tr>
<tr>
<td>noninitial</td>
<td>CV:</td>
<td>79 (33)</td>
<td>24 (39)</td>
<td>-42 (38)</td>
</tr>
<tr>
<td></td>
<td>CV:CV</td>
<td>91 (40)</td>
<td>40 (41)</td>
<td>-23 (37)</td>
</tr>
<tr>
<td></td>
<td>CVC.CV</td>
<td>84 (28)</td>
<td>34 (30)</td>
<td>-22 (29)</td>
</tr>
<tr>
<td>final</td>
<td>CV:</td>
<td>50 (23)</td>
<td>-2 (29)</td>
<td>-70 (24)</td>
</tr>
<tr>
<td></td>
<td>CV:CV</td>
<td>68 (29)</td>
<td>18 (29)</td>
<td>-48 (30)</td>
</tr>
<tr>
<td></td>
<td>CVC.CV</td>
<td>70 (30)</td>
<td>15 (31)</td>
<td>-44 (34)</td>
</tr>
</tbody>
</table>

3.2. Alignment of H

In contrast to the alignment of L, H shows a considerably amount of variation when investigated relative to segmental landmarks. Figure 5 depicts the alignment of H relative to the end of the accented syllable which at the same time refers to the end of the vowel in CV: and CV:CV and to the end of the coda consonant in the case of CVC.CV.

The rmANOVA revealed a strong effect of both factors and an interaction between them [POSITION: F(4,12)=56.88, p<0.001*, SEGMENTAL MAKEUP: F(2,6)=29.66, p<0.001*, POSITION x SEGMENTAL MAKEUP: F(4,12)=41.53, p<0.001*]. Posthoc tests showed that both the target words and phrasal positions differ from each other significantly. In CV:, the peak was aligned on average 78 ms before the end of the accented syllable, in CV:CV, it was aligned 7 ms before it and in CVC.CV it was aligned 49 ms before the end of the accented syllable.

Figure 6 depicts the alignment of H relative to the maximum opening of the onset consonant gesture.

Again, the rmANOVA revealed an effect of both factors including an interaction [POSITION: F(2,6)=8.81, p<0.01*, SEGMENTAL MAKEUP: F(2,6)=29.34, p<0.001*, POSITION x SEGMENTAL MAKEUP: F(4,12)=22.87, p<0.001*]. Taking these results together, the f0 peak is neither aligned with the end of the accented syllable nor with the transvocalic opening of the consonant.

However, the picture changes when investigating the alignment of the f0 peak relative to the articulatory target of the vocalic gesture in the tongue body trajectory (see figure 7). Here, the effects of segmental makeup disappear. Note that we had to exclude the data from CV: in phrase-final condition as we could not reliably detect the vocalic target in the tongue body trace.

The rmANOVA revealed no effect of the segmental make-up but it did reveal an effect of the phrasal position [POSITION: F(2,6)=23.96, p<0.001]: On average, the peak is aligned 86 ms after the vocalic target in phrase-initial, 52 ms in phrase-noninitial and 44 ms in phrase-final condition. To sum up the, we found a stable alignment of H relative to the vocalic gestures. Table 2 summarises our results by providing means and standard deviations for the acoustic and articulatory alignment lags for H.
4. Discussion and conclusion

In this study, we investigated segmental and positional effects on the alignment of rising nuclear pitch accents in German. We confirmed that the beginning of the nuclear rise (L) exhibits a stable alignment, both to the acoustic onset of the accented syllable and the release of the consonantal gesture. As expected, the end of the rise, the H peak, was highly variable relative to acoustic and transvocalic consonantal landmarks. Both in the acoustic analysis and the analysis of the consonantal gestures, the H peak was consistently aligned later, the greater the amount of segmental material available. However, we found a stable alignment of the f0 peak relative to the target of the vocalic gesture. Thus, we propose that the target of the actual vocalic gesture (rather than the transvocalic gesture) serves as a trigger for the f0 peak.

Our results contribute to the growing body of evidence of a direct relation between laryngeal and supralaryngeal gestures. (Gao, 2009, Katsika, 2012). Utilizing a model of coupled oscillators within the framework of Articulatory Phonology (Browman & Goldstein, 1992; Nam & Saltzman, 2003) and conceiving pitch accents as tone gestures, it has already been shown that different alignment patterns between languages can be related to underlying coupling structures (Mücke, 2012, Niemann, 2011). Our results show that not only peak alignment differences found cross-linguistically but also contextually conditioned differences can be explained in terms of gestural coordination.

5. References

Byrd, D. Articulatory vowel lengthening and coordination at phrasal junctures. Phonetics, 57(1), pp. 3-16.

Perception of dialectal variation of speakers from East Thuringia
Christina Otto, Adrian P. Simpson
Friedrich-Schiller-University Jena
otto.christina@uni-jena.de, adrian.simpson@uni-jena.de

Abstract
The present study examines the potential relationship between vowel space dimensions and the perceptual distance of speakers with a regional accent and speakers of a more standard variety of German. Vowel space properties from one sentence of six reference speakers from the Kiel Corpus and six speakers from Zeitz were analyzed (F1, F2, Euclidean Distances, vowel space size). Speaker groups differ in acoustic parameters such as the Euclidean Distance from /a/ to /ɛ/, the retraction of the open vowel /a/, and [t] as the vocalic realization of /ɔ/. Subjects rated strength of dialectality of paired stimuli from the 12 speakers. Multidimensional scaling was used to model the perceptual space of listeners. These dialectal features are found to correlate primarily with the first two perceptual dimensions of listeners judging speakers from Zeitz and the reference speakers.

Keywords: East Thuringian, vowel space, perception, Multidimensional scaling

1. Introduction
Vertical variation between Standard German and dialectal varieties in East Central Germany is affected by processes of dialect leveling. Regional differences get lost and base dialects converge towards the majority variety with higher prestige: Standard German. This leads to more uniform varieties such as Upper Saxon Vernacular or Thuringian Vernacular (Bellmann 1983; Røneland 2010). The perception and production of this vertical variation has been subject to recent studies. Anders (2010) focuses on Preston’s framework of Perceptual Dialectology (Niedzielski and Preston 2000) to evaluate perceptions towards Upper Saxon, the ability to discriminate and assign speakers of different parts of Saxon and how the dialect region is represented and perceived by nonlinguists. Schaufuß (submitted) and Kehrein (2012) investigate the perception of vertical variation in Leipzig and Dresden (Upper Saxon) with the ‘Hörerurteil-Dialektialität’. In this approach, judgments about speakers are collected on a scale from deepest dialect to Standard German without accent (Herrgen and Schmidt 1985). This subjective dialectality is often correlated with a more objective measure of dialectality. The Phonetic Distance Measurement (Herrgen and Schmidt 1989) directly compares narrow transcriptions to the codified standard pronunciation of German and takes segmental differences to quantify the distance to Standard German. While this method seems to be a reliable and handy measurement of dialectality, as replications of results in several dialect areas have shown (see Purschke 2012 for an overview), it is entirely based on auditory judgments and nothing can be said about the actual acoustic distance from standard speakers since it reflects an impressionistic measure to a normative standard pronunciation. Indeed, spectral analyses are rare. Iivonen (1987) compares Standard German as spoken in East Central Germany (Halle/Saale) with the Standard German spoken in Vienna by comparing formant values. Siebenhäar (2013) explores data of 22 female speakers from Leipzig (Upper Saxon) measuring the acoustic vowel space sizes spanned by F1 and F2 of the long vowels /iː eː uː aː/ and the short vowels /ʊ ɛ ə a/. He finds shifting vowel space dimensions and smaller vowel spaces in a formal interview situation than in an intended dialect translation or an intended standard translation for both long and short vowels. Thus the present study also investigates the vertical variation in an East Central German variety to a more standard variety and its perception, but instead of impressionistic transcriptions, acoustic measures are used. ‘Standard variety’ is determined perceptually rather than with reference to the codified norm. Speakers from Zeitz (Saxony-Anhalt, East Thuringian dialect region) were recorded and compared to a reference speaker group (Kiel Corpus). A pretest revealed that a speaker group from the Kiel Corpus is perceived as significantly different from speakers from Zeitz. Speakers from the Kiel Corpus are perceived as closer to the standard end of a scale from Standard German without accent to deepest/strongest dialect, which is the main criterion to legitimately classify them as reference speakers (see Otto, submitted, for details on the pretest). The perceived similarity or dissimilarity and thus the distance of speakers from East Thuringia to reference speakers in terms of dialectality is explored. For this, a difference measure between speakers in terms of dialectality is needed. Perceived similarity/dissimilarity has been subject to several studies focusing on different aspects of voice (Kreiman et al. 1992; Nolan et al. 2011; Baumann and Belin 2010; Clopper et al. 2006). These studies use multidimensional scaling (MDS) to model the differences between stimuli in a low-dimensional space. MDS is a statistical technique to expose relationships among a set of items. Items are arranged in a low-dimensional space such that the distances between the objects reflect their relative dissimilarity. Dissimilarities are usually obtained by judgments of participants about the similarity or dissimilarity of an object pair. The MDS models the dissimilarity by fitting the distances in a low-dimensional space (usually Euclidean) (Borg et al. 2010). A multidimensional scaling of the listeners’ average judgments may show that speakers from Zeitz are perceived distant from each other and from most speakers from the Kiel Corpus while the reference speakers are perceived as being more similar to each other. Moreover, by correlating the resulting dimensions with acoustic measures, the perceptual space can be interpreted in terms of salient dialectal features which are important cues to listeners. The present study examines the potential relationship between vowel space dimensions and the perceptual distance of speakers with a regional accent and speakers of a more standard variety of German. The questions are: (1) Do speakers from an East central German variety differ in their vowel space size and dimensions from speakers of a more standard variety of German; (2) Do acoustic features of the vowel space have an influence on the perception of “dialect depth”? If so, which ones?
2. Material and Method

2.1. Speech material

Six male speakers from Zeitz (Saxony-Anhalt, East Thuringian dialect region), aged 26-69 were recorded. East Thuringian belongs to the East central German dialect group Thuringia-Upper Saxon and is spoken in the east of Thuringia and southwestern Saxony-Anhalt and is mostly a transition zone to Upper Saxon (Spangenberg 2013 [1989]). Characteristic features of the vowel system of this dialect area, which will be examined here more closely, have been described mostly by auditory studies to be centralization of the whole vowel system, retraction of the open vowel /a/ and the pharyngealized vocalic variants of /ɛ/ (Hünecke et al. 2012; Khan and Weise 2013; but Otto et al. 2014 for an articulatory and acoustic study on pharyngealization).

Recordings were made in a living room with a Zoom H4 on a tripod placed in front of the speaker. Subjects were asked to read the 100 Berlin sentences (Sotschek 1976). The six male reference speakers (aged 25-55) come from the Kiel Corpus of Read Speech (Simpson et al. 1997) and spoke the same sentence material under laboratory conditions. Only one of these subjects (k65) does not come from Schleswig-Holstein (Low Saxon dialect region) but from Bavaria.

Out of the list of sentences the utterance ‘Die Nacht haben Maiers gut geschlafen’ (‘Last night Maiers slept well’) was analyzed acoustically and served as stimulus for the perception experiment. This sentence was chosen because it shows only minimal consonantal dialect features, while at the same time it contains a range of vowels that together define a speaker’s vowel space and can be expected to be different for the two speaker groups.

2.2. Experimental design

A perception test was conducted to get a direct comparison measure in perceived dialectality between speakers that could also be correlated to relevant acoustic parameters. In contrast to the pretest and the distance measure used in previous studies (‘Hörerurteil’), speaker pairs are directly compared to each other in terms of strength of dialectality (see also Weirich and Simpson 2014 for a detailed description of this method).

Stimulus pairs were presented via the ‘ExperimentMFC’ module in Praat (Boersma and Weenink 2013). All possible speaker combinations (in both orders) were created out of the 12 speakers, while same-speaker combinations were excluded. The resulting 132 speaker pairs were presented to 35 listeners, all native speakers of German (10 male, 25 female, age range = 20-34). Listeners were mostly undergraduate students and were paid for their time. They were asked to decide which speaker speaks with a stronger regional accent and to rate the difference in perceived regional accent between two speakers on a seven-point scale from -3 to +3. Negative numbers were used if they perceived the first speaker as more dialectal and positive numbers indicated that they perceived the second speaker as more dialectal. The extreme points of the scale (-3/3) represented the perception of a very strong difference between the speakers while ‘0’ indicated no perceived difference in dialectality.

2.3. Acoustic analysis

Acoustic labeling and analysis was done using Praat (Boersma and Weenink 2013). Vowel spaces were explored by measuring F1 and F2 (in Hz) in several vowels contained in the sentence: ‘Die Nacht haben Maiers gut geschlafen’. The formants for /a/ and /u/ were measured at the midpoint of the vowels in ‘Nacht’ and ‘gut’. Because not all speakers produced the /ɪ/ in ‘die’, the formant values from the midpoint of the second part of the diphthong /ai/ in ‘Maiers’ was taken as a suitable high front reference point. Note though, that the aim of this study was not to describe the vowel characteristics of the East Thuringian dialect extensively, but to investigate the impact of the particular acoustic characteristics of one short sentence on the perceived dialectality. Since the retraction and often pharyngealization of the vocalic correlates of /ɛ/ is a salient feature for East Central German varieties (e.g. Khan and Weise 2013), formant values for [r] in ‘Maiers’ were also measured. Vowel space sizes were measured by calculating the polygon area (in kHz²) spanned by F1 and F2 of the three vowels /a/, /i/ and /u/. To investigate vowel space dimensions the Euclidean distances (ED) between these vowels were calculated. Unfortunately formant values from only five out of the six speakers from Zeitz could be measured since the recording quality of z06 was too poor. However, z06 was still used as a stimulus for the perception experiment.

3. Results

3.1. Acoustic Analysis

Figure 1 shows formant values for all relevant vowels for the six speakers from the Kiel Corpus in black (left graph) and the five speakers from Zeitz in grey (right graph). The lower plot of the figure shows the mean vowel spaces spanned by the vowels /a/ and /u/ for speakers from Zeitz (solid grey line) and from the Kiel Corpus (dashed black line). Differences between the speaker groups are apparent in Figure 1.

The resulting F2 formant values and EDs reveal that the average vowel space is more centralized for the speakers from Zeitz compared to the reference speakers of the Kiel Corpus: F2 of /a/ (Wilcoxon rank sum test, p < .01), and thus the ED from /i/ to /u/ are significantly lower for the speakers from Zeitz than for the reference speakers (Wilcoxon rank sum test, p < .05). Furthermore, F2 of /a/ (p < .05) and [r] (p < .01) is significantly lower for the Zeitz speakers, which indicates the
expected retraction of the open vowel and the vocalic correlate of /ər/. Both centralization and the retraction of the open vowels are dialectal features of East Central German varieties (Hünecke et al. 2012; Khan and Weise 2013). F1 differences did not reach significance. Regarding the vowel space size, no significant difference was found between the two groups (p = .08). Table 1 summarizes the results of the acoustic analysis. Individual variability on the acoustic measures in question can be observed in Figure 2.

Table 1: Mean values of the relevant acoustic measures and standard deviations with significant differences (Wilcoxon rank sum test, p < .05) in bold.

<table>
<thead>
<tr>
<th></th>
<th>Kiel mean</th>
<th>Kiel SD</th>
<th>Zeitz mean</th>
<th>Zeitz SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2 a* (Hz)</td>
<td>1427</td>
<td>96.96</td>
<td>1277</td>
<td>71.81</td>
</tr>
<tr>
<td>F2 r** (Hz)</td>
<td>1823</td>
<td>78.30</td>
<td>1621</td>
<td>89.07</td>
</tr>
<tr>
<td>F2 u (Hz)</td>
<td>915</td>
<td>173.40</td>
<td>1120</td>
<td>155.35</td>
</tr>
<tr>
<td>F2 v** (Hz)</td>
<td>1482</td>
<td>76.78</td>
<td>1106</td>
<td>149.53</td>
</tr>
<tr>
<td>ED u-a (Hz)</td>
<td>632</td>
<td>195.89</td>
<td>453</td>
<td>127.62</td>
</tr>
<tr>
<td>ED a-i (Hz)</td>
<td>441</td>
<td>146.27</td>
<td>415</td>
<td>113.54</td>
</tr>
<tr>
<td>ED i-u* (Hz)</td>
<td>924</td>
<td>180.87</td>
<td>535</td>
<td>186.31</td>
</tr>
<tr>
<td>polygon (kHz)</td>
<td>0.111</td>
<td>0.024</td>
<td>0.082</td>
<td>0.023</td>
</tr>
</tbody>
</table>

3.2. Multidimensional scaling

Judgments on each of the 132 pairings of speakers were subjected to MDS. Since the perception experiment not only asked for dissimilarity but also who speaks with a stronger regional accent, the original scale (-3 to +3) was recoded such that only differences between speakers are represented (0 to 3). An aggregated dissimilarity matrix over all listeners was calculated and subjected to a classical (metric) MDS with the cmdscale() function provided in R (R Core Team 2012). The MDS solution with three perceptual dimensions was chosen. The first two dimensions explain 89.8 % of the variance. Although the third dimension accounted for only further 3 %, significant correlation with one of the acoustic parameter was found.

Figure 3 shows the two-dimensional solution of the metric MDS. The scaling confirms the finding of the pretest (see Otto submitted): the two speaker groups are perceived as very distant from each other. Furthermore, speakers from the Kiel corpus show only little inter-speaker variability while speakers from Zeitz vary greatly on both dimensions. Only k65 is perceived as distant from the other Kiel speakers which is not surprising, since he is the speaker from Bavaria. Listeners seem to be sensitive to this even though he does not produce any salient Bavarian dialect features. But he does share some vowel space properties with the speakers from Zeitz, as can be seen in Figure 2. He has a smaller ED from /u/ to /a/ and his F2 of /a/ is not as high as for speaker k61 and k63.

Resulting coordinates of the dimensions were then correlated with acoustic measures to interpret the multidimensional scaling solution. Table 2 shows correlations between the acoustic variables and the MDS perceptual dimensions. The first dimension is significantly correlated negatively with F2 of /a/ (Pearson’s r = -0.64), positively with F2 of /u/ (Pearson’s r = 0.66) and consequently negatively with the ED of these vowel categories (Pearson’s r = -0.74). This means, that if the vowel space becomes more centralized (ED decreases because F2 of /a/ decreases and F2 of /u/ increases), the speaker is located higher on dimension 1 (e.g. speaker z01). Dimension 2, however, can be interpreted in terms of the retraction both of the open vowel /a/ and the vocalic correlate of /a/, since it correlates negatively with F2 of these vowel categories. Speakers who realize a retracted /a/ or [ə] are located higher on dimension 2 (e.g. speaker z02). Dimension 3 is weakly correlated with F1 of /a/ but only explains little of the variation.

Table 2: Correlations between the acoustic variables and MDS perceptual dimensions (Pearson’s r). Significant correlations (p < .05) are in bold type.

<table>
<thead>
<tr>
<th></th>
<th>Dim1</th>
<th>Dim2</th>
<th>Dim3</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 a</td>
<td>0.31</td>
<td>0.23</td>
<td>-0.10*</td>
</tr>
<tr>
<td>F2 a</td>
<td>-0.48</td>
<td>-0.66*</td>
<td>-0.05</td>
</tr>
<tr>
<td>F1 i</td>
<td>0.11</td>
<td>-0.42</td>
<td>0.31</td>
</tr>
<tr>
<td>F2 i</td>
<td>-0.64*</td>
<td>-0.25</td>
<td>0.18</td>
</tr>
<tr>
<td>F1 u</td>
<td>0.03</td>
<td>-0.29</td>
<td>0.09</td>
</tr>
<tr>
<td>F2 u</td>
<td>0.66*</td>
<td>0.30</td>
<td>-0.04</td>
</tr>
<tr>
<td>F1 u</td>
<td>-0.07</td>
<td>-0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>F2 u</td>
<td>-0.64</td>
<td>-0.61*</td>
<td>0.13</td>
</tr>
<tr>
<td>ED u-a</td>
<td>-0.60</td>
<td>-0.34</td>
<td>-0.11</td>
</tr>
<tr>
<td>ED a-i</td>
<td>-0.20</td>
<td>0.46</td>
<td>0.08</td>
</tr>
<tr>
<td>ED i-u</td>
<td>-0.74*</td>
<td>-0.34</td>
<td>0.12</td>
</tr>
<tr>
<td>polygon</td>
<td>-0.66</td>
<td>-0.01</td>
<td>0.10</td>
</tr>
</tbody>
</table>
4. Discussion and conclusion

Previous studies investigated (dis)similarity in voice quality (Nolan et al. 2011; Baumann and Belin 2010; Kreiman et al. 1992) or dissimilarity of speakers from different dialect regions (Clopper et al. 2006), subjects in our study rated stimuli with respect to dissimilarity of the strength of regional accent. We found significant differences between the two speaker groups for two features, which are reported to be typical dialectal features of East Central German varieties. Acoustic analysis revealed that speakers from Zeitz show a more centralized vowel space and realize a more retracted open vowel /a/ and a retracted or even pharyngealized vocalic correlate of /æ/ (see Otto et al. 2014 for an articulatory study of this feature in East Thuringia).

Multidimensional scaling revealed that these dialectal features are found to correlate primarily with the first two perceptual dimensions of listeners judging speakers from Zeitz and the reference speakers. Note though that with a metric MDS, which simply takes a dissimilarity matrix averaged over all listeners and repetitions, it is assumed, that all listeners share the same perceptual space. Kreiman et al. (1992) focus on different strategies used by expert listeners and naïve listeners on judging voice similarities of normal and pathological voices. They report differences in correlations of acoustic measures with the resulting dimensions for listener groups as well as for individual listeners. Since the listener group in our study is relatively heterogeneous, especially regarding their origin, an analysis of Individual Differences (INDSCAL) might be more appropriate, since it takes dissimilarity matrices of all listeners independently and scales the perceptual space accordingly. Origin of listeners might be a factor worth taking into account, because it might have an influence on the dialectality judgments. However, a simple MDS for the whole listener group was able to offer some insights into the perception of speakers from Zeitz in relation to reference speakers from the Kiel Corpus.

We are aware of the fact that material used in this study is limited, since listeners only rated one sentence of a sample of only 12 speakers. In addition, it was not controlled for age and the age range within the speaker groups differs considerably. But since this study represents a pilot to a larger sociophonetic study, acquisition of data from a larger set of speakers controlled for age and other variables is in progress.

5. References


Spectral and lingual correlates of /r/ in East Thuringia
Christina Otto, Melanie Weirich, Adrian P. Simpson
Friedrich-Schiller-University Jena
otto.christina@uni-jena.de, melanie.weirich@uni-jena.de, adrian.simpson@uni-jena.de

Abstract
This study investigates acoustic and articulatory correlates of coda /r/ in East Thuringian using Ultrasound imaging technique (UTI). A selection of /r/ and /r/-less word pairs containing the vowels /ɛ/ and /a/ (e.g. 'Metz-März,' 'Rents-Rents') of seven male speakers were recorded. Systematic differences in lingual configurations and spectral patterns between /r/ and /r/-less vowels consistent with pharyngeal constriction were found. Acoustic analysis of the word pair 'Pelle-Perle' (/pɛɬɛr/ /pɛɬɛr/) illustrates the temporal extent of this pharyngealization extending over at least two syllables.

Keywords: ultrasound, East Thuringian, pharyngealization, /r/

1. Introduction
Varieties of German exhibit a wide range of consonantal and vocalic correlates of /r/. Furthermore, there is generally a marked difference between the onset and coda correlates of /r/, such that the coda correlates tend toward more vocalic variants (Simpson 1998). The vocalic correlates can give rise to a range of diphthongs (e.g. [iɭ uɭ] ʔer (her), ʔe (her), ʔer (clock)), but in combination with more open vowels monophongal qualities can arise from the temporal overlap of the correlates of the vowel and the correlates of the coda /r/ (e.g. /ʃpʊrt/ /ʃpʊrt/ Sport (sport)). In east central German varieties, such as Saxon (Khan and Weise 2013; Spangenberg 2013 [1989]), but also in a south western variety such as Swabian (Hiller 1995), correlates of coda /r/ are characterized by pharyngealization of the preceding vowel, but also of adjacent consonantal material. Pharyngealization is produced either by contracting the constrictor muscles of the pharynx or by retracting the tongue root to produce a narrowing of the pharyngeal passage which results in the impression of a darker quality of the segment (Laver 1994). This articulatory configuration has typical acoustic properties. The enlargement of the oral cavity lowers the strong and weak ends of an auditory rhotic continuum. Vowel retraction and pharyngealization can be observed in derhoticized variants.

The present study represents the first comprehensive investigation of the acoustic and articulatory correlates of coda /r/ in East Thuringian, contributing to our knowledge of a still poorly understood phenomenon in German dialectology and is part of a larger sociophonetic study (e.g. Otto submitted).

2. Material and method
2.1. Data collection
Seven male speakers aged between 26 and 28 from the south of Saxony-Anhalt were recorded. All speakers grew up in the East Thuringian dialect region, either in Naumburg (all but one speaker) or a town nearby (Zeitz). East Thuringian belongs to the east central German dialect group Thuringian-Upper Saxon and is spoken in the east of Thuringia and southwestern Saxon-Anhalt and is mostly a transition zone to Upper Saxon (Spangenberg 2013 [1989]). All subjects live and work in their hometown or in another east central German city and visit their hometowns more than once a month. They all belong to the same circle of friends, along with the first author, which ensured an informal situation during recording.

To evaluate the coarticulatory effect of postvocalic /l/, 15 minimal pairs of words differing in the presence or absence of coda /l/ in a range of short vowel contexts e.g. Metz (mason) vs. März (March), were embedded in meaningful sentences. The sentence frame for each pair of words was kept the same but varied between different pairs of words in an attempt to enhance naturalness and not to make the target word the main focus of interest, e.g. Sie ist nach Born/Born gefahren (She went to Born/Born), Sie hat den Metz/März geliebt (She loved March/the mason). Each sentence block was repeated five times.

Data presented here is taken from four word pairs containing the short vowels /ɛ/ in Metz-März, /a/ in Pack-Park (packpark), /ɛ/ in Bonn-Born (town names) and /a/ in Kuss-Kurs (kiss-course). The word pair Pelle-Perle (skin-pearl) was analyzed to investigate the temporal extent of the correlate of /r/ on both the preceding vowel /l/ and the adjacent lateral. A transducer type 35C20EA attached to a Mindray DP2200 Ultrasound system together with Articulate Assistant Advanced software (AAA, Wrench 2007) were used to capture and process synchronized audio data and ultrasound images of the mid-sagittal contour of the tongue (transducer frequency was 3.5 MHz with a 12.9 cm depth; interlaced video frame rate of 30 frames per second). The ultrasound probe was kept in a relatively stable position under the speaker’s chin using a purpose-built headset (for details see Scobie et al. 2008). Acoustic signal was recorded using a condenser microphone (AKG C1000S). Stimuli were presented on a screen in a randomized order using the facility provided in AAA. At the beginning of each recording the palatal contour was recorded whilst the subject swallowed mineral water.
2.2. Data analysis

2.2.1. Articulatory analysis: Frame selection and spline fitting

Three frames were selected from each VC sequence to estimate the tongue contours at two points during the vowel and one point approximately 10 ms after the start of the syllable final consonant. A fan grid consisting of 42 radial axes was superimposed on each frame. The location where the tongue surface contour intersects each of these radial axes was used to fit a spline to the visible surface of the tongue by a fitting algorithm implemented in the AAA software and was corrected manually. Fitted splines of the second time points of the vowels were exported to a workspace to calculate an average tongue contour for each speaker and each vowel including standard deviation to indicate the range of the data. Within this study we concentrate on a qualitative analysis and a visual description and comparison of tongue contours with and without pharyngealization.

2.2.2. Acoustic analysis: Formant measurements

Values of the first two formants were estimated at the approximate timepoint of the articulatory measurements during the vowel using PRAAT (Boersma and Weenink 2013). Additionally, to investigate the temporal extent of the acoustic correlates of /r/ on consonantal material, for the Pelle-Perle pair, formants were measured at five timepoints throughout the voiced /h(r)l/ sequence. For statistical analyses linear mixed models (LMMs) as implemented in the lme4 package (Bates et al. 2011) were run in R (version 2.14.1, R Development Core Team 2008) with the different formant values as dependent variable, the fixed factors vowel and /r/-condition (with and without /r/), and the
random factors speaker and repetition. Likelihood ratio tests were used for model comparisons to find the model with the best fit to the data. Comparisons were done in a stepwise fashion by adding more factors (or interactions) in each step to the model and comparing this model with the reduced model without the particular factor (or interaction) in question. We then decided on keeping a factor/interaction depending on the result of the likelihood ratio test (p-value < .05). The summaries of the final models are provided in the tables.

3. Results

3.1. Articulatory analysis

Figure 1 shows palate (black) and tongue contours (red and green) of the seven male speakers taken at the second measurement point (approximately two thirds of the way through the vowel). Each column shows a different vowel pair, each row a different speaker. Green contours represent /r/-vowels, red contours their /r/-less congener. The thick lines show the mean of five repetitions, the thinner lines represent the standard deviations. As the figure shows, overall, the green splines describe a mid-sagittal tongue contour which is flatter, more retracted and higher towards the back. However, differences vary between vowel pairs and for the /a/-pair the tongue contours do not differ except for one speaker (RG). This is not surprising, since the low vowel /a/ is already realized with a retracted tongue root which results in the typically muffled /a/ of east central German variants. Only RG retracts his tongue even further in *Park*, which is in line with the extreme differences in the other vowel pairs: his tongue is more retracted and flatter than the tongues of the other speakers and he shows only small intra-speaker variability mirrored in the very small standard deviations, reflected by the thinner lines in the figure.

3.2. Acoustic Analysis

Differences in lingual configuration are reflected in acoustic differences. Figure 2 contains a F2xF1 plot of all vowel pairs showing that the /r/-vowels (green) are positioned further back and lower in the acoustic vowel space. However, here too, differences are apparent between the vowel pairs. Analogous to the articulatory results, differences are smallest for the /a/-/ar/ pair and largest for /i/-/I/.

Figure 3 summarises the differences between the formant values separated by vowel and /r/-condition (F1 left, F2 right). While for F2 lower values are found for the /r/-condition for all vowels, the influence of /r/ on F1 varies between vowel pairs.

Likelihood ratio tests revealed that the LMM with the best fit to the data includes the interaction term /r/-condition*vowel for both formants and Table 1 shows the summary statistics for the fixed effects. The reference levels for both models is the vowel /a/ in the /r/-less condition (i.e. *Pack*). While for F2 all vowels differ significantly between the /r/-conditions, for F1, no difference is apparent for /a/-/ar/ and the effect is largest for /i/-/I/.

We have seen in the articulatory analysis that the /a/-/ar/ contrast does not seem to be realized by retracting the tongue root, which is also supported by a lack of difference in F1. A clear auditory difference in liprounding between *Pack* and *Park* could be observed in all speakers, which is perhaps reflected in the significantly different F2 values.

Table 1: Summary statistics for the fixed factors vowel and /r/-condition, dependent variable is F1 (upper part) or F2 (lower part), (observation: 277, repetitions: 6, speakers: 7)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>575.2</td>
<td>12.5</td>
<td>46.0</td>
</tr>
<tr>
<td>cond.[no /r/ vs. /r/]</td>
<td>-10.0</td>
<td>11.6</td>
<td>-0.9</td>
</tr>
<tr>
<td>vowel[ /ar/ vs. /i/]</td>
<td>-139.2</td>
<td>11.2</td>
<td>-12.1</td>
</tr>
<tr>
<td>vowel[ /ar/ vs. /e/]</td>
<td>-129.4</td>
<td>11.6</td>
<td>-11.1</td>
</tr>
<tr>
<td>vowel[ /ar/ vs. /u/]</td>
<td>-239.7</td>
<td>11.6</td>
<td>-20.7</td>
</tr>
<tr>
<td>cond.[/r/]:vowel[ /i/]</td>
<td>191.5</td>
<td>16.2</td>
<td>11.9</td>
</tr>
<tr>
<td>cond.[/r/]:vowel[ /e/]</td>
<td>65.5</td>
<td>16.5</td>
<td>4.0</td>
</tr>
<tr>
<td>cond.[/r/]:vowel[ /u/]</td>
<td>99.1</td>
<td>16.4</td>
<td>6.1</td>
</tr>
</tbody>
</table>

3.3. Temporal extent

The temporal extent of the acoustic correlates of /r/ can be seen most clearly in tokens of the *Pelle-Perle* (skin-pearl) pair displayed in Figure 4. F1 and F2 were measured at four timepoints throughout the /r(r)/als-sequence: two during /r(r)/, one in the middle of /l/, one in the middle of /a/. Each point in Figure 4 represents the mean of 35 tokens (five repetitions x seven speakers). The figure clearly shows the differences, especially in the location of F2, one of the main acoustic consequences of the different lingual configurations, not only for the vowel(/r/)-sequence but also for the lateral and even for the following schwa. In *Perle* the auditorily dark lateral is reflected in a much lower F2 value (ca. 1200 Hz) than in *Pelle* (ca. 1600 Hz). Likelihood ratio tests revealed that the LMM with the best fit to the data includes the interaction /r/-condition*timepoint for both formants and Table 2 shows the summary statistics for the fixed effects. The reference levels for both models is the first timepoint of *Pelle* (see /at/, early in Figure 4). For F1
This study has provided a first important insight into the articulatory mechanisms and acoustic correlates of a pervasive and salient feature found in a number of varieties of German. At the same time, this study also complements our knowledge of similar phonetic phenomena in other languages such as Arabic.

5. References


Bates, D., Maechler, M., Bolker, B. and Walker, S. (2011), lme4: Linear mixed-effects models using S4 classes, R package version 2.15.2.


Coordination of Acoustic Production and Postural Sway in ensemble singing

Caroline Palmer, Erik Koopmans and Frances Spidle

1Department of Psychology, McGill University

caroline.palmer@mcgill.ca

Abstract

Both auditory feedback and postural control are important for maintaining temporal coordination in singing. We examined the influence of visual and auditory cues on vocal duetts’ acoustic productions and posture. Experienced vocalists performed simple melodies in Solo, Unison (same musical part sung simultaneously by vocalists), and Round conditions (same musical part sung at a temporal delay) while standing on force plates. Visual feedback was manipulated by facing the singers Inward (with full view of each other) and Outward (with no view), while normal acoustic feedback was maintained. Measures of ground reaction forces and center of pressure indicated increased sway variability in Rounds compared with Unison performances, and when vocalists had full visual cues. Correlations of vertical ground reaction force measures indicated more agreement between singers during Unison singing than Round singing. Although visual cues about one’s partner had some influence on the variability of vocalists’ posture during duet performance, only the type of performance (Unison/Round) influenced the degree of correspondence in posture between vocal duet members, consistent with the findings from their acoustic productions.

Keywords: speech synthesis, entrainment, rhythm, interpersonal coordination, ensemble music performances

1. Introduction

Postural control is important for singing, and is constrained by the demands of breath support [1, 2]. The dynamics of maintaining balance while standing still, as vocalists must often do for long periods during performance, are complex and are often affected by the kind of visual feedback available [3]. In addition, singers often sway during performance [1], and they can use body gestures to signal to their ensemble partners [4]. The coupling of perception and action between ensemble members may be modulated by the acoustic relationships between the musical parts produced by the performers. We examine postural sway and the role that visual and auditory cues play in vocal duets.

Postural sway during quiet standing is often investigated in terms of reaction forces between the feet and the ground, and by spontaneous shifts in people’s center of pressure [5]. When maintaining a standing posture in the context of steady forces due to gravity, equilibrium of the body joints is unstable and small fluctuations in balance reflect changes in muscle activity that cause movement [3]. This sway has been interpreted as the output of a stabilization control process for the human body. The center of pressure, defined as the point location of the weighted average of the sum of vertical ground reaction forces (i.e., forces created by the vocalist’s feet on the floor) in the anterior/posterior (forward/backward) direction and the medial/lateral (side-to-side) direction, can change during standing; adjustments in center of pressure can result from biomechanical demands as well as attentional factors [3].

Several studies have shown that changes in postural sway are coupled to changes in sensory aspects of the environment. For example, changes in center of pressure were coupled to predictable sinusoidal changes in a tactile stimulus [6]. Speaking has been shown to cause an increase in body sway, relative to mental (nonverbal) tasks [7, 8]; the authors hypothesized that increased sway may be attributable to adjustments associated with breathing or with increased motor planning. We hypothesize further that increased sway in duet music performance may be attributable to communicative influences between partners as they attempt to synchronize their vocalizations. We compared the standing body sway patterns of performing vocalists, as measured by the center of pressure (COP): integrated pressure that a person exerts on a supporting surface when standing. Several studies have documented greater postural variability in the medial/lateral direction than in the anterior/posterior direction [5, 9], and therefore we hypothesize that duet vocalists may exhibit greater coupling across medial/lateral measures of sway.

We examined the postural sway of experienced vocalists who stood on force plates as they performed simple melodies in Solo (individuals performing the same melodic content alone), Unison (duet performers performing the same melodic content simultaneously), and Round conditions (duet partners performing the same melodic content with one duettist’s part temporally offset from the other). Influence of visual feedback about the partner’s movement was manipulated by facing the singers Inward (full view of each other) and Outward (no view of each other). This design allowed the comparison of the effects of visual and auditory cues on singers’ postural sway as they performed.

2. Vocal Duet Measurements

2.1 Experimental Design

Thirty amateur vocalists, recruited from the Montreal community, were matched randomly into 15 pairs, based on their vocal range. All had more than 5 years of choral experience and reported having normal hearing.

The vocalists were asked to sing the folk tune Frere Jacques from memory to the syllable “da”, to make the task easier. All participants were familiar with the song prior to the study, and sang it first in a Solo condition at a comfortable vocal range and tempo. They then sang together in Unison and Round conditions at a specified tempo of 120 quarter-note beats per minute (500 ms IOI) cued by an internal metronome. Singers stood on force plates, placed at 90 degree angles to each other, and were recorded with directional head-mounted mikes. A metronome was sounded for 4 beats at the beginning of the trials, and the acoustic and force plate measures were synchronized via SMPTE.

The experiment used a within-subjects design in which each participant sang the same melody in Solo, Unison and Round conditions. In the duet performance conditions, vocalists were assigned the role of Leader (who was responsible for
maintaining the pace of the ensemble at the metronome setting) and Follower. In addition, the vocalists were instructed to face each other (Inward) on half of the trials, and to face at 90 degree angle to each other (Outward) on the remaining trials within each condition. Halfway through the experiment, the vocalists exchanged roles of Leader/Follower and positions on the force plates.

2.2 Postural sway

We compared the standard deviations for all three posture measures from force plate measurements: anterior/posterior COP, medial/lateral COP, and vertical GRF (Z), by Performance (Unison/ Round) and Facing (Inward/Outward) conditions. Analyses of variance indicated significantly larger variability during musical Rounds than during Unison singing in anterior/posterior COP and in vertical force measures, consistent with the notion that vocalists’ musical goals influence their posture. Vocalists tended to sway less in Unison than in Round performances.

Visual cues also affected vocalists’ postural sway; vertical force and COP variability measures increased as vocalists faced Inward (full view of partner) compared with Outward (no view of partner). Facing condition interacted with Performance type in anterior/posterior COP measures; full visual cues from the vocalists’ partner affected the variability of posture more in Round than in Unison performances.

We compared vocalists’ postural sway in Solo performances with those in duet (Unison and Round) performances. Although vocalists did not sing together in the Solo performances, this condition provides a control for whether postural sway increases in variability due to demands of articulation, as documented in previous research [6], or alternatively, due to communicating or coordinating with a partner. Comparisons between the Solo condition and the two duet conditions revealed no significant differences in variability of postural sway between Performance types in anterior/posterior or medial/lateral COP. Vertical force measures indicated significant differences across the three performance types. Post-hoc comparisons demonstrated that Solo and Unison conditions did not differ; only Round performances generated larger variability in postural sway).

Thus, the vocalists’ posture during Solo performances suggests that articulatory demands alone of singing could account for the level of postural variability found in the Unison condition but not the increased variance in the Round condition.

2.3 Correspondence measures across Vocalists

To compare whether vocal duettists displayed the same pattern of postural change, we correlated the posture measures across Leader/Follower duet partners within each performance. Vertical GRF values were significantly correlated across Leader and Follower in both Unison and Round conditions. Analyses of variance on the correlation values for vertical GRF by Performance and Facing conditions indicated that Unison performances were significantly more correlated than Round performances suggesting that performers influenced each other’s vertical posture more when they sang the same musical part simultaneously (Unison) than when they sang the same melody at a temporal lag (Round). Facing in or out did not affect the level of correspondence between vocalists’ posture measures.

Periodic (cyclical) changes in Leader and Follower’s posture during singing suggested that, similar to previous research, postural sway reflected different frequency components. To test whether the two vocalists’ postural sway exhibited similar frequency components during duet performance, we computed a Fast Fourier Transform (FFT) on the frequency components present in each trial for vertical GRF measures; the FFT decomposes a time-series sequence into its component frequencies. Each vocalist’s power spectrum (in units of Nms²) was then normalized within each trial to represent the proportion of the total spectral energy; Each performance contained multiple frequency components; the overlap between components signals the degree of correspondence in postural change.

To compare the degree of correspondence across conditions, we correlated the normalized power spectra across Leader/Follower duet pairs within each trial. Significant correlations between the vocalists were found for vertical GRF values for both Unison and Round duet conditions. Analyses of variance on the power spectra for vertical GRF measures indicated equivalent correspondence in Unison and Round conditions. The effect of Facing condition on the power spectra approached significance; correlations in power spectra for vertical force were higher when vocalists faced Inward than Outward, suggesting that the frequency components of vocalists’ vertical forces changed in synchrony with their partner more in the presence of visual cues.

We compared the degree of correspondence in the power spectra for the duet performances with those in the Solo performances, as a control condition for standing body sway when singers perform separately. Analyses of variance on the correlation values for Solo performances and duet performances indicated significant differences across Performance conditions, with lower (non-significant) correlations for Solo than for Round and Unison conditions. Consistent with the correlations among vertical GRF values, these analyses suggest that vocalists’ vertical posture corresponded to vertical posture changes by their duet partner when they sang together, but not when they sang individually.

3. Discussion and conclusion

Vocalists’ changes in posture during duet performance were influenced more by the musical relationships between the parts being sung, than by the visual cues from their partners. We contrasted Unison and Round singing, in which interference from related but different parts may influence posture of vocalists who need to maintain synchrony with the other performer. We also contrasted the visual cues present by having the vocalists face inward (full visual cues of their partner) and outward (no visual cues of their partner). Measures of ground reaction forces and center of pressure indicated increased variability in postural changes when each vocalist’s part deviated from the other in the Round condition, and when the vocalists had full visual cues about their partner in Inward-facing conditions.

Postural influences of the musical parts sung were also evidenced within each performance, as one vocalist’s body sway coincided with changes in the other vocalist’s sway. Correlates of vertical ground reaction force measures indicated significant correspondence across duet performances of both Unison and Round singing. Furthermore, comparisons of the relative weighting of frequency components in the
postural changes, as measured by a power spectrum analysis of the vertical ground forces in each performance, indicated greater correspondence between Leaders and Followers' Unison and Round performances than was found between their Solo performances. Although the presence of visual cues about vocalists’ partners increased the correlations of vertical ground forces, the effects on postural change were weak in comparison to those of Performance type.

Although several well-established findings document the effects of visual cues on postural sway during standing [3], it is perhaps not surprising for music performance to be influenced more by acoustic aspects of the duets than by visual cues. Previous work [7, 8] suggested that the act of speaking (articulating) increases body sway, due to increases in motor planning or in jaw and head movements. We were able to control such effects of articulating by comparing posture measures during Solo performances by each vocalist with their posture in duet performances. Thus, the increased correspondence between vocalists’ power spectra for force measures in duet performances relative to Solo performances could not be attributed solely to articulation demands, which remained consistent across Solo and duet performances.

In sum, balance and posture control are important for successful singing. Postural sway appears to be influenced more by the musical part being produced (Union or Round) than by the presence or absence of visual cues from the other performer. These findings were based on performances of a rhythmically simple melody; further work may examine the relationship of sway frequencies to the rhythms present in the music, and whether postural sway is constrained by the style of musical performance.

4. Acknowledgements

The author would like to thank the ISSP 2014 organizing and scientific committees as well as James O’Callaghan and Peter Schubert for their assistance.

5. References


Pre-Speech Tongue Movements Recorded with Ultrasound

Pertti Palo1, Sonja Schaeffler1, James M. Scobbie1

1CASL Research Centre, Queen Margaret University, Edinburgh, Scotland, UK

Abstract

The tongue moves silently in preparation for speech. We analyse Ultrasound Tongue Imaging (UTI) data of these pre-speech to speech phases from five speakers, whose native languages (L1) are English (n = 3), German, and Finnish. Single words in the subjects’ respective L1 were elicited by a standard picture naming task. Our focus is to automate the detection of speech preparation through the analysis of raw UTI probe-return data, here captured at 201 fps. We analyse these movements with a pixel difference method, which yields an estimate of the rate of change on a frame by frame basis. We describe typical time dependent pixel difference contours and report grand average contours for each of the speakers.

Keywords: Pre-speech articulation, ultrasound tongue imaging, pixel difference, automatic data analysis

1. Introduction

Study of speech preparation has come under increasing interest in recent years. Studies based on the acoustic modality analysing e.g. questions of phonological preparation (Rastle et al. 2000), and conversation turn taking (Heldner and Edlund 2010) have been complemented with studies based on articulatory data such as the recent analysis of organisation of speech preparation processes by Tilsen and Goldstein 2012.

The most readily accessible modality of speech production – and therefore of speech preparation – is sound. However, recording speech sounds and sounds produced during speech preparation (e.g. J. M. Scobie, S. Schaeffer, and Mennen 2011) is an indirect way of observing the physical processes that produce them. Furthermore, the speech preparation movements are mostly silent and thus best observed by directly recording the articulation itself.

By varying the complexity of the tasks performed by the participants of a reaction time experiment light can be shed on the organisation of cognitive speech preparation processes (see e.g. Rastle et al. 2000; Tilsen and Goldstein 2012). For this purpose researchers are usually concerned with the onset time of acoustic speech, onset of phonation, or the onset of phonetically meaningful articulatory movements.

Ultrasound Tongue Imaging (UTI) has a long history in speech studies in general (see Minifie, Kelsey, and Zagzebski 1971, for one of the first studies). Its applications range from speech therapy (see e.g. Bernhardt et al. 2005) to silent speech interfaces (Hueber et al. 2010). In comparison with the most viable alternative, i.e. Magnetic Resonance Imaging (MRI), UTI is relatively cheaper and simpler to use and has a far better temporal resolution. On the down side UTI data is fairly noisy, often contains artefacts and the imaged area is limited to the tongue and its surface.

McMillan and Corley 2010; Drake, S. Schaeffler, and Corley 2013a; Drake, S. Schaeffler, and Corley 2013b have automated the processing of UTI data by considering the difference or amount of change between consecutive ultrasound frames. The difference was defined as the Euclidean distance of two ultrasound frames or images as they are taken to be N dimensional vectors, with each pixel presenting a dimension. A similar approach to analyse gross laryngeal motion with vector analysis has also been pursued by Moisik 2010 and applied to tongue movement by Bird et al. 2010.

2. Materials and methods

The experiment used the Snodgrass-Vanderwart picture naming task (Snodgrass and Vanderwart 1980) and it had five subjects: Four females (P1, P2, P3, and G1) and one male (S1). All of the speakers did the experiment in their native tongue – three in English (participants P1, P2, and P3), one in German (participant G1) and one in Finnish (participant S1).

The experiment was run with synchronised ultrasound, lip imaging and sound recording controlled with Articulate Assistant Advanced (AAA) software (Articulate Assistant Advanced User Guide: Version 2.14 2012) which will also be used for the analysis. The participants were fitted with a purpose-built headset to ensure stabilisation of the ultrasound probe (Ultrasound Stabilisation Headset Users Manual: Revision 1.4 2008). Attached to the helmet was a small Audio Technica AT803b microphone for high-quality acoustic recordings. Ultrasound recordings were obtained at a frame rate of 201 frames per second with the high speed SonixRP system at Queen Margaret University (Wrench and J. M. Scobbie 2011).

The sound recording was initiated and a fixation point was shown on a computer screen 1.5 seconds before the subjects were shown the stimulus on the screen. The ultrasound recording was automatically initiated at about 0.5 s or 1.0 s after the sound recording began1 thus capturing any movements related to speech preparation as well as making it possible to spot cases where the subject was moving already before the onset of the stimulus. To guarantee an informative length of common time between averaged UTI recordings any tokens with an UTI recording length of less than 2.2 s were discarded. In addition, only two of the participants completed the whole experiment of naming 260 tokens as wearing the UTI helmet does get strenuous in longer experiments. These factors resulted in sample sizes of 148, 264, 219, 260, and 216 (mean = 221) respectively for speakers P1, P2, P3, G1, and S1.

1This delay results from the experimental software and the experimental apparatus and can not at the present be controlled, but can be and has been carefully measured.
2.1. Pixel differences

We analyse the samples by automatically calculating the amount of change in the UTI data as a function of time. We use pixel difference as the change metric and evaluate it over each recording comparing raw ultrasound frames in sequence.

UTI commonly uses ultrasound probes which produce a fan shaped image of the tongue. Ordinary or interpolated ultrasound data refers to the form usually displayed by ultrasound imaging systems as seen in Fig. 1 a). The fan image of the ordinary ultrasound data is produced by linear interpolation between the actual raw data points produced by the ultrasound system as it images the tissues. The raw data points are distributed along radial scan lines with the number of scan lines and the number of data points imaged along each scan line depending on the setup of the ultrasound system.

Fig. 1 b) shows a raw ultrasound frame. In our experiments each frame has 38 scan lines (x dimension of the image) and 412 pixels along each scan line (y dimension of the image). Each pixel ranges in value from 0 (black) to 255 (white), with most pixels staying at values of below 200.

To calculate the pixel difference between two UTI frames we interpret each raw frame as a \( N = n_x \times n_y \) dimensional vector. The pixel difference \( d1 \) between consecutive UTI frames is then defined as the Euclidean distance between the two frames \( im_k \) and \( im_{k+1} \) with indeces \( i \) and \( j \) iterating over the pixels in \( x \) and \( y \) direction:

\[
d1(k) = \sqrt{\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (im_k(i, j) - im_{k+1}(i, j))^2}
\]

for \( k = \{1, 2, \ldots n_{frames} - 1 \} \). The difference can be calculated as readily for images further removed and are defined as

\[
dL(k) = \sqrt{\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (im_k(i, j) - im_{k+L}(i, j))^2}
\]

for \( k = \{1, 2, \ldots n_{frames} - L \} \). In this paper we will use the differences \( d1 \) and \( d3 \). The time stamp \( t_{dL(k)} \) corresponding to a single difference value \( dL(k) \) is defined as the average of the time stamps of the corresponding UTI frames:

\[
t_{dL(k)} = \frac{1}{2} (t_{im_k} + t_{im_{k+L}}),
\]

where the time stamp of image \( k \) is the time its acquisition ends.

3. Results

The automated production of pixel difference contours \( d1 \) and \( d3 \) (Step = 1 and Step 3, respectively, in the Figures) yielded
three distinct types of typical productions which are illustrated by examples in Figures 2 - 4 along with the corresponding sound wave forms. Fig. 2 shows a typical clear production. The participant has held still before starting linguistic articulation. From comparison with the corresponding waveform at the bottom of the figure, it can be seen that the articulation starts a good while before acoustic onset. Fig. 3 shows an example of a hesitation. The participant moves her tongue as if she was going to speak, but returns to rest and finally speaks later. Fig. 4 shows a chaotic example. The participant is moving already at the time the recording starts and continues to move throughout the whole recording.

Figures 2 and 3 show a clear preparatory stage before the linguistic movements (starting at about 1.8 s and 2.9 s, respectively). This stage is characterised by the two contours lying very close together or even practically on top of each other. In most tokens the pre-speech movements are fairly small in comparison to the linguistic movement as seen in Figures 2 and 3. In contrast, in Fig. 4 the movements are practically of the same magnitude at all stages of the production.

The fact that the contours never reach a difference value of 0, or indeed even come close to it, is caused by the fairly high pixel noise or random flicker present in the images. This results in a more or less constant noise floor (about 750 pixel difference units for the current setting) which at times hides some of the movement. Using difference with a longer step makes movement in general more prominent as seen especially in Fig 3.

Individual differences between the speakers can be seen in the grand averages shown in Figures 5 - 9. These figures show an average of the pixel difference contours over analysed tokens of each speaker. Only tokens with a UTI recording longer than 2.2 s have been taken into analysis and the averages have been calculated over the longest common time interval for the tokens of each individual speaker.

Notable features include differences in average reaction time – S1 is clearly slower to respond than P1 or P3, and differences in the separation of $d_1$ and $d_3$ contours. P2 has a clearly greatest separation of the two in the pre-speech phases with G1 and S1 having the smallest. This is probably the result of G1 and S1 holding fairly with few movements of the tongue before linguistic articulation.

4. Discussion and conclusion

Different types of productions are present in the data. Some are easier to analyse than others. Automatic analysis of such tokens as shown in Figures 2 and 3 should be fairly straightforward to produce automatic tongue movement onset times. The situation is complicated by the third type of token shown in Fig. 4. The easiest solution would be to either automatically or manually exclude such tokens from analysis. A more challenging solution would be to use acoustic analysis and more detailed articulatory analysis – e.g. using an analysis-by-synthesis approach (see...
Figure 9: Grand averages of all participant S1’s UTI recordings longer than 2.2 seconds.

e.g. Bever and Poeppel (2010) – to account for even these tokens. Roon (2013) (page 40) notes “Ideally, RT [reaction time] would be measured as the time between the presentation of the visual cue and the onset of articulatory movement associated with syllable.” He goes on to state that the acquisition and analysis of articulatory reaction time data requires a prohibitive amount of time and is too expensive to be used in a project such as his i.e. a PhD thesis. We hope that the work presented here provides a step towards cutting at least the analysis time requirement significantly.

As a next step, we will implement an algorithm which will search for the point of first movement in each token automatically and compare the resulting reaction times with a post hoc automatic voice key, which we will run on the corresponding audio data. Further, by comparing the automatic results with hand labelling of articulatory motion (see S. Schaeffler, J. Scobbie, and F. Schaeffler in press, for details), we will classify the tokens into ones which are analysable by the algorithm and ones that are not. It is very likely that tokens such as the paintbrush in Fig. 4 will not be easily analysable by such an automatic approach. The hand labelling will also be used to fine tune the movement detection algorithm.

5. Acknowledgements

The authors wish to thank Steve Cowen for assistance with the ultrasound recordings and Professor Alan Wrench for advice and help on extracting the raw ultrasound data from AAA and subsequent post-processing of the data.

6. References


Interaction between general prosodic factors and language-specific articulatory patterns underlies divergent outcomes of coronal stop reduction

Benjamin Parrell¹, Shrikanth Narayanan¹,²

¹Department of Linguistics, University of Southern California, USA
²Department of Electrical Engineering, University of Southern California, USA
parrell@usc.edu, shri@sipl.usc.edu

Abstract

Prosodically-conditioned reduction of /t/ and /d/ is examined in both English and Spanish using real-time magnetic resonance imaging. Results indicate that in both languages, the displacement of all three articulators used to create tongue tip closure (tongue tip, tongue body, and jaw) is conditioned by the duration of the movement. This suggests that reduction is a gradient rather than categorical process and argues against a rule-based allophonic account. Moreover, the results suggest that reduction in both languages arises from a similar cause. While the process is the same in the two languages, the details of the coronal constriction differ—English produces coronal stops at the alveolar ridge while Spanish produces them at the teeth. It is argued that this difference in articulatory posture underlies the divergent outcomes of coronal reduction in the two languages (flapping in English vs. spirantization in Spanish).

Keywords: reduction, flapping, spirantization, articulation, prosody

1. Introduction

Consonant reduction commonly occurs in many languages, but the processes underlying reduction and shaping its articulatory and acoustic outcomes are poorly understood. For example, both American English and Spanish show reduction of intervocalic coronal stops in prosodically weak positions, but the outcomes of this reduction are different in the two languages. Spanish reduces intervocalic /d/ (and, less often, /t/) to an approximant [l] in phrase-medial position. American English, on the other hand, reduces both /d/ and /t/ to a voiced flap [r] in prosodically weak positions such as before an unstressed vowel. In both languages, reduction has traditionally been described using a symbolic phonological alternation rule, though recent experimental work has shown that both processes are gradient rather than categorical, arguing against simple allophonic alternation (e.g. Parrell 2011; Jong 1998; Fukaya and Byrd 2005). We propose that reduction of stop consonants is the outcome of prosodically-conditioned durational shortening, with stops in prosodically weak positions produced with shorter durations and, consequently, small movements of the speech articulators used to create oral closure. Cross-linguistic differences in the outcome of this process in coronal stops are due to different postures of the tongue when creating coronal stops. We test this hypothesis using real-time MRI data from English and Spanish.

1.1. Subjects and stimuli

Four subjects participated in the current study. Two were native speakers of General American English. Two were native speakers of Peninsular Spanish. No subject reported any history of speech or hearing impairment.

Stimuli were designed to elicit coronal oral stops (/t/ and /d/) in a symmetric or near-symmetric low vowel context. The prosodic position of the consonant was varied to elicit a range of productions including both full and reduced forms. For American English, prosodic conditions included the stop in non-initial, word-initial, and phrase-initial positions. Both flanking vowels for the word- and phrase-medial conditions were /a/. For the word-medial condition, it was not possible to use the same vowels—a falling stress pattern (and reduced second vowel) is the conditioning factor for word-internal flapping. For this condition, the vowel context was /Ca/, which was chosen both to give a fairly close match to the vowels in the rest of the stimulus and to limit tongue movement between the full and reduced vowels. For Spanish, /a/ was used as the target vowel in all stimuli. As there is generally no effect of word-initial position in Spanish (e.g., Cole, Hualde, and Iskarous 1999; Parrell 2011), this condition was replaced with one placing the target word in a list (to induce an intermediate prosodic boundary). Stress variation was included in non-initial positions both at the sentence level (in both languages) and the lexical level (in Spanish) to induce additional variability.

For each language there were a total of 18 stimuli, which were randomized into two blocks of 9. Blocks were presented in an alternating fashion for a total of 6-8 repetitions per target phrase.

1.2. Real-time MRI data collection

Data were acquired using an MRI protocol developed especially for research on speech production, detailed in S. Narayanan et al. 2004. Subjects were supine during the scan with the head restrained in a fixed position to facilitate comparisons across acquisitions. A 13-interleaf spiral gradient echo pulse sequence was used (TR = 6.164 ms, Field of view = 200 x 200 mm, flip angle = 15°). A 3 mm slice located at the midsagittal plane of the vocal tract was scanned with a resolution of 68 x 68 pixels, giving a spatial resolution of approximately 2.9 mm per pixel. Videos were reconstructed with a 13-frame sliding window, with one frame reconstructed at every TR pulse. This gives an effective frame rate of 162.2 frames/s. Synchronous noise-cancelled audio was collected at 20 Hz during MRI acquisition (Bresch et al. 2006).

1.3. MRI data analysis

All measurements of speech articulator motion were extracted from the MRI images by means of pixel intensity values (Lam-mert, M. I. Proctor, and S. S. Narayanan 2010; M. Proctor et al.
This method is based on the fundamental idea that the changes in pixel intensity of a particular pixel over time reflect changes in tissue density at that point in the vocal tract. Lower intensities correspond to the absence of tissue (air) while high values signify the presence of one of the speech articulators at that particular point. In any given arbitrary region of the vocal tract, then, the average pixel intensity in that region will reflect the proportion of the region occupied by the speech articulators. By placing these regions at relevant locations in the vocal tract and measuring the average intensity over time, we are able to estimate speech articulator motion in that region. Each region was defined such that the relevant speech articulators (tongue tip, tongue body, jaw) were always present in the region, avoiding any floor effects which could be caused by the complete absence of the articulator from the region.

For the current study, we are interested particularly in the forward motion of the tongue body during the transition from vowel to coronal stop (or tap/approximant), the motion of the tongue tip towards the palate, and the raising of the jaw. Tongue body (TB) movement was measured by defining a long, horizontal region in the pharyngeal area of the vocal tract. This region has a vertical span from the top of the epiglottis to the bottom edge of the velum at it’s lowest position. The TB region spanned horizontally from the rear pharyngeal wall to a point roughly in the middle of the hard palate, including for each subject one pixel of the pharyngeal wall and two to three pixels of the tongue during production of /f/ (the most forward position of the tongue in the dataset). Because the pixel values in the pharyngeal region were found to vary substantially from sample to sample (indicated by jagged mean pixel intensity contours), the mean pixel intensity in the TB region was normalized by the mean pixel intensity in the entire image on a frame-by-frame basis. This was sufficient to remove a large portion of the noise from the signal without losing relevant kinematic information.

Jaw (JAW) movement was measured with a circle with a radius of 2 pixels that was placed at the base of the jaw between the jaw inflection point and the hyoid bone. The circle was placed such that when the jaw was closed, some part of the jaw was still in the circle and that when the jaw was maximally open the circle was not entirely filled by the jaw. This avoids possible saturation effects that might limit the accuracy of the measurement at extremes of jaw position. The precise location of the circle was manually determined for each subject.

Tongue tip (TT) movement was measured using a set of subregions. Each subregion had a horizontal width of only one pixel, with a vertical span of four pixels beginning at the palate. These regions were arranged in a horizontal array beginning just posterior to the teeth, past the alveolar ridge, to the end of the hard palate. This method crucially gives comparable results whether the constriction is apical or laminal. For each subject, the one of these regions with the highest maximum pixel intensity during production of the each target consonant was chosen to index tongue tip movement. Each speaker was highly consistent in the location of the produced constrictions. For English, both speakers produced all consonants at the same point, near at the inflection of the alveolar ridge. For Spanish, however, each consonant was produced at a slightly different location. Each subject was consistent within each segment, however, and the locations were similar between subjects. /d/ was measured at the first point on the palate, at the upper teeth and /l/ was measured at a point slightly behind the teeth. For examples of ROI locations, see Figure 1.

After measurement, all resulting signals (TT, TB, JAW) were smoothed using a locally weighted linear regression (M.

After measurement, all resulting signals (TT, TB, JAW) were smoothed using a locally weighted linear regression (M.

Proctor et al. 2011; Lammert, Goldstein, and Iskarous 2010). The weighting function used was a Gaussian kernel $K$ with a standard deviation of $h$ samples, where $h = 4$. As samples lying more than $3h$ from the center of the kernel in either direction receive weights near zero, this gives a smoothing window width of roughly 150 ms given the sampling period of 6.164 ms. Subsequently, all signals were individually normalized to a range from 0 to 1 for each speaker.

Gestural identification was conducted using an algorithm developed by Mark Tiede at Haskins Laboratories. The identification algorithm used takes as input a manually located estimate of the midpoint one derived variable (here pixel intensity contours). Using the velocity of that variable (the absolute value of the first difference of the signal), it then locates the velocity minimum crossing closest to the input point (measurement point: Time of maximum constriction). It then finds the peak velocity between that point and both the preceding and following velocity minima (measurement point: time of peak velocity). The algorithm then locates the onset of gestural motion by locating a point where the velocity signal from the preceding minima to the first time of peak velocity crosses some arbitrary threshold of the velocity difference between the two points. Gestural offset is defined as the point where the velocity falls below the same threshold from the second time of peak velocity to the velocity minimum following the point of maximum constriction. Onset and offset of the constriction proper are also defined by the points where the velocity crosses a threshold between the times of peak velocity and the point of maximum constriction. All thresholds were set to 20 percent. Movement duration was calculated as the time between gesture onset and constriction offset and movement displacement was calculated as the difference in normalized intensity between gesture onset and the point of maximum constriction.

2. Results

Statistical analysis was conducted using the lme4 package in R (Baayen, Davidson, and Bates 2008). Statistical significance of each predictor was assessed using the results of the t-tests given by the summary() function in the lme4 package. $P$ values and post-hoc tests were calculated using the lmerTest package (Kuznetsova, Christensen, and Brockhoff 2013). Each
language and articulator was analyzed separately with a model predicting the maximum displacement from movement duration. All models additionally included a fixed effect of token and a random intercept by subject. On visual inspection, productions in phrase-initial conditions in English show limited effects of duration (Figure 2). This seems likely due to saturation effects, where the articulator reaches its maximum possible position. In order to account for this and avoid fitting the effect of movement duration on displacement incorrectly, two additional terms were included in the model: a fixed effect of phrase boundary (phrase-initial or not) and an interaction between phrase boundary and movement duration. This effectively allows the phrase boundary condition to be fit with a different intercept and slope compared to the non-phrase boundary condition. A similar effect was seen in Spanish, where productions longer than roughly 185 ms show little effect of duration, though this does not line up as well with prosodic boundary as in English (Figure 3). For Spanish, the model was fit with an additional category (called “duration boundary”) to differentiate from prosodic phrase boundary, sorting short (less than or equal to 185 ms) and long productions (greater than 185 ms) as well as an interaction term between duration boundary and duration.

The same general pattern was found for tongue body. There was significantly more movement at extremely long durations. A nearly negligible (β = −0.009, t = −2.52, p < 0.05) indicates a reduced effect of duration of displacement in phrase-initial position.

### 2.1. American English

For the tongue tip movement, the statistical model showed a significant effect of movement duration (β = 0.022, t = 11.8, p < 0.0001). There was no difference between /d/ and /t/. The intercept for phrase-initial condition was substantially higher than for phrase-medial condition (β = 0.84, t = 9.4, p < 0.0001) and a significant interaction between phrase position and movement duration (β = −0.022, t = −10.9, p < 0.0001). The β term here, which effectively cancels the overall effect of duration, shows that there was essentially no effect of movement duration on displacement for phrase-initial productions.

The same general pattern was found for tongue body. There was a significant effect of movement duration (β = 0.007, t = 3.8, p < 0.0001) and a significant effect of both phrase boundary (β = 0.314, t = 3.5, p < 0.0001) as well as an interaction between phrase boundary and duration (β = −0.009, t = −2.52, p < 0.05), indicating a reduced effect of duration of displacement in phrase-initial position.

### 2.2. Spanish

For the Tongue Tip, the model showed significant effects of duration (β = 0.007, t = 2.2, p < 0.05) and duration boundary (β = 0.31, t = 3.4, p < 0.0001), indicating that movement displacement varies with duration and that there is significantly more movement at extremely long durations. A near-significant interaction effect between the two factors (β = −0.007, t = −2.0, p = 0.05) suggests that there is virtually no effect of movement duration on displacement over 185 ms. As for segment, /d/ shows significantly less movement that /t/ (β = 0.21, t = 10.8, p < 0.0001, /d//t/: β = −0.182, t = −9.3, p < 0.0001) but there is no difference between /d/ and /d/.

The results are very similar for Tongue Body and Jaw movements. For Tongue Body, displacement varies with duration (β = 0.012, t = 10.5, p < 0.0001) and there is significantly more displacement at durations above 185 ms (β = 0.29, t = 7.6, p < 0.0001), though the influence of duration on displacement at these very long durations is negligible (β = −0.010, t = −7.9, p < 0.0001). Unlike for Tongue Tip, there is no difference between /d/ and /t/. For Jaw movement displacement, there are similarly significant effects of duration (β = 0.015, t = 7.9, p < 0.0001), duration boundary (β = 0.20, t = 2.6, p < 0.01) and their interaction (β = −0.011, t = −5.1, p < 0.0001). As with...
the tongue tip, /t/ shows greater displacement than either /d/ ($\beta = 0.12, t = 7.3, p < 0.0001$).

3. Discussion and conclusion

While these results are based on only two subjects for each language, the consistency within and across languages suggest that American English and Spanish show very similar patterns of articulator movement as a function of prosodic context. In both languages, the amount of movement of the all the articulators used for producing a coronal stop (tongue tip, tongue body, and jaw) is heavily influence by duration. With the exception of word-medial /d/ in Spanish, there is consistent contact between the tongue tip and hard palate in both languages. While the patterns of contextual variation are similar, the posture of the tongue used to produce coronal stops differs radically between the two languages.

Importantly, this spatial and temporal reduction is not the consequence of a simple phonological alternation, as could be (and has been, in many phonological accounts of these reduction processes) suppose. That is, the evidence here does not support a symbolic substitution of one segment for another with unrelated articulatory and acoustic outcomes. Rather, the amount of tongue tip, tongue body, and jaw movement in both languages varies dynamically with changes in duration. This suggests that the extreme reduction is the magnitude of these movements in word-medial position is due to the very short articulatory durations in these positions.

While the patterns of contextual variation are similar between the two languages, the outcomes are different—spirantization in Spanish /d/ and flapping in both English coronals. This may be due to the posture of the tongue used to produce coronal stops, which differs between the two languages. In Spanish, the coronal stops are produced at or just behind the teeth (M: 1.7 mm posterior to teeth) while in English they are produced at the alveolar ridge (M: 12.7 mm posterior to teeth). There are two ways in which this difference in articulation might lead to different outcomes of reduction. First, there may be a difference in the speed with which stop closure can be created as the intrinsic muscles of the tongue used to front the tongue tip (necessary in Spanish) differ from those used to raise the tip (necessary in English); these differences may allow for the attainment of closure of alveolar but not dental stops at similar (short) durations. Second, the dental stops in Spanish appear to have a target constriction location at the upper teeth themselves, with subsequent extension along the alveolar ridge. The reduced stops in Spanish may in fact attain contact with this dental target (though such contact cannot be visualized using rtMRI), though the short duration and lack of secondary articulator movement would prevent spreading of the tongue along the palate and a complete seal of the vocal tract.

This of course does not explain the full productions of /t/ versus spirantization of /d/ in word-medial position in Spanish. This cannot be explained as just the consequence of durationally-conditioned spatial reduction as both /t/ and /d/ show similarly attenuated displacements at short durations. This difference, though, is consistent with the hypothesis that voiceless stops in Spanish have a large negative virtual constriction target (beyond the point of articulator contact, c.f. Löfqvist and Gracco 1997) while the voiced stops have a target just slightly beyond the point of articulator contact (Parrell 2011). If this analysis is correct, the tongue tip for /d/ on it’s own would just barely touch the teeth/hard palate when it reaches its target (as might occur at the long durations associated with occurring in phrase-initial position). This would make achieving a full closure of the vocal tract in this position crucially dependent on additional duration.

In sum, durational variation conditioned by prosodic structure plays a key role in coronal stop reduction in both English and Spanish. Short durations lead to smaller movements of the articulators involved in forming the tongue tip constriction, particularly the tongue body and jaw. Flapping versus spirantization outcomes of this process are due to differences in the articulatory posture of the coronal constriction between the two languages. These results are consistent with the view of lenition as a gradient rather than categorical process.

4. Acknowledgements

Work described in this paper was supported by NIH Grant DC007124.

5. References


The articulatory modeling of German coronal consonants using TADA

Manfred Pastätter, Marianne Pouplier
Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University, Munich, Germany
manfred@phonetik.uni-muenchen.de, pouplier@phonetik.uni-muenchen.de

Abstract

We present a modeling study on German coronals using the task dynamic synthesizer TADA to provide new aspects to the discussion whether coronals should be generally specified for a tongue body target to control the tongue-shape globally. We argue that all coronals must be specified for a tongue body target in order to avoid inappropriate dominance of the vowel during the consonant. This target must be weighted more strongly compared to the temporally co-existing vowel gesture. We further study the effects of changing the consonant:vowel stiffness ratio on CV and VC transitions. The stiffness settings employed in TADA for American English lead to general vowel diphthongization, which is inappropriate for German. Better results are obtained with a smaller consonant:vowel stiffness ratio which gives rise to faster transitions. These results raise the possibility that global stiffness settings differ between languages.

Keywords: German coronals, task dynamics, gestural model, TADA, tongue body target, stiffness

1. Introduction

In this paper, we present a modeling study of German coronals /t, d, n, s, z, l/ in CV and VC contexts. Our work specifically addresses the question of tongue body (TB) control during coronals. Besides being relevant for our understanding of articulatory synergies in speech motor control, the issue of TB control has been causally linked to exceptional phonotactic and phonological properties of liquids. For instance, Proctor (2011) proposed that TB control is cross-linguistically the defining feature of liquids setting them apart from other coronals. He observed that both Russian dark /l/ and Spanish clear /l/ coarticulated less with the vowel than other coronals. However, Geumann and colleagues do not agree with Proctor’s claim since their EMA data revealed no difference in vowel induced coarticulation for /t, d, n/ and clear German /l/ (Geumann et al. 1999). Also Mermelstein (1973) mentioned that coronal stops probably have a TB target. He observed that an intervocalic alveolar stop deflects the V-to-TB trajectory, suggesting that alveolar stops exert TB control in a synergistic manner to support tongue tip closure. Therefore it is important to gain a better understanding of how laterals, in particular clear /l/, can be differentiated from other coronals.

For this purpose we used TADA (Nam et al. 2004), a modularized, MATLAB based implementation of the linguistic gestural model and the associated task dynamic model (Browman and Goldstein 1990, Saltzman and Munhall 1989). Based on an alphabetic string input, TADA simulates utterance planning using systems of coupled oscillators. Thereby dynamically parameterized gestures are coupled to one another in a pairwise fashion, either in-phase (e.g., CV) or anti-phase (e.g., VC). Gestures specify abstract constriction goals on the basis of mass spring equations, that is, each gesture is specified for a rest position (i.e., target), as well as a stiffness and damping parameter. The stiffness parameter serves to distinguish between classes of sounds with vowel gestures having a lower stiffness setting than consonant gestures (e.g. Fowler 1980, Perkell 1969, Roon et al. 2007). Each gesture hierarchically controls an ensemble of articulators which are yoked in a task-specific fashion in order to achieve a particular constriction goal. For instance, a lip closure gesture is associated with the articulators upper lip, lower lip, jaw. If temporally overlapping gestures call on the same articulators with conflicting demands, weighting parameters specify the degree to which one gesture may dominate in its control over a given articulator. This effectively implements coarticulatory resistance. The result of the gestural planning process is a gestural score which specifies the gestures’ constriction goals and their relative timing. This gestural score provides the input to the task dynamic model which calculates the articulatory trajectories of the vocal tract variables. These are in turn the basis for the computation of time-varying area functions and formant frequencies by means of the vocal tract model CASY (The Haskins Configurable Articulatory Synthesizer; see Iskarous et al. (2003) and Rubin et al. (1996)). Finally, the CASY parameters are used to drive HLsyn (Hanson and Stevens 2002) which generates acoustic output consisting of the fundamental frequency and the first four formants.

The currently released version of TADA provides a full gestural dictionary comprising parameter settings for American English phonemes. In the American English model, coronals are generally implemented by a tongue tip (TT) constriction gesture. Manner differences are implemented as follows: coronal stops /t, d, n/ are characterized by a complete closure at the alveolar ridge; additionally, nasality of /n/ is rendered by a velum lowering gesture. Lateralization of /l/ cannot be addressed directly since CASY is a two-dimensional model of the vocal tract. Thus, as an approximation to lateralization the distance between TT and palate is narrowed, however, without producing friction. The coronal constriction is accompanied by a pronounced tongue body gesture, a well known characteristic of American English dark /l/ (Sprout and Fujimura 1993). To realize frication typical for sibilants /s, z/, the constriction degree is even more narrowed compared to the lateral. In addition, the gestural specifications of /s, z/ include a TB target at the articulator (as opposed to the gestural) level.

For our current work on German coronals, we generally adapted all parameter settings of TADA to reflect the Standard High German phoneme inventory. Below we discuss three articulatory manipulations used to model German coronal consonants in CV and VC contexts appropriately.
2. Modeling of German coronals using a TB articulator target

2.1. Coronal sibilants /s, z/
As already mentioned, for American English TADA, the sibilants are by default specified with a TB target at the articulator level. This active TB control assures that the TB retains its position during the production of sibilants, thus TB is less affected by V-to-C coarticulation during sibilants compared to other coronals (Recasens et al. 1997, Stone et al. 1992). Since the same coarticulatory resistance was found for German sibilants (Geumann et al. 1999), we adopted the same specifications from the American English sibilant specifications (i.e., tongue body constriction location (TBCL) = 110° and tongue body constriction degree (TBCD) = 10mm).

2.2. Coronal stops and nasal /t, d, n/
Following the American English TADA assumptions, we initially specified no TB target for /t, d, n/. Modeling CV coarticulation for German coronals /t, d, n/ we found that especially in low vowel contexts the absence of a TB target resulted in an unsatisfactory acoustic and articulatory quality for all coronals alike. As to the nasal and stops, it is possible that due to the simultaneous activation of the nasal/stop and vowel gestures (i.e., in-phase coordination (Löfqvist and Gracco 1999, Browman and Goldstein 2000), the underlying synthesizer interpreted the pharyngeal constriction of the vowel (which is achieved during the consonantal closure) as the primary constriction instead of the alveolar one of the coronal. For stops, this led to a simulated supraglottal decrease in pressure. The remaining intraoral pressure was then not sufficient to cause an appropriate burst and voice onset time at the release for /t, d/. For /n/ this misinterpretation of the primary constriction led to an /N/-like percept. The synthesis issue observed for CV held also for VC contexts: the consonantal closure (TBCL) of the consonant severely compromises the consonant’s acoustic and auditory identifiability across vowel contexts.

In order to reduce the dominance of the vowel gesture during the consonant we introduced an articulatory TB target for /t, d, n/. As a first approximation, we adopted the parameterization for the TB target specified in the American English TADA for /s/. Since then both the vowel and the consonant call on the TB articulator, we could use the weighting parameter of TADA to specify the degree of vocal tract control. Observe in Figures 1 and 2 how the addition of a TB target during the consonant leads to a wider pharyngeal opening during the stop, a stronger burst, and a longer VOT.

As mentioned earlier, Mermelstein (1973) argued for a TB target associated with coronal stops, since he found the TB configuration actively adjusted for the coronal stop in VCV sequences. He further stated that a consonantal TB target (in conjunction with jaw position control) avoids undue extension of the tongue blade during tongue tip raising towards the alveolar ridge. We observed exactly this effect of a reduced vertical tongue stretching during tongue tip closure in TADA when a TB target was introduced (Figure 1). Mermelstein treated nasals identically to stops which we also do in our simulations.

2.3. Coronal lateral /l/
Within the gestural model it is assumed that TB control for laterals is part of their gestural specification, i.e., both the tongue tip (TT) and the TB gesture form a part of the lexically specified coupling graph (Proctor 2011). German, as opposed to American English, has a clear /l/ which is characterized by a fronted, raised TB position rather than the post-dorsal retraction typical for dark /l/ (Ladefoged and Maddieson 1996).

Recall that we initially specified no TB target for /t, d, n/. Since Geumann et al. (1999) found that /l/ and the aforementioned stops showed no differences in coarticulatory variability we suspended the gestural TB target for the German lateral in the first instance. However we observed the same acoustic and articulatory issues as earlier described for /t, d, n/ caused by an overly dominant vowel gesture. To achieve a clear quality typical for German /l/ a TB target turned out to be necessary even though with a different parameterization than for the other coronals. We based our parameter choice on the descriptions in the literature which suggest a lower TB positioning for German /l/ compared to /t, d, n, s, z/ (Ladefoged and Maddieson 1996, Wängler 1961) by setting the TB target to TBCL = 110° and TBCD = 13mm.
3. Weighting of the TB target

The introduction of a TB target for all coronals allowed us to specify dominance relations between consonantal and vowel TB target for both CV and VC sequences. For the case of conflicting demands on the same articulator (i.e., CV), the blending parameter $\alpha$ is part of a gesture’s specification in the task dynamic model (Saltzman and Munhall 1989). To achieve complete dominance of, e.g., the TB vowel target over the consonant the blending parameter $\alpha$ would be set to $\alpha(\text{Vowel}) = 1$ and $\alpha(\text{Consonant}) = 100$, for an equal blending between the two parameter specifications, one would set $\alpha = 1$ for both consonant and vowel. Based on extensive auditory evaluation by the authors, $\alpha$ was adjusted for TBCD(Cons) and TBCD(Vow) to 10 and 1, respectively. For vowels, $\alpha$ was set to 1 for both parameters. This relative greater weighting of the consonantal TBCL suppresses the vocalic TBCD during the consonant resulting in a wider pharyngeal space (Figure 1). A dominance of consonantal specification over the vowel is again in accordance with Mermelstein’s (1973) observation that the V-to-V trajectory of the TB seemed to be perturbed by a consonantal target.

We further varied the specification of the blending/dominance parameter for coronals as a function of syllable position. In contrast to CV, for VC, there are no synchronous and thus no conflicting TB targets, since vowel and consonant are coordinated anti-phase. Thus the dominance relationship may remain well-balanced between the coronal consonant and the vowel, i.e., $\alpha = 1$ for both consonants and vowels.

4. The role of articulator stiffness

We found that the introduction of a consonantal target for the TB necessitated a modification of the consonant:vowel stiffness ratio. Recall that the relative speed with which a gesture reaches or moves away from its target is specified by the stiffness parameter of a gesture’s mass-spring equation. Consonants are differentiated from vowels based on this parameter: as an approximation, TADA employs globally two stiffness values for American English consonants and vowels with 8Hz and 4Hz, respectively; this means, respectively 4 closure-release cycles per second (Browman and Goldstein 1990, Nam et al. 2012). Hence, the consonant:vowel stiffness ratio is 2:1. Note that the stiffness parameter is a specification at the gestural task level and does not refer to muscular stiffness.

The introduction of a consonantal TB target and the concomitant decrease of anticipatory vowel coarticulation caused an increase in articulatory distance between the consonantal and vowel TB targets in CV (Figure 3). Particularly in a high front vowel context (e.g., /ni/) TB movement was then too slow to cover the distance between the /n/ and /i/ positions (see trajectory (2) in Figure 3). Thus TB passed through an articulatory position similar for an /e/ after the stop closure has already been released (see also Mermelstein (1973) on diphthongization resulting from a relatively lower peak velocity). This resulted in an audible, for German inappropriate diphthongization of the vowel (i.e., /nei/).

In the context of a low vowel the problem did not arise due to a lesser articulatory distance between the consonantal TB target and the low vowel target and a greater variability in producing an /a/ compared to /i/ (Hoole and Kühnert 1995). For American English, the problem might not occur at all, since American English vowels are generally diphthongized, in contrast to German. Since the diphthongization arose due to the CV transition being covered too slowly we increased stiffness for both consonants and vowels, resulting in an overall reduced consonant:vowel stiffness ratio (i.e., 1.6:1).

The stiffness of all tract variables associated with vowels (including TB and lips) was increased to 6Hz (Table 1). The stiffness of TT gestures for sonorants and sibilants was set to 10Hz and for stops 12Hz, respectively. The higher stiffness for stops evoked a considerably better burst quality; the possibility of a higher stiffness value for stops than for fricatives has been previously considered by Brown and Goldstein (1990). For the consonantal TB target, we specified a vowel-like stiffness level due to its hypothesized ‘vocalic’ property and its greater mass (Roon et al. 2007).

Table 1: Modified stiffness settings.

<table>
<thead>
<tr>
<th>Class</th>
<th>Class</th>
<th>Sounds</th>
<th>Articulator</th>
<th>Stiffness</th>
</tr>
</thead>
<tbody>
<tr>
<td>vowel</td>
<td>stop</td>
<td>/t, d/</td>
<td>TT</td>
<td>TB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sonorant</td>
<td>/n, l/</td>
<td>TT</td>
<td>TB</td>
</tr>
<tr>
<td></td>
<td>sibilant</td>
<td>/s, z/</td>
<td>TT</td>
<td>TB</td>
</tr>
</tbody>
</table>

5. Discussion and conclusion

We propose in the current paper that an adequate modeling of German coronal consonants requires the specification of a TB target for all coronals to provide appropriate consonant-vowel coarticulation patterns. Our modeling work shows that the tongue-shape behind the primary tongue tip constriction of the consonant cannot be entirely dominated by the following or preceding vowel. Therefore, the consonantal TB target needs to
be weighted more strongly than the vowel target. Further, we modified the relative stiffness difference between consonants and vowels in order to avoid vowel diphthongization. Our work raises the possibility that the relative stiffness of consonants and vowels and the concomitant speed of CV-VC transitions may be a language-specific setting, leading to the characteristic vowel diphthongization of American English. A language like German in which there is no general diphthongization of vowels, a different global consonant:vowel stiffness ratio may hold (cf. Laver (1978) on global language-specific articulatory settings). Generally, it is likely that stiffness should be distinguished in a much more fine-grained fashion on an articulator basis (Perkell 1969, Roon et al. 2007). Further it has been shown that the stiffness (and hence the velocity) of an articulator may vary over time and is not necessarily constant during an articulator’s motion. Thus to assign for each articulator movement a single stiffness value for each (as proposed by the task dynamic model) is a simplifying assumption (Fuchs et al. 2011). These aspects clearly remain an issue for future research.

Overall, our modeling results receive support from the literature in which it has been argued that V-to-T trajectories show evidence for a consonantal TB target for coronal nasal/stops (Mermelstein 1973). A question not addressed in our work so far is how to account for differences in TB variability among the coronals as typically induced by vowel coarticulation. It is well known for instance, that sibilants coarticulate very little with the vowel, while laterals show much less coarticulatory resistance (Recasens et al. 1997, Stone et al. 1992). For German, Geumann et al. (1999) studied based on EMA data to what extent coronals differ in vowel-context induced variability. Unsurprisingly, they found less variability for the TB for sibilants relative to other German coronals. Interestingly, TB during /l/ varied with vowel-context similarly to the other coronals /n, d, t/ and would support our hypothesis that all coronals have a TB target. This contrasts with findings in Proctor (2011) that in Spanish and Russian coronal stops show more dorsal vowel-conditioned variability compared to the liquids in both languages (Spanish has a clear /l/ while Russian has a dark /l/). Proctor assumes that all liquids are specified for a TB gesture and that this differentiates coronal liquids from other coronals. At least for German our modeling results and the data of Geumann et al. (1999) speak against generalizing Proctor’s findings for Spanish to other languages. While we agree that German /l/ should be specified for a TB target, this may not be a characteristic differentiating the lateral from other coronals in German. Whether the TB target for the lateral as much as for the other coronals should be considered to be controlled at the articulatory level or the gestural level (and thus participating in coupling relations at the planning level) is a question for future research. Conceivably, this differ between coronals as well as between languages.

6. Acknowledgements

Work supported by the ERC under the EU’s 7th Framework Programme (FP/2007-2013)/Grant Agreement n. 283349-SCSPL. We thank Julian Baumann and Julia Stiegert for their help with the evaluation of TADA.

7. References


The temporal coordination of Polish onset and coda clusters containing sibilants

Manfred Pastätter, Marianne Pouplier

Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University, Munich, Germany
manfred@phonetik.uni-muenchen.de, pouplier@phonetik.uni-muenchen.de

Abstract

Based on articulatory data of five speakers we examined the temporal coordination of Polish onset and coda clusters. Previous studies in the field of the gestural model suggest that cluster-vowel timing interacts with cluster composition – particularly in the case when sibilants form either C₁ or C₂. Thus, we investigated whether the position of a sibilant within onset/coda cluster affects the temporal organization of the cluster relative to the vowel. The results showed more cluster-vowel overlap only when the sibilants are C₁ of the onset cluster. However, the overlap pattern did not change with increasing complexity for clusters with vowel-adjacent sibilants. While we found systematic timing differences with respect to sibilant position in onset clusters, no such differences were apparent in coda clusters. We assume that cluster-vowel timing interacts with consonantal and even vowel coarticulation (resistance) and underscore that syllable timing patterns cannot be understood independently of cluster or syllable composition. Further, we provide some evidence about syllable affiliation of word-initial obstruent-obstruent clusters.

Keywords: Gestural model, articulatory timing, Polish, sibilant clusters, coarticulation, coarticulatory resistance

1. Introduction

The gestural approach of syllable organization predicts different cluster timing patterns as a function of syllable position (Browman and Goldstein 2000): onset clusters are by hypothesis globally aligned along the temporal midpoint of the consonants (the “c-center” effect) independent of onset complexity whereas coda clusters are sequentially organized. This has often been evaluated by comparing the timing of a cluster relative to a corresponding singleton onset/coda. Figure 1 (left) illustrates schematically the timing patterns of a complex onset (bottom panel) relative to the corresponding singleton onset (top panel): the vowel-adjacent consonant (indicated by the blue box) in the bottom panel starts later in time compared to the top panel. This relative rightward shift implies that the bottom /k/ in [skala] overlaps more with the vowel than the top /k/ in [kala]. The dashed line in the left panel of Figure 1 indicates the temporal midpoint of the singleton (top) and cluster (bottom) onset, while the solid line indicates the constant anchor point relative to which timing of the onset consonants is evaluated. Since the temporal distance between the dashed and the solid lines is constant in the singleton and cluster condition this is described as the “c-center” effect. In contrast, Figure 1 (right) shows the predicted timing pattern for singleton (top) and complex (bottom) coda. When coda complexity increases, the vowel-remote consonant /s/ in [laks] is simply ‘added’ to the singleton coda /k/ in [lak]. The so-called sequential organization has no effect on the timing of the /C₁/ in a VC₁C₂ sequence. This is illustrated by no temporal change between the anchor (solid line) and the vowel-adjacent /k/ (dashed line) in the bottom panel relative to the top panel.

While the predictions of the gestural model have been confirmed for several languages, previous studies have also revealed that cluster composition may interact with cluster-vowel timing in ways not accounted for by the model. This has become particularly evident in the case of clusters containing sibilants (Hermes et al. 2013, Marin 2013). Recently Marin’s study on Romanian consonant clusters revealed “c-center” organization only for sibilant-stop onset clusters but not for stop-sibilant ones. Coda clusters, however, consistently showed sequential organization regardless of the sibilant’s position. Marin attributed different timing patterns in Romanian sibilant-stop and stop-sibilant onset clusters to frequency effects, but the design of her study did not allow to come to a firm conclusion. An alternative interpretation of the Marin results was mentioned by Pouplier (2012) who suggested that the coarticulatory resistance of the vowel-adjacent sibilant may have conditioned the results. A sibilant may prevent an increase in consonant-vowel overlap associated with increasing onset complexity (Figure 1) and therefore the expected c-center organization may fail to emerge. Pouplier mentioned that this interpretation cannot explain the patterning of /šk-/ since /š/ is known to be even more resistant in dorsal coarticulation than /s/. Support for the hypothesized role of coarticulatory resistance comes from previous findings that sibilants showed less coarticulatory variability than other consonants (Recasens et al. 1997, Recasens 2012). In combination with Redford’s (1999) findings that syllable phonotactics interact with the mandibular cycle (i.e., the open-close movement of the jaw) this would imply that temporal coordination of

Figure 1: Schematic representation for the predicted syllable organization. Left: “c-center” effect of onset cluster in [skala]. Right: sequential organization of coda cluster in [laks].
sylables varies in a predictable fashion as a function of cluster composition. In this study, we systematically examine Polish sibilant clusters in onset and coda aiming to understand whether the position of a sibilant within a cluster affects the temporal coordination of the cluster with the vowel. For this purpose Polish serves as an interesting test case since it provides an unusual variety of consonant sequences, among others clusters combining stops and sibilants in both orders (e.g., /ps/, /pt/) in both onset and coda position. For onsets we expect C-center organization with more onset-vowel overlap for sibilant-initial (SC) but not for sibilant-final (CS) clusters due to the coarticulation resistance of the vowel-adjacent sibilant. For codas, however, we suppose that due to generally less vowel-coda overlap there is no such interaction of sibilant position within the cluster and coda-vowel timing. There is some controversy in the literature as to whether Polish prevocalic obstruent sequences form onset clusters at all (e.g., Gussmann 2007, Rochon 2000). Our present work will be able to shed further light on this controversy.

2. Method

2.1. Experimental setup and material

We collected EMA data (AG501, Carstens Medizinelektronik) with synchronized audio from five native speakers of Polish. We followed standard procedures for sensor calibration, placement and data postprocessing. Participants were asked to accentuate the target words (cf. Table 1) embedded in carrier phrases. The corpus contained sibilants in different syllable and cluster positions, i.e., the clusters SC={/ms/, /ps/, /sp/, /sk/} and CS={/mf/, /pf/, /ps/, /ks/} occurred in both onset and coda position. In addition we included target words with corresponding singletons as a baseline for the comparison of timing patterns under increasing onset/coda complexity, e.g., for onsets [skala] vs. [kala]; for coda [zam] vs. [sam] (see Measurements). The phonemic environment was – as far as possible – kept consistent across a group of singleton and cluster onset/coda words to preserve the comparability between singleton and cluster pairs and CS and SC clusters. The complete data set comprises (up to) four repetitions per cluster and subject. Some data points are missing for the first subject due to technical issues.

We placed coils/sensors mid-sagitally on the upper and lower lip, the tongue tip and the tongue dorsum to capture the articulatory trajectories associated with labials /p, m/, coronals /t, d, n, s, l, r/ and velar /k, g/. Articulatory events were identified semi-automatically on the basis of the tangential velocity profile. Due to its robustness we used for our analyses the timepoint of peak velocity (PVEL) of a gesture’s closing movement (Figure 2, solid red lines). We preferred PVEL since the maximum constriction appeared to be less stable.

2.2. Measurements

For each singleton and cluster target word we measured the temporal lags between onset/coda constituents and a constant anchor point. The anchor point was the consonant following and preceding the vowel in onsets and coda clusters, respectively (e.g., /l/ in onset target word [skala]; /s/ in coda target word [zam]). In keeping with previous work (Brownman and Goldstein 1988, Marin and Pouplier 2010) we indirectly determined the cluster-vowel timing as follows: for onset/coda cluster words we measured the difference in lag ratios as a function of sibilant position and the anchor point (henceforth V-adjacent lag; e.g., for onset: k+1 in [skala]; for coda: z+1 in [zam]); further we measured the distance between the temporal midpoint of the consonant(s) and the anchor point (henceforth C-center lag; e.g., for onset: sk+1 in [skala], k+1 in [underline kala]; for codas: z+1 in [zam], s+1 in [sam]). Figure 2 exemplifies for the /sk/ cluster in [skala] the measurement points for of the V-adjacent and C-center lags.

To quantify the relative timing differences between singleton and cluster words, we computed lag ratios for each cluster as follows: we averaged all lag measurements of a given singleton condition (e.g., [kala]); then we compared each occurrence of the corresponding cluster condition (e.g., [skala]) relative to the singleton mean value; finally we centered the lag ratios to 0. This was done for V-adjacent and C-center lags alike. Positive lag ratios represent a lag increase between singleton and cluster, i.e., less overlap with the vowel in cluster than in singleton target words. Negative lag ratios indicate a shift towards the anchor, i.e., more consonant-vowel overlap with increasing complexity. Lag ratios around 0 suggest no change in timing compared to the corresponding singleton.

For onsets, the gestural model predicts a negative lag ratio for the V-adjacent measure and a lag ratio around 0 for the C-center measure (Figure 1 left). For codas, the V-adjacent measure should not be affected by coda complexity while the C-center measure should be positive (sequential organization; Figure 1 right). If, as we predict, coarticulation resistance prevents increasing onset-vowel overlap in the case of vowel-adjacent sibilants, there would be an interaction of these measures with sibilant position: for SC clusters there should be a negative V-adjacent lag ratio and a C-center lag ratio around 0; for CS clusters, however, a V-adjacent lag ratio around 0 and a positive C-center lag ratio are expected. For codas we predict no difference in lag ratios as a function of sibilant position. We expect for both cluster types V-adjacent lag ratios around 0 and

---

**Figure 2: Articulatory labeling and lag measurements.**
2.3. Statistics

To test the global differences between sibilant-initial (SC) and sibilant-final (CS) clusters statistically we applied two linear mixed models, one with the C-center lag, the other with the V-adjacent lag as the dependent variable. Sibilant Position (two levels: SC and CS) and onset/coda Complexity (two levels: singleton and cluster) were fixed factors; Speaker and Set (pairs of singleton and cluster target words) were random factors. P-values were obtained by comparing, for example, one model with and one without the interaction of the fixed factors. Based on our hypotheses we predict for onsets a significant interaction of sibilant position and complexity for both lag ratios. For codas on the other hand we expect no significant difference in vowel-coda timing between SC and CS clusters.

3. Results

3.1. Temporal coordination of onset clusters

Results for onsets are given in the left plot of Figure 3. For SC clusters, we found as predicted a “c-center” organization with more consonant-vowel overlap in clusters than in the corresponding singletons. This is evident by consistent negative V-adjacent lag ratios for those clusters (light gray bars). In the case of CS clusters, the V-adjacent lag ratios are predominantly above or around 0, indicating no change in consonant-vowel overlap in clusters compared to singletons. Confirming our predictions, the interaction of Sibilant Position and Complexity was significant ($\chi^2[1] = 25.8$, $p<.001$), i.e., the V-adjacent lag ratios differed significantly between CS and SC clusters. The light gray bars suggest that within the CS and SC groups, not all clusters behave the same. We therefore ran two additional mixed models, separately for CS and SC clusters in order to determine whether the clusters within these two groups differed significantly from each other (dependent variable: V-adjacent lags; fixed factors: Set and Complexity; random factor: Subject). Results revealed a significant interaction of Complexity and CS sets ($\chi^2[3] = 16.5$, $p<.001$), i.e., the V-adjacent lag ratios differed between CS sets. Post-hoc Tukey tests revealed that consonant-vowel overlap increased significantly only for set /ks/ ($p<.05$), i.e., the vowel-adjacent consonant /s/ in [kszero] shifted towards the vowel and, thus, overlapped more with the following vowel than in the singleton condition [zero]. This is the pattern that we generally found for SC clusters and that is predicted for a “c-center” onset coordination. For the remaining CS clusters (/mʃ/, /sp/ and /pʃ/) consonant-vowel overlap did not change significantly between singleton and cluster condition. Similarly, the effect of complexity on V-adjacent lags differed significantly between SC cluster sets (Complexity × Set interaction; $\chi^2[3] = 11.8$, $p<.01$). All sets showed more consonant-vowel overlap in clusters than in singletons. Post-hoc Tukey tests revealed complexity effects for /sk/ ($p<.001$), /ʃm/ ($p<.01$) and /ʃp/ ($p<.001$) but not for /sp/.

Onset C-center lag ratios (dark gray bars, Figure 3 left) patterned more homogeneously showing consistently positive values for CS and SC clusters. Statistically, the C-center lag ratios differed significantly between SC and CS clusters ($\chi^2[1] = 13.4$, $p<.001$), i.e., there was an overall bigger C-center lag ratio for CS clusters compared to SC clusters. In analogy to the V-adjacent lag measurements, we conducted follow-up mixed models for the C-center lags since the dark gray bars suggest set-dependent variability (dependent variable: C-center lags; fixed factors: Set and Complexity; random factor: Subject). Considering sibilant position, we found that C-center lag ratios differed between CS sets ($\chi^2[3] = 9.9$, $p<.05$). Post-hoc Tukey test revealed a global shift away from the vowel in the cluster relative to the singleton condition for /mʃ/ ($p<.001$), /sp/ ($p<.001$) and /pʃ/ ($p<.01$). However, /ks/ showed “c-center” organization since the C-center lag did not increase significantly with complexity. The differences between sets were also significant for SC ($\chi^2[3] = 14.0$, $p<.01$). Post-hoc Tukey test ascertained that clusters /sk/ and /ʃp/ show “c-center” organization, since C-center lags did not change significantly between singleton and cluster condition. In contrast the remaining clusters shifted globally away from the vowel, i.e., no global alignment (/ʃm/; $p<.001$; /ʃp/; $p<.01$).

3.2. Temporal coordination of coda clusters

The lag ratios for coda clusters are shown in Figure 3 (right). Similarly to the results of the sibilant-initial (SC) onset clusters the V-adjacent lag ratios showed for sibilant-final (CS) clusters
consistently negative values (Fig 3 right, light grey bars). This indicates a slightly decreasing lag between the anchor point and the vowel-adjacent consonant and, thus, more consonant-vowel overlap in the cluster compared to the singleton condition. The V-adjacent lag ratios of SC coda clusters, however, fall around 0, i.e., there is no change in vowel-consonant timing relative to the singleton condition. Since differences concerning consonant-vowel overlap (indicated by V-adjacent measure) was small between CS and SC coda clusters, syllable position and complexity interacted only at trend level ($\chi^2[1] = 3.1$, p<.1). However, there was no interaction of Set and Complexity in the follow-up mixed models, Post-hoc Tukey tests revealed that vowel adjacent consonant of /p/-/ shifted towards the preceding vowel as the /f/ was added (p<.01). Concerning consonant-vowel overlap in SC clusters, we found no set-specific differences. Finally, both CS and SC clusters shifted globally away from the vowel, i.e., the C-center lag increased with coda complexity for CS and SC alike. In sum, with the only exception /p/-/ CS and SC clusters showed sequential organization.

4. Discussion and conclusion

We hypothesized that syllable organization of onset clusters should be affected by the sibilant’s position within the cluster since more onset-vowel overlap is expected for sibilant-initial than sibilant-final clusters. For coda clusters, however, we expected no such effect due to generally less consonant-vowel overlap. The results confirm our predictions. We found more consonant-vowel overlap when sibilants are C1 in a C1C2 onset cluster (i.e., SC). In contrast the vowel-adjacent sibilant overlapped less with the vowel in CS onset clusters. Further, our results showed no differences in vowel-consonant timing in CS and SC coda clusters. This means that our findings deviate to some extent from the predictions made by the gestural model (Browman and Goldstein 1988, 2000): concerning the V-adjacent measures, CS onset clusters showed sequential alignment instead of “c-center” organization; further, CS and SC onset clusters showed a significant complexity effect for the C-center lag measure. The gestural organization of coda clusters is, however, predominantly in line with the gestural model. The timing patterns observed in our data agree with recent results of Marin (2013) who found “c-center” organization only in sibilant-stop but not in stop-sibilant onset clusters for Romanian. That coda clusters are not affected by sibilant position in the current data is also consistent with Marin’s findings. These results can be interpreted in terms of coarticulatory resistance and aggressiveness of consonants since sibilants were found to coarticulate less and – at the same time – to trigger more coarticulation than other consonants (Recasens et al. 1997, 2012). This may account for why onset clusters with a vowel-adjacent sibilant showed generally less consonant-vowel overlap compared to clusters in which sibilants formed the edge of the target word. In sum, the timing patterns generally support the assumption that cluster-vowel timing interacts with coarticulation resistance and underscore that syllable timing patterns cannot be understood independently of cluster composition. We also found differing timing patterns between CS sets: while /m/-/, /p/-/ predominantly showed V-adjacent lag ratios above or around 0 (i.e., increasing or unaltered timing patterns with increased complexity), /ks/-/ showed a significant increase in consonant-vowel overlap compared to the singleton condition. Another set-dependent timing pattern concerns the coda cluster /p/-/: we found a significant shift of the vowel-adjacent consonant towards the preceding vowel. Interestingly, both clusters differ from the other stimuli used in the experiment in terms of vowel context, i.e., /ks/-/ and /p/-/ precede/follow the front vowel /e/ while the other clusters are adjacent to back vowels /a, o, o/. This suggests that also vowel identity may affect syllable timing.

In this respect the jaw could play a decisive role (cf. Redford 1999). It is possible that consonant-vowel overlap in onset /ks/-/ and even in coda /p/-/ can be achieved since the target position of the jaw is relatively similar for both the vowel-adjacent consonant and the vowel /e/. This would be in accordance with Redford (1999) who observed a positive displacement-duration relationship for the open-close movement of the jaw. However, whether and to what extent the syllable timing patterns is affected by jaw position constraints remains for future research. Finally, our results provide some new aspects to the controversy whether initial/final obstruent sequences are complex onsets/codas in Polish or not. From an articulatory point of view, we found evidence for syllabic organization since our results showed systematic differences in terms of sibilant position and vowel context. However, the complex interaction of gestural overlap, jaw movement and coarticulation in Polish syllable organization remains a topic for further investigation.

5. Acknowledgements

We thank Susanne Waltl for assistance with data collection. Work supported by the ERC under the EU’s 7th Framework Progamme (FP/2007-2013)/Grant Agreement n. 283349-SCSPL.

6. References


Articulatory mechanisms underlying incremental compensatory vowel shortening in German

Sandra Peters, Felicitas Kleber

Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University, Munich, Germany
sandra@phonetik.uni-muenchen.de, kleber@phonetik.uni-muenchen.de

Abstract

The main aim of this study was to investigate the articulatory mechanisms underlying incremental vowel shortening in German. Five speakers were recorded reading sentences with target words that contained either simplex (one consonant) or complex (two consonants) onsets and codas, respectively. Acoustic measurements showed that both onset and coda clusters induce vowel shortening. Two potential articulatory mechanisms that may cause this shortening are taken into account: (1) a compression of the vowel gesture’s plateau in words containing complex onsets or codas and/or (2) a shift of the vowel adjacent consonant towards the vowel in these tokens. Vowel plateaus were not significantly compressed in cluster words, but the CV overlap in onset cluster tokens increased indicating a shift towards the vowel. Despite the acoustic vowel shortening in words with coda clusters, there was no significant shift towards the vowel (only a trend). The degree of VC overlap also depended on onset’s manner of articulation. The results are discussed with respect to the predictions made by articulatory phonology regarding gestural timing.

Keywords: Compensatory shortening, coda/onset clusters, Articulatory Phonology, EMA

1. Introduction

This study addresses the interplay between acoustic durations of syllable segments and the underlying gestural coordination of their articulators. More precisely, it investigates the articulatory mechanisms leading to acoustic vowel shortening in closed syllables with an increasing number of consonants in either onset or coda position (onset vs. coda shortening). This compression effect is commonly known as incremental compensatory shortening (Munhall et al 1992), i.e. vowels are shorter in complex CCVC or CVCC syllables than in simplex CVC syllables (Katz 2012).

Our knowledge about incremental compensatory shortening partly comes from the extensive work on gestural coordination within syllables that has been conducted in the framework of Articulatory Phonology and c-center studies (e.g. Browman and Goldstein 1988; Marin 2013; Marin and Pouplier 2010) providing evidence for shorter vowel durations after complex onsets as compared to simplex onsets. Studies in favor of the c-center hypothesis argue, that this is due to the global timing of onset clusters, in which consonants are organized around a stable midpoint of the cluster – the c-center – causing vowel remote consonants to move away from the nucleus and vowel adjacent consonants to move towards the vocalic nucleus. This may then result in an increased CV overlap and therefore in acoustic vowel shortening. Coda clusters, on the other hand, are considered to be locally timed, i.e. in a sequential order, not causing any increase in overlap between the vowel and the following coda consonant (henceforth VC overlap) and therefore also no acoustic vowel shortening.

Findings from recent studies, however, challenge the general assumption of globally timed onset clusters and sequentially ordered coda clusters, as acoustic and articulatory evidence for coda-induced incremental vowel shortening increases (e.g. Katz 2012, Munhall et al. 1992) and a global timing of onset clusters seems to depend on manner of articulation (see e.g. Marin 2013 for Romanian clusters). The effect of manner on global timing may very likely be a consequence of interconsonantal timing, which can also be measured in terms of overlap. For example, Bombien (2011) found more overlap between /k/ and /l/ than between /k/ and /n/. Though he did not relate this finding to a potential c-center within these clusters, this difference in consonant cluster timing, nevertheless, may in turn affect the global organization proposed for onsets and the consonant’s shift into the vowel.

Those instances of acoustically measured incremental vowel reductions are, according to Katz (2012), caused by two possible articulatory mechanisms. It may either come about because of a compression of the vowel gesture or by a consonantal shift towards the vowel, the latter increasing the CV or VC overlap. But a combination of these two proposed mechanisms may also be conceivable.

The aims of the present study were twofold: (1) to investigate the articulatory mechanisms underlying acoustic incremental vowel shortening driven by both onset and coda clusters and (2) to test the effect of onset manner on vowel shortening. Because of the second aim, we included the same onset consonants as in Bombien’s (2011) study. The following four hypotheses were tested in the current study:

H1 Vowels are acoustically shorter in words with complex onsets or complex codas compared to their simplex counterparts.  
H2 The vowel plateau duration is shorter in words containing complex onsets or complex codas compared to their simplex counterparts.  
H3 There is more CV overlap in words containing complex onsets and more VC overlap in words containing complex codas compared to their simplex counterparts.  
H4 There is less acoustic vowel shortening, less vowel plateau compression and less CV overlap in /kl/-tokens.

2. Method

2.1. Speech Material and experimental set-up

The test items were non-existent words containing simplex and complex onsets and codas, respectively, and the tense vowel /a:/ as the nucleus. To address the research questions, the test items differed from those words typically used in c-center studies. The consonants surrounding the vowel were
kept identical for onset and coda analyses, thus preventing any other influences on vowel duration. Simplex onsets consisted of /l/ or /n/ (henceforth /l/- or /n/-words), simplex codas consisted of /p/. In order to measure articulatory tongue movements (1) between vowels and adjacent consonants and (2) between the clusters’ consonants, complex onsets contained either /kl/ or /kn/ and complex codas contained /pt/. All target words are listed in Table 1, separately for onset and coda shortening. As we were interested in both onset and coda shortening, singleton tokens were the same in both analyses.

Table 1: l- and n-target words used in the analyses of onset and coda shortening

<table>
<thead>
<tr>
<th></th>
<th>Onset comparison</th>
<th>Coda comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>l-words</td>
<td>/kla:p/</td>
<td>/na:p/</td>
</tr>
<tr>
<td>n-words</td>
<td>/kna:p/</td>
<td>/na:p/</td>
</tr>
</tbody>
</table>

For syntactical reasons, all target words were embedded in one of the two carrier phrases: “Melanie’s Omi [target word] imitiert ein Lied.” (“Melanie’s grandma [target word] imitates a song.”) was used for noun-like singleton coda and all onset tokens. “Melanie’s Omi [target word] ihm einmal.” (“Melanie’s grandma [target word] him once.”) was used for verb-like coda clusters.

Articulatory recordings were made in a sound attenuated booth at the Institute of Phonetics in Munich. The movements of speech articulators were tracked using 3D-Electromagnetic Articulography (EMA, AG 501; cf. Hoole et al. 2003). In total, ten EMA coils were attached to various parts of each speaker’s head. The sensors included in the present analysis were those attached to the tongue back (TB), tongue tip (TT) and to the upper and lower lips (LA). Seven repetitions of each sentence were presented isolated in randomized order on a computer screen.

2.2. Participants

Five speakers (3 females, 2 males) of Southern Standard German aged between 19 and 25 were recorded. None of them reported any hearing or speaking disorders. Three out of the five participants were undergraduate students of phonetics, but they were naive as to the purpose of the experiment. One participant was the first author of this paper.

2.3. Acoustic measurements

Segmentation and labeling of the speech signals were done automatically using MAUs (Schiel 2004). The segment boundaries of the target words were hand-corrected in Praat (Boersma and Weenink 1992) when necessary. Vowel duration (Vdur) was measured from the point of spectral change in F2 of the preceding nasal/lateral up to the end of F2.

2.4. Articulatory timing measurements

For the present study, we only analyzed temporal articulatory measures, i.e. specific moments in time of vertical and horizontal movements of the relevant sensors. The physiological data was labeled using Emu (Harrington 2010). The target words’ components /k/ and /a/ labels were set based on the vertical movements of the tongue back. The labels of the alveolars /l/, /n/ and /t/ were based on the tangential velocity of the tongue tip. For /p/, the lip aperture was calculated as the Euclidean distance between upper and lower lips. Plateau’s onsets and offsets were defined on the basis of changes in the articulators’ velocity, which were interpolated values and represent the 20% threshold of the difference between two adjacent maxima in the velocity signal. Prior analyses have shown that vowels vary acoustically depending on the number of consonants within a cluster (see e.g. Marin and Poupiller 2010). Because of this (potential) variability, we again did not use normalization methods usually applied in c-center studies as they normalize on anchor points using either the following coda consonant (Marin and Poupiller 2010) or the acoustic vowel midpoint (e.g. Poupiller 2012). Thus, we used the normalization method described in Bombien (2011) instead. Following this method, we first determined the lag between two neighboring segments and then normalized this lag on the entire duration of these two gestures. In order to measure the CVlag, the consonant’s (C2on) plateau offset (Poff) was subtracted by the vowel’s (V) plateau onset (Pon) and divided by the entire gesture (G) of these two syllable constituents as described in equation (1).

\[ CV_{lag} = \frac{P_{on}[V]-P_{off}[C_{2on}]}{G_{off}[V]-G_{on}[C_{2on}]} \]

The same procedure was applied to the VClag with the exception of subtracting the vowel’s (V) plateau offset by the coda consonant’s (C1on) plateau onset (cf. equation (2)). Lower lag values for complex onsets or complex codas as compared to their simplex counterparts indicate a shift towards the vowel and thus more overlap.

\[ VC_{lag} = \frac{P_{on}[C_{1on}]-P_{off}[V]}{G_{off}[C_{1on}]-G_{on}[V]} \]

The vowel plateau durations (Vplat) were calculated by subtracting the plateau onset by the plateau offset. Thus, higher values signify longer plateau durations. Figure 1 schematically displays the specific landmarks used in this analysis.

For the statistical analyses we conducted various repeated measures ANOVAs. Vplat, CVlag and VClag each served as the dependent variable. ONSET or CODA COMPLEXITY (two levels each: simplex vs. complex) and ONSET MANNER (two levels: /l/ vs. /n/) were the independent variables and speaker was entered as random factor.

3. Results

3.1. Acoustic vowel duration

Prior to the articulatory analyses, we verified whether clusters in either onset or coda position acoustically shortened the duration of the tense vowel /a/.
Both ONSET \((F[1,4]=33.7, \ p<0.01)\) and CODA COMPLEXITY \((F[1,4]=40.2, \ p<0.01)\) had a significant effect on the acoustic vowel duration. That is, in words containing complex onsets and complex codas, respectively, the acoustic vowel duration was shorter compared to their simplex counterparts (cf. Figure 2).

![Figure 2: Acoustic vowel duration in words containing either simplex (light grey) or complex (dark grey) onsets (left) and codas (right), respectively.](image)

Further, there was no effect of ONSET MANNER, indicating that there was no difference in acoustic vowel compression between /n/- and /l/-words.

### 3.2. Vowel plateau duration

In order to investigate the influence of ONSET and CODA COMPLEXITY as well as ONSET MANNER on the vowel plateau duration, we analyzed words containing both complex (/kl/ vs. /kn/) onsets and complex (/pt/) codas and their simplex counterparts.

![Figure 3: Vowel plateau duration separately for /l/- (left) and /n/-words (right) containing simplex (light grey) and complex (dark grey) onsets or codas, respectively.](image)

Neither ONSET nor CODA COMPLEXITY, nor ONSET MANNER had a significant influence on vowel plateau duration (cf. Figure 3). This result suggests that a compression of the vowel plateau seems not to be an underlying articulatory mechanism leading to the compression effects found in the acoustic measurements, though there was a tendency for vowel plateau compression in complex tokens—especially in /n/-words. These effects may not have been significant because of the higher variability in vowel plateau duration.

### 3.3. CV lag

Words containing both simplex (/l/ vs. /n/) and complex (/kl/ vs. /kn/) onsets but only simplex codas were included in the analysis of the effects of ONSET COMPLEXITY and ONSET MANNER on the timing between the vowel adjacent onset consonant (C2on) and the vowel.

![Figure 4: CVlag separately for /l/- (left) and /n/-words (right) containing simplex (light grey) and complex (dark grey) onsets.](image)

ONSET COMPLEXITY had a significant influence on the CVlag \((F[1,4]=7.8, \ p<0.05)\). Lag values were significantly smaller in complex onsets than in simplex onsets, indicating that the plateau of the onset cluster consonant C2on shifted towards the vowel plateau.

There was no significant influence of the independent variable ONSET MANNER. This means, that there was a shift towards the vowel in both /l/- and /n/-words. However, there was a strong tendency towards more CV overlap in /n/-words, recognizable in lower lag values in these tokens (cf. Figure 4). Two separate repeated measures ANOVAs revealed a significant difference between simplex and complex /n/-words \((F[1,4]=16.7, \ p<0.05)\), but no difference between simplex and complex /l/-words. This then supports the obvious trend between /l/- and /n/-words described above.

### 3.4. VC lag

In order to investigate the influence of CODA COMPLEXITY and ONSET MANNER on the timing between the vowel and the adjacent coda consonant (C1on), we analyzed words containing only simplex (/l/ vs. /n/) onsets and both, simplex (/p/) and complex (/pt/) codas.

![Figure 5: VClag separately for /l/- (left) and /n/-words (right) containing simplex (light grey) and complex (dark grey) codas.](image)
Concerning the VC lag, there was no significant effect of CODA COMPLEXITY. This finding suggests that the plateau of the coda consonant did not shift towards the vowel plateau in complex tokens – although there was coda driven compensatory vowel shortening in the acoustics. Again, ONSET MANNER had no significant influence. For /l/-words, however, there was a trend for /p/ to shift towards the vowel in complex codas compared to simplex codas (cf. Figure 5).

4. Discussion and Conclusion

There were four main findings from this study. First, the acoustic vowel duration was shorter in words containing both onset and coda clusters compared to words containing singletons, thus confirming hypothesis H1. The articulatory analysis revealed, on the one hand, that vowel plateaus did not differ in duration in simplex versus complex onsets and codas, respectively, but, on the other hand, lower CVlag and VClag values in complex than in simplex onsets and codas, respectively. The results support hypothesis H3, which predicted a shift of the vowel adjacent consonant towards the vowel, and hypothesis H2 must be rejected. In the present data both onset and coda driven acoustic vowel shortening stem from a greater overlap (indicated by lower lag values) between the vowel and its adjacent consonants and not from a compression of the vowel plateau. The fourth finding was the notable shift of /n/, but not of /l/, towards the vowel in complex onset tokens. This result may be due to the greater overlap between /k/ and /l/ than between /k/ and /n/ in German onset clusters reported in Bombien (2011). Thus, it may be the different coordination patterns between consonant gestures that affect the degree of vowel shortening: if there is more overlap between /l/ and the preceding /k/, then /l/ may not be shifted towards the vowel resulting in less CV overlap. Our lag results therefore point to a global organization of the complex onset /kn/. Whether the German /kl/ cluster is sequentially ordered or globally organized remains to be tested in future analyses.

In addition to the findings arising from the four hypotheses, we also found an effect of onset manner on the timing between the vowel and coda clusters, although the coda was kept constant in all tokens: there was more VC overlap in complex coda tokens when the onset consonant was /l/ than when it was /n/. This onset manner dependent shift of coda clusters may be a direct consequence of the degree with which the onset consonant shifts into vowel: there was no shift of /l/ towards the vowel in complex onsets, but more VC overlap in complex coda tokens with singleton /l/ in onset position. The nasal, on the other hand, shifted towards the vowel in complex onsets, but the VC overlap was about the same in complex and in simplex codas with singleton /n/ in onset position. This suggests that VC(C) organization may also be predicted to some extent by (C)CV organization, or vice versa (see also Hawkins and Nguyen 2000).

To conclude, vowel shortening is caused by an increase in gestural overlap between the nucleus and the adjacent consonant in onset or coda clusters. Both syllable position and manner of articulation determine the degree of vowel compression. The complex onset /kn/ appears to be globally organized, which would be in line with the predictions made by Articulatory Phonology. Tendencies for less CV and greater VC overlap in /l/-words, however, are not predicted by this theory. This and the acoustic shortening before complex codas suggest that coda clusters are not entirely sequentially organized. In addition to durational measures, spatial parameters such as stiffness may broaden our understanding of the articulatory mechanisms underlying acoustic vowel compression.

5. Acknowledgements

This research was supported by ERC grant number 295573 ‘Sound change and the acquisition of speech’ to Jonathan Harrington.

6. References


Working memory and stress clash resolution in Multiple Sclerosis

Caterina Petrone1,2, Céline De Looze3, Léa Gomis2, Marjorie Wims2, Alain Ghio1,2, Noémie Moreau1,4 & Laurent Renié1,4

1Laboratoire Parole et Langage, CNRS, 2Aix-Marseille Université, France
3Phonetics and Speech Laboratory, CLCS, Trinity College Dublin, Ireland,
4Neurology Department, Pays d’Aix Hospital, Aix-en-Provence, France
caterina.petrone@lpl-aix.fr, delooze@tcd.ie

Abstract

In this paper, we investigate stress clash resolution, i.e., how speakers readjust the metrical patterns of an utterance to preserve its metrical regularity. We test the effect of working memory (WM) capacities in control (C) subjects and individuals with Multiple Sclerosis (MS) on their ability to look-ahead to the metrical structure of an utterance and compute changes in prominence structure in the clash condition. We predict that if WM capacities affect stress clash resolution, MS patients with low WM capacities will not be able to look ahead and will fail to apply stress resolution. Our results do not provide evidence of stress clash resolution for both C and MS subjects, suggesting neither stress shift nor stress deletion took place. Yet, some evidence is provided that MS speakers with low WM capacities narrow down the scope of the prosodic planning unit, thus suggesting that variation in planning strategies depend on cognitive skills. We discuss our results along different models of French prosody.

Keywords: stress clash, Multiple Sclerosis, working memory, prosodic structure, French.

1. Introduction

When speakers are to produce an utterance, they plan what they want to say and the way they want to say it, e.g., they choose the segments, the words, the prosody of that utterance. Although the incremental character of language production is uncontroversial (Levlt, 1989), it is still difficult to determine how far ahead speakers do plan and whether advance planning might differ at different representational levels (Wheeldon & Lahiri, 1997; Krivokapić, 2007). It has been suggested that the scope of planning does not coincide with a fixed linguistic unit, but might be flexibly adapted by the speakers (Swets et al., 2007). In particular, speaker-specific variations would reflect differences in working memory (henceforth WM) capacities. “Working memory” refers to the memory subsystem which is responsible for the active maintenance of mental representations along ongoing processing and/or distractions. Swets et al. (2007) argues that there is an effect of WM capacities on prosodic structure, such that readers with low WM are more likely to chunk a text into smaller prosodic phrases than readers with high WM. Petrone et al. (2011) found that German speakers with high WM capacities start their utterances with higher F0 peaks than speakers with low WM capacities in order to apply F0 declination (the gradual decrease of F0 over the course of an utterance) on larger portions of the utterance.

In this study, we investigate working memory constraints on underlying variability in the scope of prosodic planning by comparing healthy subjects with individuals affected by cognitive disorders due to multiple sclerosis (MS). MS is a progressive inflammatory disease of the central nervous system which damages the myelin sheath. MS is characterized by cognitive deficits (such as WM impairments, cf. Fuso et al., 2010) which can occur independently of other disorders. A preliminary study on French (Petrone et al., 2013) found that MS patients affected by working memory impairments (but not by speech motor control disorders) were not capable of planning long prosodic phrases in French. Patients with low WM produced more prosodic breaks in unpredictable locations and their speech rate was slower than that of patients with high WM.

Since MS patients with low WM show difficulties in the temporal organization of their speech, we predict that they will also have difficulties in planning other aspects of prosody, such as the metrical structure. In the current paper, we focus on stress clash resolution, i.e., how speakers readjust the metrical patterns in a sequence [word1 word2] in order to preserve metrical regularity (Nespor & Vogel, 1986; cf. Tilsen, 2012 for a review). Phonologically, metrical regularity is the consequence of the rhythm rule, by which a stress is shifted at an earlier location in the word or deleted to avoid clash. A case in which the rhythm rule applies is when a word with late stress like JapaNESE creates a clash when followed by a word with early stress like INstitute (Keating & Shattuck-Hufnagel, 2002). The clash is resolved by reducing the prominence of the last syllable of word1 (e.g., by decreasing its duration and intensity and by deleting or shifting the pitch accent) and eventually by making the first syllable of word1 more prominent (e.g., by means of a pitch accent).

In French, the word-final (full) syllable is usually the most prominent and a primary accent can be assigned to this position. Stress clash has been traditionally defined as the occurrence of two consecutive primary accents (Post, 1999). Stress clash resolution is claimed to be obligatory when the two accents are within the same phonological phrase (Post, 1999) For instance, in a sequence [word1 word2] such as “jolLiS LYS” (nice lilies), accent resolution involves the deletion (“jolIS LYS”) or the shift (“jOlIS LYS”) of the pitch accent of the word final-syllable (Post, 1999; Avanzi & Schwab, 2013). Moreover, the duration of the word-final syllable is shorter when the stress clash resolution applies (Avanzi & Schwab, 2013).

Different accounts of stress clash in French have been proposed. In particular, Jun & Fougeron (2000) claimed the existence of an Accentual Phrase (AP), whose underlying pattern is /LHi LH*/. In this account, clash resolution would be reinterpreted in terms of variation in the tonal pattern due to a specific constraint, by which, for instance, the AP final H* would be realized as a L* to avoid a sequence of three consecutive H tones (AVOD HHH). However, the syllable associated to L* would still be lengthened.
Stress clash is a good candidate to test the effects of WM on prosodic planning, since clashes are not stored in lexical memory and thus their resolution require a more on-line building of the prosodic structure (Tilsen, 2012). When producing word1, the speaker must already take into account the metrical pattern of the upcoming word2 to apply the rhythm rule (Keating & Shattuck-Hufnagel, 2002). If WM capacities affect stress clash resolution, both healthy speakers and MS patients with high WM capacities will be able to look-ahead to the metrical structure of following words and compute changes in prominence structure in the clash condition. Following Post (1999) and Avanzi & Schwab (2013), we predict that speakers will reduce the duration of the final syllable of word1 and they will either delete the accent or shift it from word1 to word2. On the contrary, the scope of planning will be narrower in MS patients with low WM spans and their look-ahead will be reduced. As a consequence, they will fail to apply stress clash resolution. The accent on word1 final syllable will be similar in clash and no-clash conditions or the prosodic structure of the sequence will be affected (e.g. prosodic break, hesitation inserted between word1 and word2).

2. Methods

2.1. Participants

Thirty MS patients (22 women and 8 men) and thirty-one healthy (22 women and 9 men) speakers participated in the experiment. The mean age is 43 y.o. for both the MS (ranging from 25 to 63 y.o.) and the control group (29-69 y.o.). The two groups were also matched in educational level. Results for twenty speakers analyzed are reported here, i.e., ten speakers for each group. The inclusion criteria are detailed below.

2.1.1. Multiple sclerosis group (MS)

MS patients meeting with a clinically definite relapsing-remitting form of multiple sclerosis (i.e., periods of acute inflammation in which a deficit appear, followed by periods in which the patient eventually recovers from the deficit, cf. McDonald et al. 2001) took part in the study. Patients were included with (1) no clinical relapses and steroid treatment at least one month prior to the study; (2) no concomitant therapy with antidepressant or psychoactive drugs; (3) optimal visual acuity; (4) no dyslexia; (5) no dysarthria (as assessed by structural MRI scanning and clinical examination). The experiments were run in the CNRS rooms at the Neurology Department in Pays d’Aix Hospital (Aix-en-Provence), in the anechoic room of the Laboratoire Parole et Langage and at the Hospital La Timone (Marseille). The participants agreed with the experimental procedure and signed an informed consent. Experiments were conducted under ethical conditions.

2.2. Working memory tasks

Prior to the acoustic recordings, MS and healthy speakers underwent a battery of cognitive assessments including a series of working memory span tasks. The span tasks are based on the forward and backward recall of digits as well as on the recall of letters and number sequences (Brissart et al., 2008). The administration time of this battery was 15 minutes. The letter-number sequencing task better discriminated the two groups [β= -0.33. SE= 0.13, z = -2.57, p = 0.01]. Only correlations between the WM score obtained with this task and the acoustic parameters are described below. In the letter-number sequencing task, random sequences of letters and numbers are presented by the experimenter, from 2 to 8 letter-number combinations. Participants have to remember and repeat them by, first, repeating the numbers in ascending order, then the letters in alphabetical order (e.g., T-9-A-3; correct response is 3-9-A-T).

Raw scores on this WM span task were converted to standard scores (between 1 and 19). Participants were then split into high vs. low WM capacity sub-groups depending on whether their WM score was above vs. below the mean WM score of the group to which they belong to (MS = 9; C = 11). As expected, the mean WM score also differed for high (MS = 6.8; C = 8) and low (MS = 13.2; C = 16) WM speakers across the two groups.

2.3. Acoustic task

A reading corpus was collected, which consists in three sets of sentences composed of Subject-Verb-Object phrases (S-V-O) such as in “Mon frère regarde les jolies lilas dans le parc.” (My brother is looking at the nice lilacs in the park). The O constituent contains a sequence [word1 word2]φ in which word1 is a bisyllabic adjective and word2 is the head noun. In the clash-condition (CL), word2 is monosyllabic and the two primary accents are close to each other (“jolIS LYS”, “nice lilas”); in the no-clash condition (no-CL), word2 is bisyllabic and one unaccented syllable intervene between the two accented ones (“jolIS liLAS”, “nice lilacs”). The two target words were always placed within the same phonological phrase in order to induce speakers to solve clash in the clash condition. The target sentences were interspersed among fillers and presented three times randomly. The acoustic experiment lasted 15 minutes. To sum up, a total of 960 observations were analyzed (20 speakers x 2 groups x 2 clash conditions x 4 sets x 3 times). Both MS and C groups were recorded by means of the EVA2 workstation (Ghio et al., 2012). The acoustic task lasted 15 minutes for each speaker.

2.4. Acoustic annotation

The syllables in word1 and word2 were manually aligned and annotated. The two syllables in the adjective (A) are labeled IA and FA, respectively for their Initial and Final position in the Adjective.

The sequence [word1 word2]φ containing dysfluencies were annotated R and were excluded from the statistical analyses. In total, 6R are found for the control group C (among 4 speakers) and 23R for the MS group (among 7 speakers). R stands for omissions (jolIs -las), hesitations (jolIs heu lilas), lengthening

Figure 1: Pitch track for the clash sequence “nouVEAUX MONTS” (new mountains). Straight lines indicate syllable boundaries.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Gender</th>
<th>Mean age (y.o.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>30</td>
<td>22 women, 8 men</td>
<td>43</td>
</tr>
<tr>
<td>Control</td>
<td>31</td>
<td>22 women, 9 men</td>
<td>43</td>
</tr>
</tbody>
</table>

Table 1: Description of the two groups.
(de nouveaux mmmmmonts), repetitions (jo-jolis lilas), prosodic breaks ([jolis][pause][lilas]) and errors (beaux lilas).

2.5. Acoustic measurements

Acoustic measurements were automatically carried out using the Praat software (Boersma & Weenink, 2014). For each syllable, the syllable duration (ms) and the f0 maximum (maxf0, Hz) were measured. The ratios between IA and FA maxf0 and syllable duration (ratioPitch_IA/FA, ratioDur_IA/FA) were also computed. A ratio equals to 1 indicates similar duration and maxf0 while a ratio > 1 a shorter duration and lower maxf0 for FA and a ratio < 1 a longer duration and higher maxf0 for FA. In no-CL condition, FA is expected to be longer and higher than IA. In CL condition, FA is expected to be shorter and lower than IA. As a consequence, the ratio is expected to be >1 in CL and <1 in no-CL.

2.6 Statistical analyses

A series of linear mixed models were carried out to test the effect of WM capacities (high vs. low), Group (C vs. MS) and Condition (CL vs. no-CL) on the duration and f0 of the word1 last syllable (FA) and on the ratios IA/FA. For f0 analysis, Gender (male vs. female) was also included as the fourth fixed factor. The f0 data were log-transformed to meet normality assumption. Speakers and sentence sets were the random intercepts. Forward selection of random slopes and intercepts based on likelihood-ratio tests was used to decide which random terms should be retained in the models. Non-significant interactions among fixed factors were also excluded based on likelihood-ratio tests. The standard mixed models do not provide p-values when random slopes and intercepts are correlated. We assumed that a fixed factor is significant if its t value is greater than 2 (Baayen et al., 2008).

3. Results

3.1 Duration

Fig. 2 shows the FA duration across group, WM and clash conditions. No significant differences on FA duration between CL and no-CL conditions were found in C group, independent of WM capacities. Similarly, MS speakers with high WM showed no significant differences across the clash conditions. However, for MS speakers with low WM, FA duration is significantly longer in CL (246 ms) than in no-CL conditions (188 ms). [β= -0.014, SE= 0.006, t = 2.325].

Concerning the ratioDur_IA/FA, no significant differences in duration between IA and FA in regards to conditions, groups and WM.

3.2 F0

Our analyses on the FA maxf0 do not show any effect of Group, WM and Condition. The mean F0 maximum was in the upper F0 range for both female (C = 217 Hz; MS = 234 Hz) and male (C = 147 Hz; MS = 100 Hz) speakers, suggesting the presence of a pitch peak on FA. Fig. 3 shows that ratioPitch_IA/FA is around 1 in the CL condition, meaning that the maxf0 had similar value in IA and FA. However, the ratio was significantly lower in no-CL than in CL for both MS and C subjects ([β= -0.07, SE= 0.01, t = -5.33]).

Figure 2: FA duration in regards to group (C vs. MS) and WM capacities (high vs. low).

4. Discussion and conclusion

Our results do not provide evidence of stress clash resolution in the way predicted by Post (1999) and Avanzi & Schwab (2013), both for duration and f0 measurements. First, the duration of FA was similar in both clash and no-clash condition for the control group, suggesting no changes in the degrees of stress. Moreover, contrary to our expectations, FA duration was even longer in the clash than in the no clash condition for the MS group with low WM, indicating specific prosodic strategies adopted by these speakers. The ratioDur_IA/FA also showed no difference across the two clash conditions between the initial and final syllable of the adjective. This further supports the idea that neither stress-shift nor stress deletion took place. Moreover, if clash resolution would have applied at tonal level, the accent on the final syllable of word1 should have been deleted or shifted to the initial syllable. However, no significant differences in maxf0 within FA were found in the clash and no-clash condition. This was true for both the C and MS groups, independent of WM capacities. The ratioPitch_IA/FA showed that, for both groups, the maxf0 has the same value for both IA and FA in the clash condition, whereas FA carried a higher F0 value in the no-clash condition.

The negative results for clash resolution can be explained by taking into account other models of French prosody than Post (1999). Jun & Fougeron (2000) claimed that the lower constituent of French prosody is the Accentual Phrase (AP). The AP is the domain of primary stress and is characterized by an obligatory final rise (LH*), preboundary lengthening and an
optional initial rise (LHi). Phonetically, the H* of the final rise is consistently realized at the end of the last full syllable of the AP. The Hi of the early rise is not anchored to a specific landmark but it can be realized on the first or second content word syllable (Welby, 2006). The L’Hi LH*/H pattern in both clash and no-clash conditions (Fig. 4). Specifically, speakers of the C group and speakers of the MS group with high working memory could have consistently placed an initial Hi on word1 since in both conditions there is no violation of AVOID HHH. Given the ‘loose’ association of Hi with the left edge of the AP, its phonetic target could have been variably aligned with the segmental string. It could have been realized on FA in no-CL condition (which resulted in PitchRatio < 1). However, in the CL condition, its alignment could be slightly earlier, such as at the end of IA/beginning of FA (which resulted in PitchRatio = 1; cf. also Fig.1). That is, the proximity with a following H target would influence only alignment details in the phonetic implementation of Hi (maybe reflecting tonal repulsion; Silverman & Pierrehumbert, 1990). This could explain the high values of maxf0 in FA in both clash and no-clash conditions, as well as the lack of differences for ratioPitch_IA/FA in the CL-condition. Moreover, the lack of durational results can be explained by the fact that the syllable carrying the initial rise is not marked by lengthening.

A different prosodic pattern might have been chosen by MS speakers with low WM. For these speakers the final syllable was lengthened in the CL-condition. A possible hypothesis is that the adjective and the noun have been uttered in two separate APs: [word1]AP [word2]AP. Thus, FA may become the phrase-final syllable and, as such, it may be marked by a final rise and preboundary lengthening. In other words, MS speakers with low WM could have narrowed down the scope of prosodic planning compared to the other speakers because of their reduced cognitive resources. Note also that planning difficulties also resulted in a higher number of dysfluencies in the MS than in the C group. This suggests that the planning of the prosodic unit is flexibly adapted by speakers depending on their cognitive skills (Swets et al. 2007).

In conclusion, the prosodic patterns found in our data are temptatively reanalyzed following Jun & Fougeron (2000)’s model of French intonation. While speakers might have not planned clash resolution, the comparison between healthy and MS speakers provide evidence of individual differences in the scope of the planning unit. A more detailed phonetic study is needed to quantify the amount of speaker-specific behavior in the realization of the prosodic structure and its link with cognitive capacities.

5. References


Changes in breathing activity under different focus conditions
Caterina Petrone¹, Susanne Fuchs² and Laura Koenig³

¹Laboratoire Parole et Langage, CNRS & Aix-Marseille University, Aix-en-Provence
²Zentrum für Allgemeine Sprachwissenschaft, Berlin
³Haskins laboratories & Long Island University

caterina.petrone@lpl-aix.fr, fuchs@zas.gwz-berlin.de, koenig@haskins.yale.edu

Abstract
This work investigates the relationships among between thoracic volume changes, measured via Respitrace, subglottal pressure variation, and sentence stress in ten speakers of German. The plateau of the intraoral pressure (Pio) during productions of /t/ in target words, measured via a pressure transducer on the palate, was taken as an indication of subglottal pressure. Acoustic measures of vowel duration and intensity verified that speakers realized the stress as intended (viz., left-focused, right-focused, or not focused). The Pio data showed that sentential stress patterns were reflected in the Pio peak measures and in the Pio slope in the target #t#t sequence; respitrace measures of thoracic volume did not show reliable differences across focus conditions, perhaps because the method is not sufficiently sensitive.

Keywords: subglottal pressure, intraoral pressure, thorax kinematics, sentence stress.

1. Introduction
Speech production requires more active control of the respiratory system than quiet breathing. For instance, in quiet breathing air expiration is mainly the passive result of the elastic recoil force, while in speech, specific respiratory and laryngeal muscles might be activated during expiration to prevent subglottal pressure (Psub) from decreasing too rapidly over the utterance. Past studies have pointed out that breathing and, in particular, subglottal pressure also affects the production of prosodic parameters. For instance, an increase in subglottal pressure leads to an increase of intensity, which in turn is used, e.g., to signal prosodic prominence (e.g., Ladefoged and McKinney, 1963). An increase in subglottal pressure will also yield a small increase in f0 (Titze, 1989). This paper focuses on the involvement of subglottal pressure in the implementation of stress. Though the belief that stress is implemented with greater respiratory effort is quite old, the physiological mechanisms underlying the realization of stress are still controversial (cf. Ohala, 1990 for a review). Different studies have provided evidence for the existence of local peaks of subglottal pressure co-occurring with the stressed syllables (e.g., Ladefoged, 1968). A possible hypothesis is that such local peaks are the result of the fast contractions of respiratory muscles. Ladefoged and Loeb (2009) specifically claimed that prominent syllables (i.e., stressed at either the lexical or sentence level) are realized with an increase of subglottal pressure due to additional activity of the internal intercostalis muscles. The internal intercostalis are expiratory muscles situated between the ribs of the thorax, deep to the external intercostals. Their contraction would pull the ribs together, thus forcing air out. Ladefoged and Loeb (2009) replicated an old experiment, summarized in Ladefoged (1968), with new methods in order to provide more solid quantitative results.

Specifically, by means of EMG data, they found that peaks of internal intercostals activity occurred right before each lexically stressed syllable of a sentence, and that the peak with largest amplitude was associated with the syllable carrying the sentence stress.

Assessing respiratory contributions to prosodic variation has historically been difficult, due to the invasiveness of the methods which have been adopted to measure subglottal pressure (e.g. tracheal puncture; swallowing a balloon into the oesophagus) and to record the activity of the respiratory muscles (e.g., electromyography). In this paper, we evaluate the relationship between respiratory activity and local prosodic effects using a method which is more comfortable for the subject and allows a larger sample and longer recordings. It is based on the assumption that intraoral pressure (Pio) quickly rises to subglottal pressure when the vocal tract is completely closed and the glottis is open (Löfqvist et al., 1982) as in voiceless aspirated stops. The pressure equalization usually corresponds to a plateau phase in the Pio profile (see Fig. 1). Based on this assumption we have established a new experimental design combining the recording of intraoral pressure (by means of a piezoresistive pressure sensor, see Fuchs & Koenig, 2009 for further details) and respiratory activity (using Respitrace). While Respitrace doesn’t allow us to make predictions about the specific muscles activated during stress production, it provides measurements of the thoracic/abdomen volume changes.

Figure 1: Example of different intervals in the Pio profile (from left to right): Part I: Pressure rise determined by laryngeal-oral coordination, followed by plateau phase where subglottal pressure is equal to intraoral pressure. Part II: pressure drop determined by laryngeal-oral coordination

The relationship between subglottal pressure and stress production was investigated by focusing on utterance-level prominence. That it, we manipulated the position of the most prominent syllable in the sentence, i.e., carrying the sentence stress. In fact, the local pulse of subglottal pressure is expected...
to be higher (and thus, more easily measurable) at this location of the sentence (Ladefoged and Loeb, 2002). To do so, a set of sentences was created, in which the position of the constituent carrying a contrastive focus was varied. Contrastive focus was chosen because previous studies on German have reported an increase of prominence-lending cues in contrastive focus compared to other focus conditions, which has been related to increased articulatory effort (Mücke & Grice, in press). We hypothesized that if sentence stress is achieved via subglottal pressure, we would find traces of higher Pio during voiceless stops, in the focused location. Given a sequence of [word1 word2], we predict that if the focus is realized on the left (word1), the Pio slope would be close to zero or negative, since the pressure of the stop should be higher on the left than on the right side. If focus occurs on the right side (word2), we expect an increase in Pio pressure at this location and steep, positive pressure slope. Concerning the involvement of breathing in this process, we predict increased thoracic volume changes associated with the focused constituent.

2. Methods

2.1. Corpus and participants

The speech material consisted of five sets of German sentences with a monosyllabic verb ending with /t/ (word1) followed by a bisyllabic noun starting with /h/ and lexically stressed on the initial syllable (word2), e.g., “Er malt Tanja, aber nicht Sonja.” (He paints Tanja, but not Sonja). The target words varied across the five sets, but they always contained this /h/+/t/ sequence. The /h/+/t/ sequence was chosen since German speakers usually realize it as one /h/ with a long closure duration, similar to a geminate (Fuchs & Koenig, 2006). As a consequence, it is also expected to correspond to a long Pio plateau. Readings with the desired location of contrastive focus were obtained by setting up a question-answer paradigm where the experimenter read a question and the participant read the corresponding answer making felicitous use of prosody to disambiguate. Three focus conditions were created: (1) focus on the left side [word1], hereafter lf; (2) focus on the right side [word2], hereafter rf and (3) no focus on the [word1 word2] sequence, but on the preceding personal pronoun (e.g., Er), hereafter nf. The no focus condition served as a baseline, since early focus in German requires post-focal deaccentuation (Féry & Kügler, 2008). Ten speakers were recorded: seven women and three men (ranging between 22-36 years old). None of them reported any history of speech, language or hearing difficulties or heavy smoking habits. All subjects spoke Northern Standard German. All together five different target items were used in three different focus conditions. They were repeated four times in a randomized order leading to 600 sentences (5 targets * 3 focus conditions * 4 repetitions * 10 speakers).

2.2. Recording the data

Acoustic, aerodynamic and breathing data were simultaneously recorded to a multi-channel data recording system using a sampling frequency of 11025 Hz. Subjects were standing in front of a music stand where they could read the respective sentences. The microphone was positioned about 30 cm from the mouth. Although mouth-microphone distance was not controlled except to ask speakers to minimize movement during recording, any variation in distance should be randomly distributed over utterance types, since speakers read from fully randomized lists.

Pio data were obtained using a pressure transducer (Endevco 8507C-2) affixed to a small plastic tube glued on to the posterior end of the hard palate. This placement provides Pio data for the alveolar stops used here. The pressure signal was calibrated in steps of 2 cm water column using PCQuirer. Pio data were converted to Matlab and smoothed using the filtfilt function and a kaiser window with 40 Hz passband and 100 Hz stopband edges, and a 50 dB damping factor to eliminate glottal pulses as well as effects of low-frequency electrical interference. From the smoothed signal, the first derivative (velocity) and second derivative (acceleration) were calculated.

Respiratory movements were obtained using Respiratory Inductance Plethysmography. Two elastic bands, one around the rib-cage and one around the abdomen, are used to register volume changes via changes in the electrical resistance of small wires located in the bands. We will here focus on the thoracic movements, since they have been more widely discussed in the literature. Subjects were standing in an upright position and they were asked to read the target sentences without moving their arms or legs to avoid distortions of the breathing signals. The recording took approximately 15-20 minutes.

2.3. Measurements

Based on the acoustic signals, we labeled the following in Praat (Boersma & Weenick, 2014): (1) the duration of the stressed vowels in word1, word2 and on the preceding subject pronoun; (2) the duration of the closure and aspiration for the /h/; (3) the intensity of the vowels in word1 and word 2. The acoustic measurements were made to ensure the correct elicitation of focus. In the Pio data we labeled the onset and offset of the pressure plateau in the /t/ sequence (see Figure 1, area where Pio=Psusb). Both parameters were labeled based on the velocity signal and corresponded to the onset and offset time points where the velocity is close to zero. We calculated the Pio slope between these time points. Furthermore we searched automatically for the pressure maximum in the time between the two vowels.

From the whole database we excluded 7.7% of the data where no clear plateau phase could be defined since speakers either realized two /h’s (left column in Figure 2) or they spirantized the stop (right column in Figure 2).

In the respiratory data, we calculated the slope of thorax movement from the beginning to the end of V1 (the stressed vowel in the word1) and V2 (the stressed vowel in word2).
After visual inspections we did not consider the closure and aspiration part of the /t/ because this may have affected breathing automatically due to glottal aperture.

2.4. Statistical analyses

Linear mixed models as implemented were run in R (R Development Core Team, 2011) with one of the following parameters as dependent variables: a) acoustic data: Vowel duration difference (V2-V1), Mean intensity difference (V2_I-V1_I); b) intraoral pressure data: pressure slope, pressure maximum; c) breathing data: difference in thorax slope (V2_thorax-V1_thorax).

As fixed effects we included focus condition and as random intercepts speaker, word and repetition. A random slope by focus was also included. We assumed that a fixed factor is significant if its $t$ value is greater than 2 (Baayen et al., 2008).

Table 1 gives a summary of the results of the fixed effects. Since in the regression models all pairwise comparisons of the levels in the focus conditions were made with respect to a reference level (the intercept lf), the same models were run again by changing the reference level (nf) in order to obtain different pairwise comparisons.

3. Results

Table 1: Summary statistics of the linear mixed model

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std.Err</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voweldur difference (ms)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept lf)</td>
<td>56.08</td>
<td>29.68</td>
<td>-1.89</td>
</tr>
<tr>
<td>Lf-nf</td>
<td>32.65</td>
<td>11.74</td>
<td>2.78</td>
</tr>
<tr>
<td>Lf-rf</td>
<td>54.51</td>
<td>19.89</td>
<td>2.74</td>
</tr>
<tr>
<td>(Intercept nf)</td>
<td>-23.43</td>
<td>19.35</td>
<td>-1.21</td>
</tr>
<tr>
<td>Nf-rf</td>
<td>21.86</td>
<td>9.54</td>
<td>2.29</td>
</tr>
<tr>
<td>Intensity difference (dB)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept lf)</td>
<td>-5.37</td>
<td>1.02</td>
<td>-5.26</td>
</tr>
<tr>
<td>Lf-nf</td>
<td>2.83</td>
<td>0.63</td>
<td>4.48</td>
</tr>
<tr>
<td>Lf-rf</td>
<td>9.28</td>
<td>1.04</td>
<td>8.94</td>
</tr>
<tr>
<td>(Intercept nf)</td>
<td>-2.55</td>
<td>0.84</td>
<td>-3.0</td>
</tr>
<tr>
<td>Nf-rf</td>
<td>6.45</td>
<td>0.75</td>
<td>8.6</td>
</tr>
<tr>
<td>Pio slope (Pa)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept lf)</td>
<td>0.005</td>
<td>0.019</td>
<td>0.29</td>
</tr>
<tr>
<td>Lf-nf</td>
<td>0.067</td>
<td>0.011</td>
<td>5.87</td>
</tr>
<tr>
<td>Lf-rf</td>
<td>0.136</td>
<td>0.019</td>
<td>6.98</td>
</tr>
<tr>
<td>(Intercept nf)</td>
<td>0.073</td>
<td>0.015</td>
<td>4.75</td>
</tr>
<tr>
<td>Nf-rf</td>
<td>0.069</td>
<td>0.016</td>
<td>4.32</td>
</tr>
<tr>
<td>Pressure maximum (Pa)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept lf)</td>
<td>680</td>
<td>49</td>
<td>13.98</td>
</tr>
<tr>
<td>Lf-nf</td>
<td>-85</td>
<td>22</td>
<td>-3.93</td>
</tr>
<tr>
<td>Lf-rf</td>
<td>12</td>
<td>26</td>
<td>0.47</td>
</tr>
<tr>
<td>(Intercept nf)</td>
<td>595</td>
<td>34</td>
<td>17.69</td>
</tr>
<tr>
<td>Nf-rf</td>
<td>98</td>
<td>21</td>
<td>4.63</td>
</tr>
<tr>
<td>Difference in thorax slope in V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept lf)</td>
<td>0.026</td>
<td>0.035</td>
<td>0.75</td>
</tr>
<tr>
<td>Lf-nf</td>
<td>0.013</td>
<td>0.053</td>
<td>0.24</td>
</tr>
<tr>
<td>Lf-rf</td>
<td>0.059</td>
<td>0.073</td>
<td>0.81</td>
</tr>
<tr>
<td>(Intercept nf)</td>
<td>0.039</td>
<td>0.033</td>
<td>1.16</td>
</tr>
<tr>
<td>Nf-rf</td>
<td>0.046</td>
<td>0.051</td>
<td>0.90</td>
</tr>
</tbody>
</table>

The acoustic results show that speakers realized the vowel under focus with a longer duration in comparison to the other one (i.e., V1 in lf condition and V2 in rf condition). Furthermore, speakers also produce the vowels under focus with a higher mean intensity in comparison to the vowels without focus. Thus, the difference between the first and the second vowel for the vowel under focus was greater than the vowel without focus (see Figure 3).

These acoustic results provide evidence that speakers realized the contrastive focus successfully. What effect did this have on the pressure data?

In Figure 4 and 5, the slope and the maximum intraoral pressure are displayed. The slope parameter varied as a

![Figure 3: Boxplots for the mean intensity difference between V2 and V1 in dB; all data are pooled together.](image)

![Figure 4: Boxplots for pressure slope in the different focus conditions (lf=left focus, nf = no focus, rf = right focus), all data are pooled together.](image)

![Figure 5: Boxplots for pressure maximum in Pa in the different focus conditions, all data are pooled together.](image)
function of the focus condition with a progressive increase of the slope steepness from lf to rf. These results confirm that subglottal pressure differences are involved in the production of contrastive focus. This is also evident in the pressure maxima data, where values are always higher when focus is produced in comparison to the non-focus condition. The involvement of subglottal pressure in the production of contrastive focus could, however, not be found in the breathing data. Here, we found some small effects in certain speakers, but no consistent pattern in general.

4. Discussion and conclusion

The acoustic measurements confirmed the correct elicitation of contrastive focus. Vowels V1 and V2 were significantly louder and longer in the focus (lf, rf) than in the no-focus condition (nf). This is line with the literature, which reported that contrastive focus in German causes an increase in intensity and duration on the focused constituents (Mücke & Grice, in press).

Changes in contrastive focus position had consequences on intraoral pressure. Intraoral pressure was measured in correspondence of the /t/#/t/ sequence, as it is assumed to equal subglottal pressure at this location. The intraoral pressure profile was mostly characterized by a long plateau phase, whose maximum value was higher in the focus (lf, rf) than in the no-focus (nf) conditions. This confirms that sentence stress (as a result of contrastive focus) is achieved via higher peaks of subglottal pressure realized in the focused constituents (Ohala, 1990).

Moreover, although the /t/#/t/ sequence was usually produced as a single [t], the intraoral pressure plateau has different slopes depending on the position of the focus. The slope difference reflects a shift in the location of the maximum value (Pio_max) within the plateau. Specifically, in rf, the pressure maximum is located at the offset of the plateau as word2 is under focus. As a consequence, the plateau assumes a steep raising slope. In the nf condition, neither word1 nor word2 are under focus, and the plateau has only a slightly raising slope, with lower values than in the rf condition. In lf, the plateau is flatter (or even slightly falling) but it has much higher values than in the nf condition. This means that, in connected speech, subtle differences in the plateau realization (height, slope) convey different information about the involvement of subglottal pressure in sentence stress realization. While the height of the pressure maximum would indicate the presence/absence of sentence stress, the slope measurements in rf also suggest that the subglottal pressure quickly reaches a maximum within the focused word and decay after the implementation of sentence stress.

The analysis of thoracic data, on the other hand, revealed that changes in contrastive focus position had consequences on subglottal pressure. Intraoral pressure was measured in intraoral pressure. Intraoral pressure was measured in

in rf also suggest that the subglottal pressure quickly reaches a maximum within the focused word and decay after the implementation of sentence stress. The analysis of thoracic data, on the other hand, revealed that they are not correlated with focus. There are at least two possible explanations for this null finding. First, sentence stress could be implemented through physiological mechanisms other than the activity of respiratory muscles. For instance, it has been suggested that subglottal pressure changes are affected by glottal resistance to airflow (Ohala, 1990 and references therein; cf. also Demolin, 2010). Moreover, the respiratory data we obtained cannot be considered as a direct mirror of the activity of specific respiratory muscles (such as the internal intercostals) which have been associated with stress production. Since Respirtrac gives us only the information of thoracic volume changes, the coordinated contraction of different muscles could have affected the breathing data in an unpredictable way. That is, we cannot discount a role for specific respiratory muscles in subglottal pressure control whose activity has not been directly tracked by the inductance plethysmograph.

In conclusion, the relationship between subglottal pressure and stress has been investigated through a new methodology, involving the simultaneous recordings of intraoral pressure and breathing data. Intraoral pressure data suggest that changes in contrastive focus position involve a rise of subglottal pressure on the syllable carrying the sentence stress. This supports the idea that breathing control can be exerted at a local level. More sophisticated measurements are needed to better investigate the role of specific respiratory muscles in stress production.

5. Acknowledgements

This work was supported by a grant from the Ministry of Education and Research (01UG1411) to the second author and a BQR grant from Laboratoire Parole et Langage to the first one. Thanks to Joerg Dreyer for technical support.

6. References


Acoustic correlates of prosodic marking in spontaneous self-repair in Dutch

Leendert Plug
University of Leeds, United Kingdom

l.plug@leeds.ac.uk

Abstract

This paper reports on a phonetic analysis of instances of lexical and phonological self-repair sampled from Dutch spontaneous speech. The focus is on the relationship between f0, intensity and articulation rate characteristics of the repairs on the one hand, and their perceived status as ‘prosodically marked’ or ‘unmarked’ on the other. The concept of prosodic marking was first introduced in psycholinguistic work on self-repair, and its phonetic correlates have remained unclear, although available descriptions suggest they may be different from those of other forms of prominence marking, in particular ‘hyper-articulation’. This paper confirms that prosodic marking in self-repair is a form of prominence marking whose phonetic features warrant its distinctive label.

Keywords: self-repair, prosody, hyper-articulation

1. Introduction

This paper reports on a phonetic analysis of instances of self-repair, including lexical repairs such as *Thursd- uh Friday* and phonological repairs such as *expela- explanation*, sampled from Dutch spontaneous speech. The focus of the analysis is on the relationship between f0, intensity and articulation rate characteristics of a repair on the one hand, and its perceived status as prosodically ‘marked’ or ‘unmarked’ on the other.

The concept of prosodic marking in self-repair was first introduced by Cutler (1983) and Levelt & Cutler (1983), as a terminological improvement on Goffman’s (1981) distinction between ‘strident’ and ‘flat’ repairs. Cutler (1983) describes an ‘unmarked’ repair as one in which the pitch, intensity and speaking rate of the repair component — in the case of lexical repair, the second lexical item; in the case of phonological repair, the correct pronunciation of the target word — are not noticeably different from those of the reparandum — the first lexical item, or the erroneous target word attempt. A ‘marked’ repair, on the other hand, ‘is distinguished by a quite different prosodic shape from that of the original utterance’ (Cutler 1983: 81). By leaving a repair unmarked, the speaker ‘minimises the disruptive effect of the error on the utterance as a whole’, while marking assigns ‘salience’ to the correction (Cutler 1983: 80).

Cutler’s (1983) description of marked repairs suggests that prosodic marking in self-repair can be achieved in a variety of ways. Levelt & Cutler (1983: 206) suggest the same, stating that it can be implemented through ‘a noticeable increase or decrease in pitch, in amplitude, or in relative duration’. Interestingly, this is somewhat at odds with Goffman’s description of ‘strident’ repairs: according to Goffman, these involve raised pitch, loudness and tempo — not the opposite. Unfortunately, none of these sources present results of systematic acoustic analysis.

The acoustic correlates of prosodic marking in self-repair are of interest because a pertinent question from the perspective of speech production modelling is whether prosodic marking is any different from other forms of prominence marking — or in other words, whether its distinctive label is justified. If Goffman (1981) is right, it probably does, as pitch and intensity raising generally go together with temporal expansion, not compression. If Levelt & Cutler (1983) are right and prosodic marking can be achieved through, for example, a noticeable drop in intensity alone, it probably does too. If, on the other hand, most instances perceived as prosodically marked have a repair component produced at a higher pitch and intensity and a lower tempo than the reparandum, ‘prosodic marking’ would seem a redundant synonym of ‘hyper-articulation’ (Lindblom 1996).

More recent studies into the phonetics of self-repair have shown that an increase in pitch and intensity between reparandum and repair is the norm (Howell & Young 1991, Nakatani & Hirschberg 1994, Cole et al. 2005), as is an increase in articulation rate (Plug 2011). However, these studies present results of acoustic analysis only, without considering which of the repairs in their data sets sound prosodically marked. The current paper represents a first attempt to establish how prosodic marking is implemented in spontaneously produced self-repairs.

2. Data and method

2.1. Data

The data for this paper comprise 580 instances of lexical (N=214) and phonological (N=366) error repair extracted from four sub-corpora of the Spoken Dutch Corpus containing spontaneous speech. We extracted instances of speech which were coded as mispronounced or interrupted and did a number of additional, unsystematic data trawls. We only included lexical repairs in which one word was retroactively replaced by another, and phonological repairs containing at least one consonant and one vowel with primary or secondary lexical stress. Representative examples include *met de au- met de bus* ‘by ca- by bus’, *een leuke k- een mooie keuken* ‘a nice k- a beautiful kitchen’, *[b]aarbij – [w]aarbij ‘with which’, and *vana[1] de – vana[1] de ‘from the’. We left aside repairs occurring in utterance-initial and utterance-final positions in order to minimize the possible effect of prosodic boundary marking.

2.2. Acoustic analysis

We segmented all instances, placing boundaries at the start and end of the crucial lexical items in the reparandum and repair — the word that is subsequently replaced and its replacement in the case of lexical repair; the two attempts at the target word in the case of phonological repair. We delimited all vowel portions within these domains.

We measured f0 (in Hertz) and intensity (in decibels) at every millisecond across the segmented vowel portions, and log-transformed f0 values. We then calculated mean, median and maximum values. In each case we calculated a delta value by...
subtracting the value derived from the reparandum from that derived from the repair. This yields a measure of the prosodic difference between the crucial components of the repair, as well as introducing some speaker normalization. Analysis not reported in detail here (but see Plug & Carter 2013) revealed that the mean, median and maximum delta values are tightly correlated, with the maximum delta values most informative in subsequent modeling. In what follows, we therefore restrict our attention to the maximum delta values, which we will refer to by their variable names, F0 max delta and Intensity max delta.

In addition, we calculated the articulation rate for each segmented portion by dividing the number of surface segments articulated during the portion by its raw duration. We square-root-transformed rate values to normalize their distribution, and calculated a delta value for each instance by subtracting the (transformed) value derived from the reparandum from that derived from the repair.

2.3. Auditory analysis

Following Levelt & Cutler (1983), we classified all instances as prosodically marked or unmarked based on auditory analysis. The question in each case was whether the correct target word realization sounds particularly salient because of its prosody, relative to the erroneous attempt. We allowed for the intermediate classification of ‘possibly marked’ (see Plug & Carter 2013).

The classification was done by two raters: the second author and a Dutch discourse analyst with no particular knowledge of the phonetics of self-repair. The two raters classified all instances independently. They reached the same judgment in 250 cases (77%). Of the 75 instances for which the raters proposed a different classification, 24 involved one rater proposing ‘possibly marked’ and the other ‘marked’. In order not to overestimate the proportion of prosodically marked repairs, we coded these instances as ‘possibly marked’. The remaining 51 instances either involved ‘possibly marked’ vs ‘unmarked’ or ‘marked’ vs ‘unmarked’. All of these instances were reconsidered independently by both raters. In nine cases, this resulted in straightforward agreement, while in 42, the raters confirmed their initial judgments. Remaining cases of ‘possibly marked’ vs ‘unmarked’ were coded as ‘unmarked’; cases of ‘marked’ vs ‘unmarked’ as ‘possibly marked’. In the final coding, 385 instances (66%) are ‘unmarked’, 81 (14%) ‘possibly marked’ and 114 (20%) ‘marked’.

In what follows, we will refer to the marking classification by its variable name, Prosodic marking. For the purpose of the quantitative analysis reported here, this variable was transformed into a binary one. Exploratory modeling suggested that collapsing ‘possibly marked’ and ‘marked’ results in a better fit with the acoustic measurements than collapsing ‘possibly marked’ and ‘unmarked’. We will therefore report on the former.

2.4. Quantitative analysis

We investigated the relationship between the prosodic marking judgments and the acoustic measurements using cluster analysis (run in SPSS) and mixed effects regression modeling (using the lme4 package in R). For the purpose of modeling, we transformed the delta values into Z-scores. We incorporated Speaker as a random effect where relevant. Exploratory modeling showed that Repair type — lexical or phonological — has no explanatory value, so this variable is not further discussed here.

3. Results

3.1. Correlations among acoustic parameters

Figure 1 provides a visual impression of the relationship between the three acoustic parameters. Intensity max delta is significantly correlated with both F0 max delta (unstandardized: Spearman’s ρ=−0.37, p<0.001) and Rate delta (ρ=−0.10, p=0.02). F0 max delta and Rate delta are not significantly correlated (ρ=−0.04, p=0.38). The correlation between Intensity max delta and F0 max delta is positive, while that between Intensity max delta and Rate delta is negative: the higher the intensity of a repair compared with its reparandum, the higher its f0, but the lower its articulation rate. Delta values are more often positive than negative across the three parameters (Intensity max delta 65%>0, F0 max delta 67%>0, Rate delta 70%>0), and 32% of instances have positive delta values only. Instances with negative values only are rare (4%), and instances with positive and negative deltas tend to show intensity and f0 clustering together against rate.

![Figure 1: Scattergrams of Intensity max delta against F0 max delta (top) and Rate delta (bottom). Dotted lines mark delta values of 0. Grey points represent unmarked instances; white points marked.](image)

3.2. Modeling the marking judgments

Looking now at the distribution of grey and white data points in the scattergrams in Figure 1, representing prosodically marked and unmarked repairs respectively, it seems clear that there are systematic relationships between the three acoustic parameters and the marking judgments. In the top scatter, white points predominantly cluster in the top right corner, while grey points have a wider spread. This means that prosodically marked instances are more strongly associated with positive values for Intensity max delta and F0 max delta than unmarked ones. In the bottom scatter, white data points cluster on the right accordingly, and appear to lean somewhat towards the bottom. The latter means that prosodically marked instances are associated with lower values for Rate delta than unmarked ones, although most are positive. These
observations are confirmed by the boxplots in Figure 2 and associated unpaired comparisons: marked instances have significantly higher values for Intensity max delta (standardized: t(578)=9.40, p<0.001) and F0 max delta (t(578)=6.58, p<0.001), and significantly lower values for Rate delta (t(578)=−3.61, p<0.001).

Figure 2: Boxplots for Intensity max delta, F0 max delta and Rate delta with Prosodic marking as grouping variable (‘no’ meaning unmarked; ‘yes’ marked).

To gain further insight into the predictive value of the three acoustic parameters, we built a mixed effects regression model with Prosodic marking as target variable, (standardized) Intensity max delta, F0 max delta and Rate delta as predictors and Speaker as a random factor. The model is shown in Table 1. The strongest single predictor was Intensity max delta. Given its significant correlations with the other two predictors, we orthogonalized the latter by replacing them with the residuals of simple linear models predicting their values on the basis of Intensity max delta. The model shows main effects consistent with the results of the paired comparisons cited above, and confirms that despite the significant correlations among the three acoustic parameters, each has predictive value of its own. In addition, a three-way interaction improves the model fit.

Table 1: Mixed effects regression model with Prosodic marking as target variable and (standardized and orthogonalized) Intensity max delta, F0 max delta and Rate delta as predictors.

| Factor | Est.  | SE   | z     | p>|z| |
|--------|-------|------|-------|--------|
| Intercept | −0.93 | 0.11 | −8.33 | <0.001 |
| Intensity max delta | 1.04 | 0.13 | 8.15 | <0.001 |
| F0 max delta | 0.43 | 0.11 | 3.99 | <0.001 |
| Rate delta | −0.33 | 0.11 | −3.08 | 0.002 |
| Intensity−F0−Rate | −0.28 | 0.12 | −2.38 | 0.017 |

The model is visualized in Figure 3 in the form of an conditional inference regression tree (Tagliamonte & Baayen 2012). The tree algorithm establishes which subdivisions in the data provide the most homogeneous groupings of observations with respect to a target variable — in this case Prosodic marking. Partitioning is recursive, so that predictors can feature more than once in the tree.

The tree in Figure 3 reflects the strength of Intensity max delta as a predictor: it yields three splits at the top of the tree (Nodes 1, 2 and 7), dividing the data into instances with substantial negative deltas, of which less than 10% are perceived as marked (Node 3), instances with high positive deltas, of which over 60% are marked (Node 11), and two sets of instances with intermediate delta values which are further split according to their values for F0 max delta (Node 4) and Rate delta (Node 8). Both F0 max delta and Rate delta split subsets of instances into two further subsets: one with around 20% and one with around 50% perceived as marked (Nodes 5 and 6, and 9 and 10). For F0 max delta, the subset with around 50% perceived as marked is associated with relatively high, positive values (Node 6); for Rate delta it is associated with relatively low ones, including negatives (Node 9). The asymmetry in the tree, with F0 max delta occurring in the left half only and Rate delta in the right half, is consistent with the significant three-way interaction in the linear model in Table 1.

Figure 3: Conditional inference regression tree with Prosodic marking as target variable and Intensity max delta, F0 max delta and Rate delta as predictors.

3.3. Cluster analysis

The findings presented so far clearly confirm that intensity, f0 and articulation rate are systematically manipulated in the production of prosodic marking in self-repair. Evidence for acoustically distinct subtypes of marking, as suggested by Levetl & Cutler’s (1983) definition, is less apparent: there is no obvious cloud separation in Figure 1, or evidence of multimodality in the distributions in Figure 2. Still, the regression tree in Figure 3 reflects some clustering among instances perceived as marked: these can have very high values for Intensity max delta, or intermediate values combined with relatively high values for F0 max delta, or low values for Rate delta. In order to assess the significance of this clustering, we carried out a cluster analysis on the standardized delta values. We first performed an unconstrained hierarchical cluster analysis using Ward’s method. Inspection of the agglomeration schedule suggested that the most informative model should contain three to seven clusters. Subsequent constrained fitting was done through k-means clustering (Zellners & Ogden 2012). We report the model with five clusters, as it provides the best fit with the marking judgments. It also closely matches the regression tree in Figure 3. The model is summarized in Table 2.

Table 2: Results of a k-means cluster analysis with (standardized) deltas as input, plus frequencies for Prosodic marking. Cluster centres are marked by distance from the mean for the relevant parameter: ‘++’ >1 SD above, ‘−−’ <0.5 SD above or below, ‘−−‘ <0.5 below, ‘−−‘ <1 SD below.

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity max delta</td>
<td>++</td>
<td>...</td>
<td>...</td>
<td>--</td>
<td>...</td>
</tr>
<tr>
<td>F0 max delta</td>
<td>...</td>
<td>++</td>
<td>...</td>
<td>--</td>
<td>++</td>
</tr>
<tr>
<td>Rate delta</td>
<td>--</td>
<td>...</td>
<td>...</td>
<td>--</td>
<td>++</td>
</tr>
<tr>
<td>N</td>
<td>75</td>
<td>65</td>
<td>168</td>
<td>127</td>
<td>145</td>
</tr>
<tr>
<td>% marked</td>
<td>59%</td>
<td>52%</td>
<td>36%</td>
<td>12%</td>
<td>29%</td>
</tr>
<tr>
<td>% out of total marked</td>
<td>23%</td>
<td>17%</td>
<td>31%</td>
<td>8%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 2 shows that in two of the five clusters (Clusters 1 and 2), more than 50% of instances are perceived as marked. Instances in these clusters have high values for either Intensity max delta or F0 max delta, and either low or average values for Rate delta. (Recall that on average, Rate delta is positive; in Cluster 1, its mean is just below zero.) Over a third of
instances in Cluster 3 are perceived as marked, and this cluster accounts for the largest proportion of marked instances overall. The prosody of these instances shows little difference between reparandum and repair in intensity and f0, but a relatively low Rate delta average, just below 0. Cluster 5 is the same except that Rate delta is high: instances in this cluster are predominantly characterized by an increase in articulation rate between reparandum and repair. Nearly a third of instances in this cluster are perceived as marked. Cluster 4 has the lowest frequency of marked instances, and accounts for less than a tenth of marked instances overall. Instances in this cluster are characterized by a decrease in intensity and f0 between reparandum and repair, and little difference in articulation rate.

To assess the predictive value of cluster membership, we entered it as a predictor in a mixed effects model with random factor Speaker. This confirms that cluster membership is a significant predictor (Est=−0.44, SE=0.07, z=−5.99, p(>|z|)<0.001). However, log-likelihood comparison shows that its impact on model fit is considerably weaker than that of Intensity max delta alone, and not significantly stronger than that of F0 max delta or Rate delta. In other words, the observed data clustering is not strong enough to invalidate the linear model in Table 1, and Intensity max delta remains the strongest predictor of marking judgments.

4. Discussion

This paper has reported on an attempt to establish the acoustic correlates of prosodic marking in self-repair, using spontaneous repairs drawn from Dutch speech. Cutler’s (1983) and Levetl & Cutler’s (1983) descriptions suggest that in prosodic marking, speakers do not aim to produce a repair with a particular set of prosodic characteristics: rather, speakers can produce the repair in a number of ways, as long as it is noticeably different from the reparandum on any one or more prosodic parameters. Goffman’s earlier description of ‘strident’ repairs suggests that these do have a recurrent overall prosodic shape, combining raised pitch, loudness and tempo.

The findings presented above are consistent to some extent with both descriptions. The implementation of prosodic marking in our data is less variable than Cutler’s (1983) and Levetl & Cutler’s (1983) descriptions might suggest: the prosodic marking judgments are best modeled linearly, with each of intensity, pitch and tempo contributing predictive value through predominantly unimodal relationships with prosodic marking. A clear majority of marked instances have a repair component with a higher maximum f0, a higher maximum intensity and a higher articulation rate than the reparandum. Still, while marking through a noticeable drop in maximum f0 or intensity is rare, it does occur, and so does marking through a noticeable drop in articulation rate.

In relation to Goffman’s (1981) description, it is worth pointing out that tempo raising is not a feature that distinguishes marked from unmarked instances: in unmarked instances, too, the norm is for a repair to be produced at a higher articulation rate than its reparandum. In fact, the average tempo increase is smaller for prosodically marked repairs than for unmarked ones. This means that while prosodically marked repairs are mostly temporally compressed relative to their reparanda, they are temporally expanded relative to unmarked repairs. The latter makes prosodic marking similar to ‘hyper-articulation’ (Lindblom 1996), which combines pitch and intensity raising with temporal expansion. Still, if our findings are representative, it would be inappropriate to equate ‘prosodic marking’ to variation along a ‘hypo-hyper continuum’: a typical prosodically marked repair has a repair component that is higher in pitch, intensity and tempo than the reparandum — and is therefore neither locally hypo-articulated, nor locally hyper-articulated.

Why should ‘prosodic marking’ in self-repair be distinct from ‘hyper-articulation’ in this way? It is worth noting that a prevalence for speeding up at the repair tallies well with the observation in conversation-analytic work that speakers tend to initiate self-repair quickly, while delaying other-repair (see Plug 2011). Seen in this light, speeding up is consistent with a drive by speakers to get self-repair work done as soon as possible, and return to normal fluency. If a drive to allow for a soonest possible resumption of post-repair speech constrains the production of most instances of self-repair, including those that speakers do not aim to attach particular salience to, then it is arguably not surprising that the temporal characteristics of instances that are marked as salient are distinct from those of speech marked as salient outside of the context of self-repair. The upshot of this line of argument is that ‘prosodic marking’ in self-repair can be seen as a form of ‘hyper-articulation’ — but its result is not a ‘hyper-articulated’ repair as such, as the latter’s characteristic of slow articulation is moderated by a drive to get the repair done as quickly as possible. Whether this account is on the right lines or not, it should be clear that prosodic marking in self-repair is an intriguing form of prominence marking whose phonetic correlates warrant its unique label.

5. Acknowledgements

This work was supported by ESRC grant RES-061-25-0417. I would like to thank Paul Carter and Christina Englert for their contributions to the research reported here.

6. References


Perceptual and articulatory factors in German fricative assimilation

Marianne Pouplier¹, Philip Hoole¹, Jona Cederbaum², Sonja Greven², Manfred Pastätter¹

¹Institute of Phonetics and Speech Processing, LMU Munich, Germany
²Department of Statistics, LMU Munich, Germany

pouplier|hoole|manfred@phonetik.uni-muenchen.de
Jona.Cederbaum|sonja.greven@stat.uni-muenchen.de

Abstract

We present data from an EPG experiment on German fricative assimilation. It has been claimed that fricative sequences other than sibilant clusters do not assimilate due to perceptual constraints. We demonstrate that /f/sibilant sequences show in principle the same kind of temporal overlap as sibilant clusters do, but due to the labial constriction dominating the acoustics, this temporal overlap is acoustically and perceptually less salient. Our data further reveal an order asymmetry: sibilant/f behave differently to /f/sibilant clusters; there is no evidence in sibilant/f clusters for the labio-dental constriction overlapping the sibilant in our data. We consider perceptual and articulatory accounts of this asymmetry. We also investigate whether lexical stress affects assimilation patterns. To that effect we discuss a new statistical method for analyzing functional data with a mixed model allowing for multiple covariates and crossed random effects. We find that primarily stress of the word-final but not the word-initial syllable interacts with assimilation.

Keywords: fricative assimilation, German, EPG, acoustics, overlap, perception, mixed models, functional data

1. Introduction

Assimilation in fluent speech continues to be a main topic of research in the speech sciences since a complex interaction of articulatory, perceptual, and grammatical aspects conditions assimilatory patterns. Our current focus is on German fricative assimilation since this area is particularly well-suited to gauge the contribution of each of these factors to word-boundary assimilation. Fricatives have been hypothesized to be generally resistant to assimilation. For instance, in German /f#ʃ/ or /f#s/ will not assimilate to a sibilant sequence /ʃʃ/ or /ss/. This has been ascribed to perceptual constraints on assimilation: Final fricatives are perceptually salient therefore blocking assimilation, with the premise being that assimilation preferably occurs in perceptually low-salient environments (Byrd, 1996; Hura, Lindblom, & Diehl, 1992; Kohler, 1990).

Thereby sibilants are an exception among the fricatives: an alveolar sibilant assimilates regressively to a following palatal sibilant (e.g., aus Schalke (‘from Schalke’) => auʃalke, but: auf Schalke => auʃalke; not *aufʃalke). Here reduction of articulatory complexity seems to override perceptual constraints, yet it is difficult to explain the difference between /f#ʃ/ and /f#s/ sequences. Final labio-dental fricatives are arguably not very salient (Miller & Nicely, 1955), and thus other factors may be at play in assimilating /s#ʃ/ and apparently non-assimilating /f#ʃ/ sequences. To what extent final fricatives other than sibilants assimilate has, to our knowledge, not been tested systematically. In the present study, we pursue the possibility that /f#s/ and /f#ʃ/ fricative sequences may show articulatory overlap just like /s#ʃ/ sequences, yet due to the difference in articulators involved there may be few acoustic and auditory consequences. The absence of perceived assimilation in such /f+sibilant sequences might then simply be due to nonlinearities in the articulator-acoustic relationship rather than perceptual constraints on cross-word boundary coordination of fricatives. To test our hypothesis, we recorded electropalatography (EPG) and acoustic data for German with a variety of abutting fricatives. In order to investigate the further contribution of articulatory and perceptual factors to assimilation patterns, we also included two vowel and four stress conditions. Particularly stress should contribute to the relative perceptual and articulatory salience of a consonant. We investigate the possibility that in a C1#C2 sequence, an unstressed C1 should more likely be subject to assimilation whereas unstressed C2 should be less likely to trigger assimilation.

2. Methods

We recorded synchronized acoustic EPG data from 9 native speakers of German. All of the speakers were colleagues at the Institute of Phonetics, but naive as to the purposes of the experiment except for S1, the first author of this paper. EPG data were sampled at 200 Hz, acoustic data at 32768 Hz. Subjects were given plenty of time to practice with their palate and an accommodation phase preceded the actual recording.

2.1. Stimuli and experimental procedure

Stimuli consisted of noun-noun compound phrases embedded in a neutral carrier sentence. The stimuli combined the fricatives /ʃ, s, f/ as C1#C2 sequences in final and initial position rendering three cluster conditions: /s#ʃ, ʃ#s, s#f, f#s/; and /s#ʃ, ʃ#s, f#f/. Homorganic combinations served as controls (/ʃʃ, ʃʃ, ʃʃ/). Note that in Standard High German, the initial alveolar sibilant is voiced, but voiceless in the Southern dialectal regions. Since all of our subjects realized the fricative as voiceless regardless of their dialectal background, we uniformly use the symbol /ʃ/ for the alveolar sibilant here. In the following, we will refer to any analyses involving combinations of /s/ and /ʃ/ in either order as sibilant condition (/ʃʃ, s#ʃ), and combinations of sibilant and /f/ in either order as f-condition (/ʃf, ʃ#f, f#ʃ, f#f/). All target words were bisyllabic; we will refer to the syllables containing the fricatives as final and initial target syllable. Two different vowel contexts (i-a, a-i) were included. In the i-a condition, the vowel preceding the cluster was /i/ or /j/ and the vowel following the cluster was /a/ or /a/ (e.g., [‘ku.ʃım.məl], and correspondingly for the a-i condition (e.g., [ku.ʃım.məl]). Lexical stress of the target syllables was varied to be either stressed or unstressed, which we will abbreviate here as S (strong) and W (weak), rendering four stress conditions (SW, WS, SS, WW). The experimental variables were fully crossed. Stimuli were presented 5 times in blocks randomized differently per subject per block, while ensuring that there were no immediately consecutive trials of identical clusters in different stress conditions.
Subjects read the target sentences as presented to them on the screen at a self-chosen rate. They were familiarized with the sentences ahead of the experiment and completed a practice round containing all stimuli wearing their EPG palate. Targeted token total amounted to 9 consonant sequence x 2 vowel x 4 stress conditions x 5 repetitions x 9 subjects = 3240 tokens. Due to a coding error, S1 had no data for the sibilant control condition ([θf]/ , i-a, strong-strong, with the consequence that experimental conditions associated with these controls had to be excluded (i-a, strong-strong, [θf, fθ], sθf, fθs)). Across subjects and conditions another 8 tokens were missing due to technical failure, leaving a total of 3202 tokens for analysis.

2.2. Data treatment

The acoustic data were downsampled to 24kHz and Thomson multi-taper spectra were computed with a 21.3ms window length, 75% overlap (Thomson, 1982). For each trial, the intervocalic fricative interval was segmented acoustically and scaled onto a time interval of [0, 1]. We used the 25% time point for statistical analyses. Both EPG and acoustic data were normalized following Pouplier et al. (2011): The temporal midpoint of the control condition served to created a reference pattern for each consonant relative to which all samples were normalized such that the data ranged between -1 and 1. We further corrected for inherent differences in how well the reference pattern mapped onto the control conditions.

For the sibilant condition, a value of 1 indicates close proximity to the /s/ reference pattern, while a value of -1 indicates close proximity to the /ʃ/ reference pattern. For the f-conditions, a value of 1 indicates close proximity to the /f/ reference pattern, while a value of -1 stands for reference /s/ or reference /ʃ/, respectively, depending on condition.

3. Results

3.1. Overall results

Our first step at data evaluation was to investigate the presence of assimilation for the sibilant and f-conditions overall, collapsing across all vowel and stress conditions. In order to identify whether assimilation had occurred at the 25% time point of the fricative interval, a classification algorithm in the form of a support vector machine (svm; Baayen (2008)) was trained in R on the extracted parameters of the control data (closed test). We then tested the homorganic conditions against the controls. For example, for the sibilant cluster condition, the svm was trained on /θs/θ/ and /θf/ θ control sequences. For a given experimental /θθθ/ token, at the 25% time point of the fricative interval, the svm algorithm should classify the token as /θ/ if regressive assimilated, otherwise the token should be classified as /s/. The training/classification was performed for each of the three cluster conditions (iθθθ/, sθθθ/, θθθθθ) separately, once for the EPG and once for the acoustic data. Table 1 gives the results of the svm classification for all conditions, for both EPG and acoustic data. For the sibilant condition, the table gives percent classified as /s/; a percentage of 100 means that all tokens were classified as /s/, a percentage of zero means that all tokens were classified as /θ/. The homographic control conditions show excellent classification accuracy for both signal types (Table 1 cells A1-2; B1-2). For heterorganic sibilant sequences (Table 1 A3-4, B3-4), the pattern corresponds to what has been found in recent studies for English (Niebuhr, Clayards, Meunier, & Lancia, 2011; Pouplier et al., 2011): there is a strong tendency for regressive assimilation for /θθθ/ (only 57 and 58% percent of data are classified as /s/ at the 25% time point, Table 1 cells A3, B3). For /θθθ/ clusters, there is some influence of /s/ on the palatal sibilant, yet only in about 10% of cases (9%, 12% in Table 1, cells A4, B4). Importantly, the results for the EPG and acoustic data are in close correspondence, the assimilation is evident in both the articulatory and acoustic domain.

We now turn to the θθθθθ sequences. For all sequences, Table 1 gives the percent classified as /θ/. First we take note of the classification accuracy for the controls for both cluster conditions and both signal types (Table 1, rows C-F, columns 1-2). For the heterorganic conditions, first consider /θθθ/. The acoustic data (Table 1, C3) show a comparatively small tendency for assimilation to occur with 77% of tokens having been classified as /θ/, the remaining 23% as /ʃ/. Yet there is a marked discrepancy between acoustic and EPG data (Table 1, D3): For the latter, only 47% of tokens were classified as /θ/ at the 25% time point, i.e. 53% were classified as /ʃ/. This assimilation rate is comparable to the one of /θθθθθ/ sequences (58% classification as /s/, 42% as /ʃ/, cell B3). We interpret this discrepancy to the effect that while the /θ/ constriction is being formed behind the labio-dental constriction, the anterior labio-dental constriction dominates the acoustics. We now turn to the opposite order /θθθθθ/ to find a third pattern of results. Note that there is no assimilation whatsoever in our data: the percentage of tokens classified as /θ/ is on the same scale as the control conditions (Table 1, rows C-F, column 4); there is close agreement between acoustic and EPG classification results.

Table 1: Svm classification results for acoustic and EPG data for all cluster conditions.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td><strong>I</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>homorg.</td>
<td>100</td>
<td>57</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heterog.</td>
<td>100</td>
<td>58</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sθθθ</td>
<td>100</td>
<td>0</td>
<td>57</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fθθθ</td>
<td>100</td>
<td>0</td>
<td>58</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acoustic</td>
<td>100</td>
<td>0</td>
<td>57</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPG</td>
<td>100</td>
<td>0</td>
<td>58</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acoustic</td>
<td>100</td>
<td>0</td>
<td>77</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPG</td>
<td>99</td>
<td>0</td>
<td>47</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>E</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acoustic</td>
<td>99</td>
<td>0</td>
<td>84</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPG</td>
<td>100</td>
<td>0</td>
<td>53</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2. Evaluating the data using mixed modelling for functional data

We included two vowel and four lexical stress conditions in order to investigate whether assimilation would be more likely in certain vowel contexts and stress conditions. Evaluating such a complex experimental design with multiple covariates and crossed random effects statistically has been a longstanding problem in the speech sciences, particularly when aiming to take into account the articulatory/acoustic dynamics throughout the entire consonant interval. We present a new approach to functional data using mixed models based on an extension of Greven et al. (2010). Due to space limitations, we demonstrate for the sibilant cluster condition only (acoustic data) how the effects of order, vowel, and stress can be assessed within a single model with crossed random effects.
discussed now evaluates the evolution of the acoustic index over the entire fricative interval for each token. Only the heterogenic conditions are used for statistical modelling; the experimental conditions are evaluated relative to each other (N=707). In contrast to other functional data analysis methods (e.g., Ramsay & Silverman, 2005), the model is able to accommodate irregularly spaced functional data, i.e. while the time scale of each index curve was scaled to a $[0, 1]$ interval, the data were not resampled to an across-tokens regular grid.

The stress conditions were dummy coded into two covariates: Stress1 (stress of the final target syllable (C1): 0 strong, 1 weak) and Stress2 (stress of the initial target syllable (C2): 0 strong, 1 weak). The other dummy coded covariates were Consonant Order (0: /s#ʃ/, 1: /ʃ#s/), and Vowel (0: i-a, 1: a-i).

Recall that we calculated the index values such that they range between -1 and 1 with -1 being a reference /ʃ/ and 1 denoting a reference /s/ acoustic pattern. In order to be able to compare the index time series for the two consonant orders (/s#ʃ/, /ʃ#s/) directly, the index time series of /ʃ#s/ was mirrored along the time axis such that both C1#C2 conditions showed in principle an index dynamic ranging from 1 for C1 to -1 for C2. Figure 1 shows the resulting index curves for the two consonant orders by speaker, across vowel and stress conditions.

The model used to analyze the data is given in (1):

$$Y_{ijht} = \mu(t, x_j) + B_i(t) + C_j(t) + E_{ijht} + \epsilon_{ijht}$$  \hspace{1cm} (1)

with $Y_{ijht}$ being the index over time for speaker $i$, item $j$, and repetition $h$ observed at time $t \in T_{\subseteq [0, 1]}$, $\mu(t, x_j)$ is a curve specific smooth mean function, $x_j$ is known covariates and possible interactions of covariates. $B_i(t)$ and $C_j(t)$ are random functional intercepts for speaker and items, respectively. $E_{ijht}(t)$ is a speaker-, item-, and repetition-specific smooth random deviation and also includes the interaction between speaker and item. $\epsilon_{ijht}$ is white noise measurement error. The mean function $\mu(t, x_j)$ is specified as in (2):

$$\mu(t, x_j) = \mu_0(t) + f_1(t) \cdot \text{Order} + f_2(t) \cdot \text{Stress1} + f_3(t) \cdot \text{Stress2} + f_4(t) \cdot \text{Vowel} + f_5(t) \cdot \text{Order} \cdot \text{Stress1} + f_6(t) \cdot \text{Order} \cdot \text{Stress2} + f_7(t) \cdot \text{Order} \cdot \text{Vowel},$$  \hspace{1cm} (2)

with $\mu_0(t) + f_1(t), ..., f_7(t)$ as unknown fixed functions; functional random effects were modelled using functional principal components analysis (Ramsay & Silverman, 2005). Point-wise confidence bands show a significant effect for all covariates. Of the interactions, significant effects were observed only for Order $\cdot$ Stress1. Figure 2 shows the estimated reference group mean function, $\mu_0(t)$, and the effects of covariates Order, Stress1, and the effect of their interaction, each with point-wise confidence bands. The reference group mean corresponds to dummy coding 0 of all covariates (see equation (2)), i.e. Order = /ʃ#s/, Vowel = i-a, Stress1 = strong, Stress2 = strong. The effect on the index trajectory is denoted by $\Delta$ Index values ($\Delta Y$). The effect of a covariate on the mean can be obtained by multiplying the covariate effect with the dummy coding (1 or 0) and adding it to the reference group mean. The covariates Vowel and Stress2 (not shown here) mostly affected the transition between the sibilants.

The effect of Order is positive over the whole sibilant interval (Fig 2 top right). Recall also that we mirrored the /ʃ#s/ curves such that both C1#C2 conditions have a reference index dynamic ranging from 1 for C1 to -1 for C2. Since Order /ʃ#s/ was dummy coded with 1, Figure 2 shows that the mean curve is pulled during C1 more towards the ideal reference pattern (index = 1) for /ʃ/ than for /s/; for C2, /ʃ#s/ is slightly further away from the ideal reference pattern (-1). This is also evident in the estimated means in Figure 3 in that the /ʃ#s/ curves consistently lie above the /s#ʃ/ curves. For unstressed final target syllables (Stress1=1; Fig 2 lower left), index values in the beginning of the sibilant interval are (on average) lower...
and values in the end of the interval are (on average) slightly higher compared to stressed final target syllables. The effect is stronger on the final syllable (C1) than on the initial syllable (C2). This may be considered to be an artefact of the stress grouping, since Stress1 groups according to the stress of the final target syllable. Yet the Stress2 pattern which groups by initial target syllable stress (not shown here) speaks against this interpretation: Stress2 has an entirely different effect on the index trajectory compared to Stress1 in that there are no significant effects at either the beginning or end of the sibilant interval, rather, the transition in the temporal mid-region of the interval shows slightly raised index values, i.e. a shallower transition. Overall, we see a pronounced difference in index values between stressed and unstressed final target syllables. The Order · Stress1 interaction can be interpreted with reference to Figure 3: At the beginning and end of the sibilant interval, Stress1 has a slightly different effect on the mean curve for Order /f/ than for Order /ʃ/. Overall, there is a greater stress effect on C1 for /ʃ/ than for /f/ from which we conclude that stress interacts with assimilation.

4. Discussion and conclusion

We have presented acoustic and EPG data on German fricative assimilation with the aim of uncovering the contribution of articulatory and perceptual factors to fluent speech phenomena. For one, we could show that in the case of word final labio-dental fricative /ʃ/ and /f/, the sibilant overlaps the labio-dental in time as has been shown to be the case for stop articulations (Browman & Goldstein, 1990). We found a pronounced asymmetry between articulation and acoustics: while the articulatory data revealed the overlapping fricative constrictions, the acoustic classification showed a much lesser effect of the sibilant: fewer tokens were classified as assimilated acoustically than articulatorily. Overall, our data support the assumption that there is a similar degree of gestural overlap for /ʃ/ and /f/ sequences, yet due to the different articulators being involved, there are few acoustic consequences of the sibilant constrictive formation during the /ʃ/, since the sound source of /ʃ/ is anterior to the overlapping sibilant. For sibilant sequences, both consonants call on the same articulators, leading to blended articulations or, presumably due to its relatively greater dorsal control, a dominance of /ʃ/ (Pouplier et al., 2011).

There was further an order asymmetry in the data which at first blush may point to perceptual constraints: For sibilant/f/ clusters, there was no evidence for the labial fricative encroaching on the sibilant. Several recent publications have looked into the order asymmetry of sibilant assimilation and have found that /ʃ/ is prone to dominate in case of gestural overlap (Niebuhr et al., 2011; Pouplier et al., 2011; Recasens & Mira, 2013). For English /ʃ/ and /f/ sequences, the (near-)lack of regressive assimilation was attributed to /ʃ/ overlapping /f/ having almost no consequences for the articulation / acoustics of the palatal sibilant due to its tighter, more holistic tongue control. Our results for the sibilant condition are consistent with this interpretation; the German pattern is quite similar to the English one. Yet this scenario obviously cannot apply to the f-conditions, since in that case largely independent articulators are involved. For one, perceptual factors may be at play: assimilating a sibilant to /ʃ/ might be perceptually salient, therefore the sibilant may be protected from temporal overlap by the labio-dental. For the reverse order this argument is void, since the labio-dental fricative due to its anterior constriction is protected from assimilation acoustically. This would support the role of perceptual factors in fricative assimilation, even though in a more differentiated manner than has been proposed previously. However, with EPG data we have no positive information about the constrictive formation during /ʃ/. While the labio-dental constriction is beginning to be formed during the sibilant, we will see no record of this in the EPG data. Also acoustically, this will have virtually no consequences because the acoustic properties of the friction noise will greatly depend on the point of biggest pressure drop-off which is the point of narrowest constriction. That is, while the sibilant constrictive is at its maximum and /ʃ/ is in the process of being formed, there will be little evidence for formation in the acoustics. It may thus very well be the case that /ʃ/ overlaps a preceding sibilant to some degree, but we cannot tell from our data. In short, it is possible that some kind of planned overlap has in our type of data observable articulatory consequences for /ʃ/ and sibilant/f/ sequences. This issue will have to be pursued using recording techniques which give information on jaw and lip movement or articulatory modelling. Finally, we have presented a novel statistical approach for mixed modelling of irregularly sampled functional data. This analysis revealed that sibilants are generally sensitive to lexical stress in final position, but less so in initial position. Stress also had a more pronounced effect of final /s/ than on final /ʃ/, supporting the assumption that stress interacts with assimilation.

5. Acknowledgements

Work supported by the ERC under the EU’s 7th Framework Programme (FP/2007-2013)/Grant Agreement n. 283349-SCSPL and DFG Emmy Noether grants PO 1269/1-1 to M. Pouplier and GR 3793/1-1 to S. Greven. Thank you to our colleagues who took part in our experiment and to S. Waltl for help with acoustic segmentation.

6. References


335
Evaluation of an OPG-controlled animated vocal tract model as a biofeedback system

Simon Preuß, Christiane Neuschaefer-Rube, Peter Birkholz

Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital RWTH Aachen

Abstract

We recently proposed an animated vocal tract model controlled by optopalatography (OPG), a technique to track tongue movements inside of the mouth cavity. In this paper, we present a study on the benefit of an improved version of this model in a biofeedback application. To that end, three subjects articulated 16 sustained phonemes, which were recorded using our OPG system. Two of these subjects were then tasked with approximating these 48 target tongue contours and voicing the resulting sounds in two settings, once without visual feedback of their own tongue movements and once with real-time feedback. The results were then evaluated both geometrically (in terms of the distance between the target and approximated contour) and perceptively. We found that adding real-time feedback did reduce the geometric error significantly for one subject, yet this improvement did not lead to a significantly improved perceptive recognition rate. The second subject did not significantly improve neither the geometric nor the perceptive error. However, this pilot study did deliver valuable insights for future improvements on our biofeedback system.

Keywords: OPG, optopalatography, animated vocal tract, biofeedback

1. Introduction

Real-time biofeedback is becoming more and more popular in research, diagnostics, and therapy of speech disorders and pathologies (see, e.g., Engwall 2012 or Richmond and Renals 2012). All currently available techniques to measure speech movements (e.g., cineradiography, electromagnetic articulography, magnetic resonance imaging, ultrasonography), however, have major drawbacks (e.g., cost, precision, complexity) or limitations with respect to temporal and/or spatial resolution. For a biofeedback system to become accepted, it needs to be easily employable with minimum expert knowledge, reliable, precise, and cheap. In our recent and ongoing research (i.e., Birkholz and Neuschaefer-Rube 2011; Birkholz and Neuschaefer-Rube 2012; Birkholz and Neuschaefer-Rube 2012; Preuß, Neuschaefer-Rube, and Birkholz 2013b), we found the technology of optopalatography (OPG) to be a good compromise of these usually mutually exclusive requirements. Using this comparatively simple measurement technique as the data acquisition device, we developed a 2D animation model of the vocal tract that is able to visualize speech movements of a subject in real-time (Preuß, Neuschaefer-Rube, and Birkholz 2013b). In this paper, we present an experiment that was designed to explore the effectiveness of this system for biofeedback. The driving questions were: (1) Can one subject adopt the visually presented midsagittal tongue shape of a sound produced by another subject, (2) does it result in the same sound, and (3) does visual feedback on the tongue movement improve the results? A similar task is quite common in speech therapy, where the therapist demonstrates a certain tongue movement or position and the patient is asked to imitate it. Without instrumental support, this exercise is quite difficult as neither the patient can exactly see the target nor can the therapist directly assess the precision of the patient’s movement. Therefore, only descriptive means and acoustic properties can be applied to evaluate success or failure. Our study was meant to explore whether a visual target and real-time feedback in terms of the 2D tongue shape helps with this difficult imitation task.

2. Optopalatography

Optopalatography (OPG), also called glossometry, is a technique first introduced by Chuang and Wang (1978) and further refined by Fletcher et al. (1989) and Wrench, McIntosh, and Hardcastle (1996). It measures the midsagittal palatolinguinal distance by means of multiple optical sensors mounted on an artificial palate individually fitted to the subject. Each sensor consists of an infrared LED and a phototransistor. The LEDs are switched on in sequence and send out narrow beams of infrared light which are then diffusely reflected by the tongue below the hard palate. The light intensity at the corresponding phototransistor varies with the distance from the tongue to the hard palate, because less light is reflected to the phototransistor the further the tongue moves away from it. After determining the geometry of the pseudopalate and the optical axes of the sensors, the tongue contour is approximated by linear interpolation between the points marked at the measured distances along the respective sensor axes. This basic principle was further developed by Birkholz and Neuschaefer-Rube (2012) who added another sensor mounted facially of the maxillary incisors to measure labial aperture.

Our current OPG prototype (see Figure 1) consists of five optical sensors along the midsagittal palate contour and a sixth sensor to measure lip movement: The more light is reflected to that latter phototransistor, the more closed the lips must be. In this study, however, this sensor is not considered, because it can only capture the opening of the lips but not their protrusion (for a more detailed discussion of this problem see Preuss, Neuschaefer-Rube, and Birkholz 2013b). The sensor data are gathered by a microcontroller board and sent to a PC via a serial connection. Our system has a sampling rate of 100 Hz and can therefore acquire data of speech movements in real-time during running speech.
3. The animated vocal tract model

In most OPG systems the tongue contour is constructed by marking a point along the optical axis of each sensor at the respective measured distance and linearly interpolating between these points. In this way, for each sensor unit one point on the tongue surface below the hard palate is obtained. Our system, however, uses a multiple linear regression model to construct the entire tongue contour. The basic principle was shown in Preuß, Neuschaefer-Rube, and Birkholz (2013b) for five points and was further developed in Mumtaz et al. (accepted) to obtain 20 points along the entire tongue contour from tongue tip to hyoid, which are then connected by linear interpolation. For this experiment, we implemented the multiple linear regression model in our 2D animated vocal tract model (see Preuß, Neuschaefer-Rube, and Birkholz 2013b for basic information on the animation model). Also, the ability to load and display previously recorded tongue contours in addition to the currently measured contour was added. The speaker adaptation scheme described in Mumtaz et al. (accepted) was implemented as well to adapt tongue contours recorded with one speaker to the hard palate shape of another speaker.

4. Experimental setup

Three subjects (2 male, 1 female, age 28-35) were outfitted with individually created OPG pseudopalates. The subjects were than asked to voice the eight German vowels /aː, eː, iː, oː, uː, Eː, øː, yː/ and eight consonants (/f, s, ç, ɾ, n, m, l/) for a total of 16 sustained phonemes. During articulation of these sounds, the tongue contours were recorded using our OPG prototype. Each recording consisted of about 2 s of frames of the subjects’ tongue contour. To compensate noise and slight variations within the sustained articulation, the average tongue contours were determined by temporal averaging over 500 ms of recorded data. These average contours were used as the 48 target shapes (16 sounds, 3 subjects) for the following task. The two male subjects (age 28 and 35) were then presented with these target shapes in randomized order, adapted to their palate geometry by the speaker adaptation scheme mentioned above. The subjects did not know if the presented target tongue contour was from one of their own recordings, or which sound the contour corresponded to. They were tasked with approximating the targets in two settings: once without real-time feedback of their own tongue and once with real-time feedback (see Figure 3). When there was no feedback provided, the subjects had to adjust their tongue position only by intuition. In the setting with feedback, the subjects were given 30 seconds at most to align their tongue contour using the visual aid. Once they deemed the approximation a best fit, the subjects started voicing the resulting sound, and the adopted tongue contour and the acoustic utterance were recorded.
5. Evaluation

To quantify the difference between the two test conditions (without and with feedback), we used two error measures: The average geometric distance between the adopted and the target contour and the perceptive recognition rate. The latter was determined by an experienced phonetician who identified the produced phonemes perceptively. The perceived sounds were compared to the actual sounds corresponding to the target contours. As the geometric error for each of the 20 points, the closest Euclidean distance $d_i(s = 1...20)$ between that point and the target contour was determined. Therefore, the average error as a measure of the goodness of fit was defined as $\sum_{i=0}^{20} d_i$.

6. Results and discussion

Table 1 summarizes the results of the experiment broken down by sound. Even though the recognition rate slightly rises for both subjects when feedback is provided, neither improvement is significant (as determined by Fisher’s exact test). The geometric error was also only non-significantly reduced for subject 1 by adding feedback but significantly so for subject 2 (see Figure 4). However, this means that a significant improvement of the geometric error does not necessarily result in an improved overall recognition rate. A possible reason for this is that a single bad fit greatly influences the mean of the distribution. It can be assumed, that a subject achieves the closest approximation when trying to assume a target shape, which was recorded by the same subject. To alleviate the effect of bad fits we therefore now only consider these samples in the following.

![Figure 4: Distribution of the average error in cm over all contours. The numbers below each box are the respective mean of the distribution. Subject 2 achieved a significant improvement in this setting.](image)

Table 1: Average geometric error in cm over all approximated contours and perceptive recognition rate. The $p$-values were determined with Fisher’s exact test (two-tailed).

<table>
<thead>
<tr>
<th>Sound</th>
<th>Subject 1</th>
<th>Subject 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>without feedback</td>
<td>with feedback</td>
<td>without feedback</td>
</tr>
<tr>
<td>/æ/</td>
<td>0.411</td>
<td>0.392</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>0.268</td>
<td>0.223</td>
</tr>
<tr>
<td>/ɔ/</td>
<td>0.276</td>
<td>0.254</td>
</tr>
<tr>
<td>/u/</td>
<td>0.698</td>
<td>0.545</td>
</tr>
<tr>
<td>/i/</td>
<td>0.31</td>
<td>0.683</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>0.274</td>
<td>0.386</td>
</tr>
<tr>
<td>/a/</td>
<td>0.347</td>
<td>0.354</td>
</tr>
<tr>
<td>/e/</td>
<td>0.139</td>
<td>0.140</td>
</tr>
<tr>
<td>/o/</td>
<td>0.193</td>
<td>0.184</td>
</tr>
<tr>
<td>/ɔ/</td>
<td>0.626</td>
<td>0.626</td>
</tr>
<tr>
<td>/u/</td>
<td>0.172</td>
<td>0.395</td>
</tr>
<tr>
<td>/ɪ/</td>
<td>0.258</td>
<td>0.309</td>
</tr>
<tr>
<td>/u/</td>
<td>0.553</td>
<td>0.449</td>
</tr>
<tr>
<td>/ɔ/</td>
<td>0.236</td>
<td>0.358</td>
</tr>
<tr>
<td>/i/</td>
<td>0.362</td>
<td>0.267</td>
</tr>
</tbody>
</table>

Table 1: Average geometric error in cm over all approximated contours and perceptive recognition rate. The $p$-values were determined with Fisher’s exact test (two-tailed).

Table 2 shows the results when the subjects were tasked with approximating their own previously recorded target contours. Adding feedback did not improve the recognition rate for subject 1 at all and could not be found to make a significant difference for subject 2. The error distribution shown in Figure 5 also exhibits no significant improvement of the average geometric error.

The results show that the presented task was much harder than we had anticipated. The overall low recognition rate of no more than 33% necessitates a much larger number of subjects and/or items. It was also observed that the adopted tongue contour changed slightly once the subject started voicing the sound (after adjusting it to the target contour without voicing). This aspect is known to be a problem in silent speech interfaces (see, e.g., Denby et al. 2010) and might be investigated further using our feedback system.

The phonetician who evaluated the recordings also noted that the sounds /m/ and /n/ were very hard to discriminate from the audio recordings, further worsening the recognition rate. In a future study, these sounds could be merged to a single class to avoid confusion.

Even though the improvements could not be proven to be significant, we are confident that a larger study will give a more satisfying answer to the question, if this feedback system is useful for this sort of task. For example, some of the subjects showed a very good intuition when approximating some of the target contours even without feedback. More naive subjects without any phonetic knowledge should therefore be considered for the effect of the feedback to manifest more clearly.

In conducting the study we also realized that the information on lip protrusion would greatly benefit the subjects’ ability to approximate not only the correct tongue shape but also the correct sound. By further developing the lip feedback, we expect the relationship between geometric error and perceptive recognition rate to become more apparent. The subjects also had no cue, whether to articulate the sounds voiced or unvoiced. By providing this information, the confusion of phonemes could be further avoided. Also, the lateral dimension was completely missing in the feedback, which made discriminating /l/ and /n/ very difficult for the subjects. By combining OPG with electropalatography (EPG) (e.g., as described in Preuß, Neuschaefer-Rube, and Birkholz 2013a), this information would be available. A future study could also include speech impaired subjects and acoustic targets alongside the visual targets to assess the effect of feedback in a therapeutic context.

Another worthwhile investigation would be to further improve...
Table 2: Average geometric error in cm and perceptive recognition rate when approximating own contours. The p-values were determined with Fisher’s exact test (two-tailed).

Figure 5: Distribution of the average error in cm when approximating own contours. The numbers below each box are the respective mean of the distribution.
A New Technique for Automatic Shape Extraction from Vocal Tract MRI

Zeynab Raeesy1, Sylvia Rueda2, Jayaram K. Udupa3, John Coleman1

1Phonetics Laboratory, University of Oxford, Oxford, U.K.
2Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, U.K.
3Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104

1(zeynab.raeesy, john.coleman)@phon.ox.ac.uk
2sylvia.rueda@eng.ox.ac.uk, 3jay@mail.med.upenn.edu

Abstract

Automatic extraction of information from large MRI databases of articulation is challenging due to the variation in vocal tract shape and the dynamics of the articulators. This paper presents a new framework for automatic extraction of the vocal tract shape from MR images of articulation. The framework is based on an image segmentation technique, Oriented Active Shape Models (OASM), that combines the model-based and boundary-based image segmentation techniques. The segmentations are extensively analysed using both image-based evaluations, such as measures of similarity to manually segmented images, and phonetics assessments such as area function calculations and formant frequency estimations. The evaluation results suggest that the method is a promising approach to automatic extraction of the shapes that represent the prominent features of the vocal tract from large databases of MRI.

Keywords: vocal tract, speech production, MRI, automatic segmentation, shape extraction

1. Introduction

The non-invasive and safe nature of MRI has made it a popular choice for studying the hidden process of human articulation. The advent of technology for capturing articulation at higher rate has led to generation of large MRI databases, making the necessity of automatic techniques for information extraction and analysis inevitable.

A few approaches have been proposed for automatic extraction of information from vocal tract MRIs by tracking and modelling the articulators or contours [2, 3]. For extracting the entire shape, however, mostly manual or semi-automatic approaches such as thresholding have been used [13]. Vasconcelos et al. [17] proposed segmenting and modelling vocal tract shape using active shape models (ASM) [5] and active appearance models (AAM) [4]. The trained models were successfully used in recognising the shape in new images; however, as ASM relies on statistical information in training images, the method is not reliable in finding the accurate geometric boundaries in unobserved images.

Automatic extraction of the entire shape from MR images requires addressing numerous challenges such as (i) high between phonemes and between speakers variations in the shape, (ii) significant anatomical changes during the articulation due to movement of the articulators, and (iii) the connectivity of the tract airway to other channels of air through lip opening, velum opening or larynx. Although traditional methods such as thresholding can address challenges (i) and (ii) above, the connectivity of the airway limits their application to small databases, where manual supervision is applicable. A shape-based segmentation approach on the other hand can address challenge (iii) by defining a shape prior. However, to be able to address significant anatomical changes, it requires training on an extensive amount of data. In addition, since it uses only the statistics accumulated from training data, it does not take into account the properties of the target image. The fuzzy nature of the boundaries in images of an MRI database of dynamic speech requires the segmentation method to be able to handle the fuzziness of the boundaries on the target image. It is desirable to have a method that can address all of the challenges above by considering both the shape properties of the vocal tract airway and the pixel intensity gradient and variation at the boundaries.

We previously introduced a new framework for automatic shape extraction [10]. We applied oriented active shape models (OASM) [9] to overcome the challenges above by considering both the shape structure and orientation in segmenting new images. In this paper, we extend the framework by adapting the original OASM method to vocal tract shape extraction and creating multiple models for different speakers. Training multiple OASMs helps reduce the dimensionality of variation by ignoring inter-speaker variances. In addition, the need for automatic recognition of approximate position of the shape in the image is relaxed and replaced by ASM recognition, as the position of the vocal tract is mostly constant across the images of a speaker.

In the rest of this paper, we first present an overview of the segmentation framework, and continue by explaining the automatic landmark tagging approach and the shape modelling technique (Section 2). The experiments settings and evaluation metrics are presented in Section 3. Finally, in Section 4, we analyse and discuss the results from image-processing and phonetic aspects.

2. Segmentation Framework

The approach proposed here is based on OASM segmentation technique that combines the powerful statistical shape modelling aspects of ASMs and globally optimal capabilities of live wire [6]. Similar to most shape modelling techniques, OASM requires a training set annotated in terms of landmarks. Manual positioning of landmarks is a labour intensive and time-consuming task, not to mention its vulnerability to errors and subjective judgments. Therefore, automatic approaches for tagging the landmarks on the training shape are more desirable. We use recursive boundary subdivision (RBS) landmark tagging approach [11] to find the corresponding landmarks in the set of training images automatically. Figure 1
shows the overview of the segmentation framework. Below we explain the main components of the framework.

2.1. Automatic Landmark Tagging
The RBS approach [11] recursively and simultaneously subdivides the boundaries in a target set of shapes, and mathematically finds and characterises the landmarks on the subdivided segments. The recursive subdivision of the boundaries is a hierarchical procedure performed on all training shapes simultaneously. At each recursive step, the correspondence between the boundaries and subdivided segments is maintained using a specified similarity criterion, thus ensuring homology among landmarks. In this work, principal component analysis (PCA) was used to find the new landmark in each boundary subsegment. An important feature that made RBS with PCA a suitable choice for this application is the fact that the corresponding landmarks are found on the boundaries of the corresponding subsegments taking into account their variability at each iteration. Furthermore, with this approach there is no need for prior registration of the shapes, unlike most automatic landmark tagging methods which call for image/shape registration operations.

2.2. Oriented Active Shape Models
The OASM segmentation method combines shape-based and boundary-based segmentation techniques and thus is powerful in statistically modelling the shape structure and in delineating the boundaries optimally. OASM applies active shape models (ASM) for constructing the shape model and live wire for delineating the optimum boundary. ASM applies statistical information to model the vocal tract in terms of a mean shape that includes the characteristic pattern of a shape class and a linear combination of eigenvectors to reflect the variation around the mean shape. Live wire finds the globally optimal boundary of the shape by considering the features of the boundary elements between every two landmarks including boundary orientation and location. OASMs are applied in two stages, (a) training and shape model construction (both ASM and live wire training), and (b) segmentation and delineation of the target object in images. For training, first, similar to ASMs, the shape statistics are used to construct the shape model. An oriented boundary cost structure is then generated using live wire cost function training, capturing the intensity features of the boundary with regards to its orientation. For segmentation, in this work, ASM recognition is used to initialise the shape instance in the image. Live wire cost function is then utilised to find the optimal oriented boundary by minimising the sum of the costs of live wire boundary segment between each two successive landmarks on the shape.

3. Experiments and Evaluation metrics
We used a dynamic MRI database of British English articulations, collected in a previous project [1]. The images were collected with the gated MRI acquisition technique. The subject repeated a phrase of maximum two words while lying in supine position inside the scanner. A sequence of 68 2D images of 256×256 pixels was obtained at the end of each round of acquisition by accumulating the pseudo frames captured during each repetition of the phrase. We used the MR images of 6 different subjects (4 females and 2 males) for these experiments. For each speaker, two sets of images were selected separately for training and testing purposes. While for training the images were chosen to include maximum variety in terms of articulation manner, the test set was randomly sampled from the training images. For each speaker, 12 training images were used, one for every phone in the database including { /\ i \ , /l i , /l t , /l u , /l o , /l a , /l = , /l , /l s l , /l t l , /l d l , /l n l }. The same number of images were set aside for evaluating the performance.

The initial training images were semi-automatically segmented by live wire user-steered segmentation, using the 3DVIEWNIX software [15]. The vocal tract user-steered segmentations were then used as the input for the RBS automatic landmark tagging approach described in section 2.1. OASMs were trained using the generated sets of landmarks and their corresponding MR images. An OASM shape model was trained for each speaker separately, and the models were trained to include sufficient number of modes to embed 90% of variation in the original vocal tract shapes. Testing images of each speaker were segmented using the OASM trained specifically for that speaker.

3.1. Evaluation Metrics
Our evaluation experiments were carried out on the segmentations of test images that were not included in the training set. A set of user-steered “gold standard” segmentations was produced by an expert using live wire segmentation [15] on the testing images. The qualitative analysis shows how good the segmented shape is, given the expected vocal tract shape for each phoneme. The quantitative analysis uses distance-based and region-based metrics to numerically compare the segmentation results with the gold standard. We use region-based evaluation [16] to calculate the percentage of Dice similarity (D) of the segmentation results. The distance between gold standard segmentations and OASM segmentations is reported in terms of the Root Mean Square Symmetric Contour Distance (RMSD) metric [7] reported in millimetres (mm). While higher value of Dice similarity coefficient (closer to 100%) indicate a better performance, smaller values of RMSD (closer to 0) suggest smaller deviation of OASM segmentations from the gold standard.

For articulatory evaluations the formants are estimated from

\[ \text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2} \]

\[ \text{Dice} = \frac{2 \times \text{overlap}}{\text{area1} + \text{area2}} \]

\[ \text{Dice} = \frac{\text{intersection}}{\text{union}} \]

\[ \text{Intersection} = \text{area1} \cap \text{area2} \]

\[ \text{Union} = \text{area1} \cup \text{area2} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Intersection} = \text{area1} \cap \text{area2} \]

\[ \text{Union} = \text{area1} \cup \text{area2} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Intersection} = \text{area1} \cap \text{area2} \]

\[ \text{Union} = \text{area1} \cup \text{area2} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Intersection} = \text{area1} \cap \text{area2} \]

\[ \text{Union} = \text{area1} \cup \text{area2} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Intersection} = \text{area1} \cap \text{area2} \]

\[ \text{Union} = \text{area1} \cup \text{area2} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Intersection} = \text{area1} \cap \text{area2} \]

\[ \text{Union} = \text{area1} \cup \text{area2} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]

\[ \text{Dice} = \frac{2 \times \text{Intersection}}{\text{Area1} + \text{Area2}} \]
the area functions of the OASM vocal tract segmentations. Since our images were 2-dimensional, we used a transformation function to estimate area functions from the mid-sagittal distances. For area function computation, a semi-polar gridline system was fitted to the image in the mid-sagittal plane, and the mid-sagittal widths were calculated at different distances from the glottis along the vocal tract. We then used the classic \((\alpha, \beta)\) function by Heinz and Stevens \([8]\) to calculate the area functions at each distance from the glottis. The transformation parameters \(\alpha\) and \(\beta\) are usually adapted by researchers for each particular study and, in general, the optimum parameters are speaker specific. However, we did not have the means to optimise and adapt the parameters, we therefore simply used the parameters for male and female speakers proposed in \([12]\). We estimated the formant frequencies from the obtained area functions using VTAR software \([18]\).

4. Results and Discussion

The results are reported individually per speaker as speaker-specific models were constructed. We first explain the qualitative analysis, followed by quantitative evaluations and phonetic assessments.

4.1. Qualitative Evaluations

We performed a subjective evaluation on all 72 test images. Figure 2 shows a few examples of the automatically obtained and user-steered contours plotted over the original images. Our observations suggest that if OASM successfully finds the initial position of the shape, it can precisely delineate the vocal tract. The details can indeed be very fine at sensitive areas such as the opening of the lips or the sublingual cavity. The generated boundaries include most of the features crucial in describing the vocal tract shape, such as the width of the gap between the lips, the posture of the velum, the height of the tongue, and the width of the cavity all along the vocal tract from glottis to lips. In some instances, however, some features such as length of the cavity cannot be as precisely estimated by OASM segmentations (see the difference in Figure 2 (e) and (f)).

4.2. Quantitative Evaluations

Quantitative evaluations were carried out separately for each speaker and were then averaged over all speakers. Table 1 shows the region-based and distance-based evaluations between the OASM and user-steered contours. The Dice similarity values vary from 76.5\% to 87.71\% for different speakers, suggesting that the minimum overlap between the OASM segmentations and the gold standards is above 75\%. The distance-based measurements provided in the table suggest that the average RMSD error between the boundaries in gold standard and OASM segmentations is 2.49 mm. We further examine these differences by performing a phonetic analysis on the extracted vocal tract shapes, as described below.

4.3. Formant Frequency Estimations

Table 2 shows the estimated values of the first two formants generated from the vocal tract shapes automatically extracted using the OASM framework. We compare our estimations with the formant frequencies of speaker BS in \([14]\). Note that the absolute values are not actually comparable as the speakers are different, and the glottal source model is different. In addition, the values in \([14]\) are obtained from static rather than running speech. We use these values to perform only a general comparison and assess the applicability of the model in generating relevant formants in speech production studies.\(^2\) For most of the speakers, the estimated results are consistent with the expected articulation of the vowels. The first formant (F1) frequencies are consistently higher in low vowels such as /a/ and /i/ compared to high vowels such as /ü/ and /u/. The second formant, F2, follows a consistent pattern where it has lower frequencies for back vowels such as /u/ and /a/ and higher frequencies for front vowels such as /i/ and /e/.

---

\(^2\)Observing the formant frequencies and absolute values individually is not very informative, as the ratio of the formant frequencies to each other is what is important in generating and distinguishing sounds.

Table 1: The mean and standard deviation of region-based and distance-based evaluations for speaker-specific OASMs.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Region-Based(%)</th>
<th>Distance-Based (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker A</td>
<td>76.51 ± 4.84</td>
<td>2.25 ± 0.49</td>
</tr>
<tr>
<td>speaker C</td>
<td>86.17 ± 2.74</td>
<td>1.82 ± 0.80</td>
</tr>
<tr>
<td>speaker L</td>
<td>85.01 ± 2.87</td>
<td>2.01 ± 0.48</td>
</tr>
<tr>
<td>speaker M</td>
<td>87.71 ± 4.09</td>
<td>1.65 ± 0.41</td>
</tr>
<tr>
<td>speaker P</td>
<td>81.75 ± 4.88</td>
<td>3.07 ± 0.67</td>
</tr>
<tr>
<td>speaker R</td>
<td>78.24 ± 3.59</td>
<td>3.17 ± 0.47</td>
</tr>
</tbody>
</table>
Table 2: First two formants’ frequencies (Hz) based on the area functions of our speakers, and of speaker BS from [14]. Note that some of the vowels were not available for all speakers (‘NA’). The values with † are the exceptions to the expected formant distribution pattern.

<table>
<thead>
<tr>
<th>speaker</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>876</td>
<td>826</td>
<td>666</td>
<td>681</td>
<td>406</td>
<td>441</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1386</td>
<td>1631</td>
<td>1141</td>
<td>1971</td>
<td>1901</td>
<td>1621</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>696</td>
<td>846</td>
<td>651</td>
<td>601</td>
<td>496</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>1306</td>
<td>1641</td>
<td>1296</td>
<td>2226</td>
<td>2016</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>731</td>
<td>711</td>
<td>421</td>
<td>421</td>
<td>NA</td>
<td>471</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>1576</td>
<td>1641</td>
<td>1061</td>
<td>2661</td>
<td>2141</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>851</td>
<td>1041</td>
<td>821</td>
<td>441</td>
<td>601</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>1596</td>
<td>2206</td>
<td>1411</td>
<td>1646</td>
<td>2311</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>551</td>
<td>621</td>
<td>516</td>
<td>446</td>
<td>NA</td>
<td>401</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1271</td>
<td>1446</td>
<td>1121</td>
<td>1621</td>
<td>2141</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>721</td>
<td>656</td>
<td>626</td>
<td>451</td>
<td>291</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>1451</td>
<td>1531</td>
<td>1466</td>
<td>2251</td>
<td>2546</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>826</td>
<td>706</td>
<td>641</td>
<td>516</td>
<td>301</td>
<td>326</td>
<td></td>
</tr>
<tr>
<td>BS</td>
<td>1146</td>
<td>1301</td>
<td>1061</td>
<td>2021</td>
<td>2496</td>
<td>1121</td>
<td></td>
</tr>
</tbody>
</table>

vowels /i/ and /u/ in the majority of speakers. The lip rounding in /a/ leads to the lowest F2 among different vowels. The outliers to the patterns above are marked by a † in the table. The obtained values allow a clear classification between categories of vowels on IPA chart, e.g. back vowels, front vowels, high vowels and low vowels.

Overall, our evaluations suggest that the proposed framework is a promising approach for automatically extracting the vocal tract shape and its important features. The extracted shape not only can be used for studying articulation, but also for generating vocal tract shape models from large databases. Some inconsistencies in pattern within the categories of vowels (noticed in phonetic evaluations) are caused by the coarticulatory effects, as the vowels are not articulated in isolation. The inconsistencies can also be related to the qualitative observation that the vocal tract lengths with OASM segmentation are occasionally shorter than the actual length of the vocal tract, and also the fact that the dentition information are not available. The method can be possibly improved by placing some landmarks manually which is essential to the effectiveness of live wire, and could be further analyzed by providing more information about the anatomical structure of the tract, such as dentition information.

References


We present a method to derive a small number of speech motor control “primitives” that can produce linguistically-interpretable articulatory movements. We envision that such a dictionary of primitives can be useful for speech motor control, particularly in finding a low-dimensional subspace for such control. First, we use the iterative Linear Quadratic Gaussian (iLQG) algorithm to derive (for a set of utterances) a set of stochastically optimal control inputs to a dynamical systems model of the vocal tract that produces desired movement sequences. Second, we use a convolutive Nonnegative Matrix Factorization with sparseness constraints (cNMFsc) algorithm to find a small dictionary of control input primitives that can be used to reproduce the aforementioned optimal control inputs that produce the observed articulatory movements. Such a primitives-based framework could help inform theories of speech motor control and coordination.

Keywords: speech motor control, motor primitives, synergies, dynamical systems, iLQG, NMF.

1. Introduction

Mussa-Ivaldi and Solla (2004) [1] argue that in order to generate and control complex behaviors, the brain does not need to solve systems of coupled equations. Instead a more plausible mechanism is the construction of a vocabulary of fundamental patterns, or primitives, that are combined sequentially and in parallel for producing a broad repertoire of coordinated actions. An example of how these could be neurophysiologically implemented in the human body could be as functional units in the spinal cord that each generate a specific motor output by imposing a specific pattern of muscle activation [2]. Although this topic remains relatively unexplored in the speech domain, there has been significant work on uncovering motor primitives in the general motor control community. For instance, [3, 2] proposed a variant on a nonnegative matrix factorization algorithm to extract muscle synergies from frogs that performed various movements. More recently, [4] extended these ideas to the control domain, and showed that the various movements of a two-joint robot arm could be effected by a small number of control primitives.

In previous work [5, 6], we proposed a method to extract interpretable articulatory movement primitives from raw speech production data. Articulatory movement primitives may be defined as a dictionary or template set of articulatory movement patterns in space and time, weighted combinations of the elements of which can be used to represent the complete set of coordinated spatiotemporal movements of vocal tract articulators required for speech production. In this work, we propose an extension of these ideas to a control systems framework. In other words, we want to find a dictionary of control signal inputs to the vocal tract dynamical system, which can then be used to control the system to produce any desired sequence of movements.

2. Data

We analyzed synthetic VCV (vowel-consonant-vowel) data generated by the Task Dynamics Application (or TaDA) software [7, 8] – which implements the Task Dynamic model of inter-articulator speech coordination with the framework of Articulatory Phonology [9]. We chose to analyze synthetic data since (i) articulatory data is generated by a known compositional model of speech production, and (ii) we can generate a balanced dataset of VCV observations. TaDA also incorporates a coupled-oscillator model of inter-gestural planning, a gestural-coupling model, and a configurable articulatory speech synthesizer [10, 11] (see Figure 1). TaDA generates articulatory and acoustic outputs from orthographical input. The orthographic input is syllabified using the Carnegie Mellon pronouncing dictionary, parsed into gestural regimes and inter-gestural coupling relations using hand-tuned dictionaries and then converted into a gestural score. The obtained gestural score is an ensemble of gestures for the utterance, specifying the intervals of time during which particular constriction gestures are active.
This is finally used by the Task Dynamic model implementation in TaDA to generate the task variable and articulator time functions, which are further mapped to the vocal tract area function (sampled at 200 Hz).

We generated 972 VCVs corresponding to all combinations of 9 English monophthongs and 12 consonants (including stops, fricatives, nasals and approximants). Each VCV can be represented as a sequence of articulatory states. In our case, the articulatory state at each sampling instant is a 10-dimensional vector comprising the 8 articulatory parameters plotted in Figure 1 and 2 additional parameters to capture the nasal aperture and glottal width. We then downsampled the articulatory state trajectories to 100 Hz. We further normalized data in each channel (by its range) such that all data values lie between 0 and 1.

3. Dynamical systems model of the vocal tract

We need to define a suitable model of vocal tract dynamics which we can then use to simulate vocal tract dynamics on application of control inputs. We will adopt and extend the Task Dynamics model of speech articulation (after [13]):

\[ M \ddot{z} + B \dot{z} + Kz = \tau \]

where \( z \) refers to the task variable (or goal variable) vector, which is defined in TaDA as a set of constriction degrees (such as lip aperture, tongue tip constriction degree, velic aperture, etc.) or locations (such as tongue tip constriction location). \( M \) is the mass matrix, \( B \) is the damping coefficient matrix, and \( K \) is the stiffness coefficient matrix of the second-order dynamical system model. \( \tau \) is a control input. However, in motor control, we typically cannot directly control these task variables. We can

\[ z = f(\phi) \]

\[ \dot{z} = J(\phi)\dot{\phi} \]

\[ \ddot{z} = J(\phi)\ddot{\phi} + \dot{J}(\phi, \dot{\phi})\dot{\phi} \]

\[ \ddot{\phi} + (J^\top M^{-1}BJ + J^\top J)\dot{\phi} + J^\top M^{-1}Kf(\phi) = \tau \]

Figure 1: A visualization of the Configurable Articulatory Synthesizer (CASY) in a neutral position, showing the outline ... value was chosen empirically as the mean of \( \omega \) values that TaDA uses for consonant and vowel gestures respectively.

Figure 2: Schematic illustrating the proposed method. The input articulatory state trajectories and a simple second-order model of the system dynamics are used as input to the iLQG algorithm to generate a matrix \( V \) of control inputs. This matrix \( V \) is then passed as input to the cNMFsc algorithm to obtain a three-dimensional matrix of articulatory primitives, \( W \), and an activation matrix \( H \), the rows of which denote the activation of each of these time-varying primitives/basis functions in time. In this example, we assume that there are \( M = 7 \) control trajectories. Each vertical slab of \( W \) is one of 5 primitives (numbered 1 to 5). For instance, the white tube represents a single component of the \( 3^{rd} \) primitive that corresponds to the first control dimension (\( T \) samples long).

only control so-called articulatory variables, \( \phi \), which can be nonlinearly related to the task variables using the so-called ‘direct kinematics’ relationship (Equation 2-4). Using this relationship, we can derive the equation of the corresponding dynamical system for controlling the model articulators (for e.g., see Figure 1):
4. Computing control synergies

4.1. Computing optimal control signals

To find the optimal control signal for a given task, a suitable cost function must be minimized. Unfortunately, when using nonlinear systems such as the vocal tract system described above, this minimization is computationally intractable. Researchers typically resort to approximate methods to find locally optimal solutions. One such method, the iterative linear quadratic gaussian (iLQG) method [14, 15, 4], starts with an initial guess of the optimal control signal and iteratively improves it. The method uses iterative linearizations of the nonlinear dynamics around the current trajectory, and improves that trajectory via Riccati equations.

In our case, we pass as input to this algorithm articulator trajectories (see Section 2), and obtain as output a set of control signals (timeseries) $\tau$ that can effect those sequences of movements (one timeseries per articulator trajectory).

4.2. Extraction of control primitives

Modeling data vectors as sparse linear combinations of basis elements is a general computational approach (termed variously as dictionary learning or sparse coding or sparse matrix factorization depending on the exact problem formulation) which we will use to solve our problem [16, 17, 18, 19, 20]. If $\tau_1, \tau_2, \ldots, \tau_N$ are the $N = 972$ control matrices obtained using iLQG for each of the 972 VCVs, then we will first concatenate these matrices together to form a large data matrix $V = [\tau_1 \mid \tau_2 \mid \ldots \mid \tau_N]$. We will then use convolutive non-negative matrix factorization or cNMF [18] to solve our problem. cNMF aims to find an approximation of the data matrix $V$ using a basis tensor $W$ and an activation matrix $H$ in the mean-squared sense. We further add a sparseness constraint on the rows of the activation matrix to obtain the final formulation of our optimization problem, termed cNMF with sparseness constraints (or cNMFsc) [5, 6]:

$$\min_{W,H} \| V - \sum_{t=0}^{T-1} W(t) \cdot H_i^t \|^2 \text{ s.t. sparseness}(h_i) = S_{h_i}, \forall i. \tag{6}$$

where each column of $W(t) \in \mathbb{R}^{20 \times M \times K}$ is a time-varying basis vector sequence, each row of $H \in \mathbb{R}^{20 \times K \times N}$ is its corresponding activation vector ($h_i$ is the $i^{th}$ row of $H$), $T$ is the temporal length of each basis (number of image frames) and the ($^{\top}$) operator is a shift operator that moves the columns of its argument by $i$ spots to the right, as detailed in [18]. Note that the level of sparseness ($0 \leq S_{h_i} \leq 1$) is user-defined. See Narayanan et al. [5, 6] for the details of an algorithm that can be used to solve this problem.

5. Experiments and Results

The three-dimensional $W$ matrix and the two-dimensional $H$ matrix described above allows us to form an approximate reconstruction, $V_{\text{recon}}$, of the original control matrix $V$. This matrix $V_{\text{recon}}$ can be used to reconstruct the original articulatory trajectories for each VCV by simulating the dynamical system in Equation 5. The root mean squared errors of the original movements and controls were approx. 0.34 and 0.4, respectively, on average. The cNMFsc algorithm parameters used were $S_{h_i} = 0.65, K = 8$ and $T = 10$. The sparseness parameter was chosen empirically to reflect the percentage of gestures that were active at any given sampling instant (~ 35%), while the number of bases was selected based on the Akaike Information Criterion or AIC [21], which in this case tends to prefer more parsimonious models. The temporal extent of each basis was chosen to capture effects of the order of 100ms. See [6] for a more complete discussion on parameter selection.

Note that each control primitive could effect different movements of vocal tract articulators depending on their initial position/configuration. For example, Figure 3 shows 8 movement sequences effected by 8 control primitives for one particular choice of a starting position. Each row of plots were generated by taking one control primitive sequence, using it to simulate the dynamical system learned using the iLQG algorithm, and visualizing the resulting movement sequence$^2$.

6. Conclusions and Outlook

We have described a technique to extract synergies of control signal inputs that actuate a learned dynamical systems model of the vocal tract. Work described in this paper can help in formulating speech motor control theories that are control synergy- or primitives-based. The idea of motor primitives allows us to explore many longstanding questions in speech motor control in a new light. For instance, consider the case of coarticulation in speech, where the position of an articulator/element may be affected by the previous and following target [22]. Constructing internal neural representations from a linear combination of a reduced set of modifiable basis functions tremendously simplifies the task of learning new skills, generalizing to novel tasks or adapting to new environments [23].

References


$^2$The extreme overshoot/undershoot in some cases could be an artifact of normalization. Having said that, it is important to remember that the original data will be reconstructed by a scaled-down version of these primitives (weighted down by their corresponding activations)
Figure 3: Spatio-temporal movements of the articulator dynamical system effected by 8 different control primitives for a given choice of initial position. Each row represents a sequence of vocal tract postures plotted at 20 ms time intervals, corresponding to one control primitive sequence. The initial position in each case is represented by the first image in each row. The cNMfsc algorithm parameters used were $S_p = 0.65$, $K = 8$ and $T = 10$ (similar to [6]). The front of the mouth is located toward the right hand side of each image (and the back of the mouth on the left).


Modelling voicing assimilation in Catalan consonant clusters
Daniel Recasens¹, Meritxell Mira²

¹Departament de Filologia Catalana, Universitat Autònoma de Barcelona & Laboratori de Fonètica, Institut d’Estudis Catalans, Barcelona, Spain
²Departament de Filologia Catalana, Universitat Autònoma de Barcelona, Spain

daniel.recasens@uab.es, meritxell.mira@gmail.com

Abstract
This study explores the phonetic validity of a Catalan regressive voicing assimilation rule applying to one or two syllable-final obstruents before a syllable-initial voiced consonant in C#C and CC#C sequences where # stands for a word boundary. Electroglostographic data do not confirm a strong version of the assimilation rule since syllable-final obstruents may be realized as mostly voiced or mostly voiceless depending on speaker, and their voicing degree is affected considerably by cluster complexity and segmental composition, as well as the position of the target obstruent within the cluster. In accordance with the rule, regressive voicing effects triggered by syllable-final obstruents onto the syllable-initial voiced consonant were found to be small or absent. This scenario suggests that languages where voiced stops have considerable voice-leading (Catalan, French, Russian) differ from those where this is not the case (English, German) regarding the extent of the regressive voicing effects relative to that of the regressive voicing effects.

Keywords: voicing assimilation, electroglostography, consonant clusters, Catalan

1. Introduction
This paper presents a theoretical account of electroglostographic (EGG) data reported in Recasens and Mira (2012, 2013) dealing with voicing adaptation of syllable-final obstruents to a syllable-initial voiced consonant in Catalan C#C and CC#C sequences where # stands for a word boundary. A major goal of these two studies was to verify the phonetic validity of a phonological rule of regressive voicing assimilation according to which one or two syllable-final obstruents, i.e., C1 in C#C sequences and C2 or C1 and C2 in CC#C sequences, assimilate fully in voiced to the syllable-initial voiced consonant (Wheeler 2005). According to this phonological rule, word final /s/ and /ps/ are realized, respectively, as [z] and [fs] before word initial /b/, as in vas bo “good glass” and taps bons “good corks”. Since underlying voiced obstruents undergo final voicing in Catalan, the presence of voicing during the production of syllable-final obstruents in the C#C and CC#C sequences of interest should be attributed essentially to the syllable-initial voiced consonant trigger (this expectation is also based on EGG data for the same speakers analyzed in the present investigation showing that the voicing lag associated with the vowel preceding the cluster is less than 20% of the C1 duration). The present study was intended to test the validity of this phonological rule on the following consonant sequence structures:

(a) C#C combinations with a stop or a fricative C1, as in vas bo “good glass” and tap bo “good cork”;
(b) CC#C combinations where C2 is always an obstruent and C1 may be either an obstruent or a sonorant, as in tups bons “good corks” and rams bons “good branches”.

In principle, the Catalan scenario ought to parallel that of languages such as French and Russian where voiced stops keep the vocal folds vibrating extensively during the closing phase (and therefore exhibit a long negative VOT value in postpausal position), and voiceless stops are unaspirated and thus produced with a positive VOT lasting for no more than about 40 ms. This scenario is in contrast with that for English and German where voiceless stops are aspirated, voiced stops do not show high negative VOT values, and voiceless-voiced obstruent consonant clusters may be implemented through regressive voicing, regressive voicing or no voicing assimilation (Westbury 1979, van Dommelen 1983). A major issue to be addressed in the present investigation is whether the two language groups behave so differently by exploring the following topics:

(a) Whether the regressive voicing assimilation process applies categorically and obligatorily and, therefore, causes syllable-final obstruents to become completely voiced for all or most Catalan speakers, or else varies in phonetic implementation depending on speaker, number of consonants in the cluster, C1 or C2 obstruent position in CC#C sequences, and place and manner of articulation of the syllable-final obstruent and the syllable-initial consonant voicing trigger. There appear to be some language-dependent differences in this respect, with on the one hand languages such as Russian where regressive voicing yields fully voiced syllable-final obstruents (Burton and Robblee 1997) and on the other languages such as French where voicing assimilation may proceed in a more gradient manner (Halle and Adda-Decker 2007).

(b) The extent to which voicing adaptation in Catalan takes place not only at the regressive level but at the progressive level as well. If so, syllable-final obstruents not only should become voiced before a syllable-initial voiced consonant but also ought to cause the voiced consonant in question to devoice to a greater or lesser degree.

Another goal of the present study is to investigate in detail the patterns of voicing coarticulation in CC#C sequences. In particular, the following issues will be looked into:

(a) If, as claimed by the Catalan voicing assimilation rule, C3-dependent regressive voicing effects operate on both C1 and C2 or just on C2 when C1 and C2 are obstruent.
(b) Whether voicing effects may take place between intrasyllabic C1 and C2. Thus, it could be that the voicing degree for C2 (which is always an obstruent in CC#C sequences) varies depending on whether the preceding consonant is a sonorant or an obstruent.

348
2. Method

The material to be recorded included the C#C and CC#C sequences listed below where the last consonant is always voiced underlying, C1 is an obstruent in C#C combinations, and C1 may be either an obstruent or a sonorant and C2 is always an obstruent in CC#C sequences. The CC#C sequences have been split depending on whether C2 is a fricative or a stop.

[C#C sequences]
C1=/p, t, k, f, s, j/;
C2=/b, d, g, z, m, n, l, r, ň, j/;

[CC#C sequences with a fricative C2]
C1C2=/ps, ts, ks, ňs, ls, ms, tf, mf, lf, rf/ +
C3=/b, d, g, m, n, l, ň, j/;

[CC#C sequences with a stop C2]
C1C2=/lp, rp, sp, lk, rk, sk/ +
C3=/b, d, g, m, n, l, z, r, ň, j/;

In the consonant sequences under analysis, /r/ is realized as an alveolar trill in syllable-initial position but shows a more tap-like articulation syllable-finally, and /s/ is an alveolar or alveolo-palatal lateral; moreover, /b, d, g/ may exhibit an approximant realization when preceded by a fricative consonant. Obstruents appearing in syllable-final position are transcribed with the phonetic symbol for voiceless stops and fricatives irrespective of whether they are phonologically voiced or voiceless since, as mentioned above, the voicing distinction for obstruents is neutralized syllable-finally in Catalan.

All consonant clusters embedded in meaningful sentences with the same number of syllables and stress pattern were read seven times by the native Catalan speakers SO, MA, LO, VA, PE, EV, MO and DR. Simultaneous EGG and acoustic recordings were carried out using an EGG-2 glottograph unit manufactured by Glottal Enterprises plugged into the Computerized Speech Laboratory (CSL) data acquisition system. Both signals were acquired at 44100 Hz, and then downsampled to 500 Hz and 11025 Hz for the EGG and acoustic signals respectively. The EGG signal was smoothed and analyzed using the MatLab script Peakdet 2. Consonant onsets and offsets were identified on spectrographic displays, and voicing onsets and offsets were labelled at positive peaks of the first derivative of the glottal waveform (see Recasens and Mira 2012 for details). The consonant voicing measure will be expressed as a percentage of voicing over consonant duration.

In order to correlate the degree of voicing for syllable-final obstruents in consonant clusters and the VOT values for voiced stops, an additional set of sentences with postpausal /b, d, g/ was also recorded and the time interval between voicing onset and the stop burst was measured on the EGG waveform.

3. Results

The EGG data for the individual cluster tokens reveal that regressive voicing assimilation in heterosyllabic consonant clusters does not operate categorically. As shown by Figure 1, voicing percentages for syllable-final obstruents computed across consonant context conditions and speakers are lower than expected, i.e., they range between about 20% and 60%. They also vary with several factors: they are lower for syllable-final fricatives than for stops (which is in accordance with the difficulty involved in combining glottal opening for the passage of airflow with uninterrupted vocal fold vibration for fricatives), lower for CC#C than for C#C sequences and thus for more vs. less complex clusters, and lower for C2 than for C1 in CC#C sequences and thus for consonants placed in the middle of the cluster than for those at its onset.

As shown in Figure 2, regressive voicing was also conditioned significantly by the manner of articulation of the syllable-initial voiced consonant trigger, with syllable-final stops and fricatives in C#C sequences and in C2 position in CC#C sequences being more extensively voiced when the voicing trigger was an obstruent (i.e., a fricative or a stop) than when it was a sonorant (i.e., a nasal, a lateral or a rhotic though not /j/ which causes considerable voicing to occur during the preceding obstruent). This outcome is not in agreement with the fact that sonorants exhibit more voicing than obstruents, and may be attributed to the conflicting requirements involved in the production of consonant combinations such as fricative + nasal where anticipatory velar lowering and voicing for the nasal could impair the high intensity friction noise for the preceding fricative (Ohala and Solé 2010). Moreover, in CC#C sequences, C3-dependent voicing differences were barely transmitted to C1, e.g., the degree of C1 voicing did not always vary significantly as a function of whether C3 was an obstruent or a sonorant.

Figure 1: Voicing percentages for the syllable-final stops /p, t, k/ (left graph) and fricatives /j, s, f/ (right graph) across contextual consonants, tokens and speakers plotted as a function of their position in C#C and CC#C sequences. The position of the syllable-final obstruent in the consonant cluster is indicated in boldface and underlined. Error bars represent one standard deviation.

Figure 2: Voicing percentages for syllable-final obstruents placed in C1 position in C#C sequences and in C2 position in CC#C sequences plotted as a function of syllable-initial sonorants (filled bars) and obstruents (empty bars). Data have been averaged across target and contextual consonants, tokens and speakers.
In agreement with the formulation of the regressive voicing assimilation rule, progressive devoicing effects exerted by syllable-final obstruents onto the syllable-initial voiced consonant were small or absent. Thus, as shown by Figure 3, voicing percentages for syllable-initial consonants computed separately as a function of preceding C1=p, t, k, f, s, ъ in C#C sequences and C2=p, k, f, s, ъ in CC#C sequences amount to 70-80% of the consonant duration in all cases. These high percentage values indicate that progressive devoicing is practically absent, mostly so if we take into consideration the fact that the intrinsic manner of articulation demands rather than the preceding consonant account for some devoicing in syllable-initial consonants such as trills and voiced stops.

Progressive voicing effects were found to occur intrasyllabically in CC#C sequences: an obstruent C2 turned out to exhibit more voicing after a sonorant C1 than after an obstruent C1, e.g., /s/ after /m/ in the sequence /msC/ than after /p/ in the sequence /psC/. This difference achieved significance (see Recasens and Mira 2013) and, as shown in Figure 4, was more obvious when C2 was a stop than when it was a fricative. This finding suggests that there is some principle of voicing organization at the syllable level.

Figure 3: Voicing percentages across syllable-initial voiced consonants, tokens and speakers plotted as a function of the immediately preceding stops /p, t, k/ (left graph) and fricatives /f, s, ъ/ (right graph) in C#C and CC#C sequences. Error bars represent one standard deviation.

Figure 4: Voicing percentages for fricatives and stops in C2 position plotted as a function of preceding sonorants (filled bars) and obstruents (empty bars) in CC#C sequences. Data have been averaged across target and contextual consonants, tokens and speakers.

Speakers turned out to differ a great deal regarding voicing degree for syllable-final obstruents. Thus, obstruct C1 voicing in C#C sequences computed across contextual consonants ranged from about 20%-40% for four speakers (SO, MA, PE, EV) to 60-80% for three other speakers (LO, VA, DR) (see thin continuous line in Figure 5). Moreover, these speaker-dependent voicing differences were highly correlated with speaker-dependent voicing differences for the following syllable-initial voiced C2 (r = 0.935), i.e., C1 exhibited more or less voicing depending on the amount of voicing occurring during C2 (compare the thick and thin continuous lines in Figure 5). A similar scenario held in CC#C sequences, i.e., speaker-dependent voicing differences for C2 and C3 were positively correlated. A comparison between the continuous and discontinuous lines in the figure also shows that, when data for speaker DR are excluded, a close relationship is obtained between C1 and C2 voicing degree in C#C sequences, on the one hand, and VOT duration, on the other: correlation values were r = 0.680 for the C1 voicing/VOT relationship and r = 0.567 for C2 voicing/VOT. This finding suggests that regressive voicing assimilation in Catalan consonant clusters is to a large extent mechanically conditioned by the degree of voicing for the consonant voicing trigger.

Figure 5: Voicing percentages for C1 and C2 in C#C sequences, and negative VOT values for postnasal /b, d, g/, according to each of the Catalan speakers subject to investigation. VOT values have been rendered positive for comparison with the C1 and C2 voicing percentages.

4. Discussion and Conclusion

The Catalan voicing data presented in the Results section agree with the prediction of the phonological voicing assimilation rule that in languages where voiced stops exhibit prevocing C#C and CC#C clusters with a syllable-initial voiced consonant should show regressive voicing but little or no progressive devoicing. However, they run against the assimilation rule in that the degree of regressive voicing varies considerably with speaker as well as other factors such as cluster complexity, segmental composition of the cluster and position of the syllable-final obstruent. The degree of regressive voicing assimilation could also vary as a function of language and dialect; thus, judging from the available published data on the subject, the process in question applies categorically in Russian but not necessarily in Catalan, where speakers may exhibit less syllable-final obstruent voicing than in French (Burton and Robblee 1997, Snoeren, Hallé and Segui 2006, Hallé and Adda-Decker 2007). This scenario also indicates that differences in regressive voicing assimilation between these languages and those without prevoced stops such as English and German involve both the regressive and progressive components: in languages with prevoced stops regressive voicing may be extensive and progressive devoicing is minimal, while in languages without prevoced stops regressive voicing effects are often small and progressive devoicing may be quite substantial. Differences in degree of
regressive and progressive assimilation between the two language types follow presumably from language-specific differences in the articulatory and aerodynamic mechanisms involved in stop and fricative production (see Solé 2011 and Solé and Sprouse 2011 in this respect). Moreover, the fact that low voicing values for single fricatives and stops have been found to occur in other Romance languages with stop prevocing besides Catalan (see data for Portuguese in Pape and Jesus 2011) suggests that the patterns of voicing assimilation in consonant clusters in those languages are less homogeneous than previously thought.

The voicing data reported in the present investigation disagrees with the Catalan phonological voicing assimilation rule in other respects. On the one hand, at least for some speakers, variations in C1 voicing have been found to proceed largely independently of changes in C3 voicing in CC/C sequences, i.e., C1 remains most insensitive to voicing differences that may occur during the production of syllable-initial consonants differing in manner of articulation. This finding may be taken to indicate that, for the subjects in question, regressive voicing assimilation in Catalan CC/C sequences is a short-range process affecting immediately adjacent but not distant syllable-final obstruents in line with the demanding articulatory and aerodynamic constraints involved in the production of three consonant in succession. It could parallel data on lingual coarticulation revealing that anticipatory effects decrease over time (Hardcastle and Hewlett 1999).

Another disagreement between the voicing adaptation scenario in consonant clusters reported in this paper and the Catalan voicing assimilation rule concerns the voicing relationship between C1 and C2 in CC/C sequences. In our study, progressive voicing effects were found to operate from C1 onto C2 for all speakers (i.e., an obstructed C2 turned out to exhibit more voicing after a sonorant than after an obstruent), which indicates that voicing coarticulation is prone to involve phonetic segments which belong to the same syllabic unit. Other phonetic processes have been reported to also apply intrasyllabically rather than across a syllable boundary such as stop insertion between C1 and C2 in the sequence /ns/ in American English and Dutch (Warner and Weber 2002, Yoo and Blankenhang 2003).

The presence of speakers showing less than about 40% voicing during syllable-final obstruents was also unexpected since, in principle, subjects could put into action specific maneuvers (e.g., cavity enlargement, nasal or oral leakage) in order to facilitate the presence of voicing throughout the consonant clusters under investigation. Moreover, a positive relationship between speaker-dependent degrees of voicing for syllable-final obstruents and syllable-initial voiced consonants suggests that regressive voicing in C1/C and CC/C sequences is largely conditioned by low-level demands imposed on the speech production system.

The Catalan voicing data reported in this paper indicate that the application of the regressive voicing assimilation process in consonant clusters is subject to principles similar to those operating on lingual coarticulation: the size and temporal extent of coarticulation vary with the articulatory and aerodynamic requirements involved in the production of specific target obstruents, as well as with their position within the consonant cluster, cluster complexity and the manner of articulation of the syllable-initial voiced consonant. Regarding the latter aspect, obstructed voicing turned out to be much less than expected before nasals and laterals due presumably to production and perceptual factors. The failure of sonorants to act as regressive voicing triggers in other languages besides Catalan (e.g., Russian) appears to be related to this finding.

The present contribution to the voicing assimilation literature shows that formal representations of assimilatory phenomena based on auditory perception do not always correspond to the production mechanisms performed by speakers. In a good number of cases, these representations need to be redesigned in the light of production data and should incorporate information about phonetic detail.

5. Acknowledgements

This research was supported by project FFII2009-09339 of the Spanish Ministry of Science and Innovation and FEDER, an ICREA Academia award, and project 2009SGR3 of the Catalan government.

6. References


A Question of Scope? Direct Comparison of Clear and In-Focus Speech Productions

Melissa A. Redford1, Jessica N. Stine1, Eric Vatikiotis-Bateson2

1University of Oregon, USA
2University of British Columbia, Canada
redford@uoregon.edu, jstine@uoregon.edu, evb@mail.ubc.ca

Abstract

Parallels in the production of clear speech and words under prosodic focus suggest that both may be realized in the same way: as hyper-articulated speech. To directly investigate this possibility, school-aged children and college-aged adults produced target words in a default conversational style, a clear speech style, and with prosodic focus. The results were that children and adults both produced target vowels more distinctly and with greater mouth opening in the clear speech and prosodic focus conditions than in the default condition. Whereas the temporal scope of production changes varied as a function of condition in adults’ speech, there was no evidence of this in children’s speech.

Keywords: speech style, prosody, acquisition

1. Introduction

Clear speech has been characterized as hyper-articulated based on acoustic and perceptual findings of increased phonemic vowel contrast relative to a default conversational or casual speech style (Lindblom, 1990; Johnson et al., 1993). Empathic stress due to prosodic focus has been similarly characterized based on acoustic and kinematic evidence that suggests more extreme vowel articulation for in-focus productions (de Jong, 1995). A clear speech style is also typically characterized by slower articulation rates (Picheny et al., 1986), just as prosodic focus is associated with lengthening (i.e., slowing; Turk & White, 1999). In spite of the parallels, clear speech and in-focus productions have yet to be directly compared. The current study makes this comparison to test the hypothesis that clear speech and in-focus productions are realized in the same way: the difference is only in the temporal scope of hyper-articulation, which is broader in clear speech. Because a single production strategy deployed for different linguistic ends would have advantages for acquisition, we chose to investigate the hypothesis in both child and adult speech.

School-aged children and college-aged adults produced target items embedded in meaningful sentences in a default conversational style, a clear speech style, and with prosodic focus. Vowel quality, movement of the lip-jaw complex, and durations associated with target word productions were analyzed as a function of speaking condition and age. The predictions were that phonemic vowel contrasts and peak displacements would be the same in the clear speech and in-focus conditions, and larger relative to the default condition. This prediction is consistent with the characterization of both clear speech and prosodic focus as hyper-articulated speech. Acoustic and kinematic durations of the word onset and vowel were expected to be longer relative to the total duration of the target word for in-focus and default productions than for clear speech productions, where an overall slowing effect was expected to lengthen word offsets. This prediction is consistent with the hypothesis that hyper-articulation has a broad temporal scope in clear speech. It is also consistent with the finding that the syllabic onset and nucleus are disproportionately affected by accentual lengthening in target words produced with prosodic focus (Turk & White, 1999). The effects of condition were expected to be the same in child and adult speech only if children control the settings that render hyper-articulated speech, and can appropriately vary the temporal scope over which these settings apply. This situation was deemed unlikely given evidence of prolonged speech motor control development (see, e.g., Walsh & Smith, 2002).

2. Methods

2.1. Participants

Sixteen native speakers of the standard west coast variety of American English participated in the study. Eight speakers were children (4 female), ranging in age from 7;2 to 7;8 with a mean age of 7;5. Children were in 2nd grade at the time of study and reading at grade level. The other 8 speakers (4 female) were college-aged adults.

2.2. Stimuli

The stimuli consisted of 16 meaningful sentences, each with a target word that occurred either early (subject position) or late (object position) in the sentence. All target words were adjectives that began with a bilabial stop, all modified a subsequent noun (bag or boots), and each contained an initial stressed syllable with 1 of 4 monophthongal American-English vowels (/i/, /ɛ/, /a/, or /ɑ/). Each vowel was represented by 2 target words: beaded/peach, matte/black, modern/Prada, blue/puce. Sentences were presented on cue cards. Font styling was varied as an additional cue to speech style. The cards were shuffled to create different random orders.

Table 1: Example stimuli.

<table>
<thead>
<tr>
<th>Style</th>
<th>Position</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>Early</td>
<td>The beaded boots were brand new.</td>
</tr>
<tr>
<td></td>
<td>Late</td>
<td>Mickey wanted the beaded boots.</td>
</tr>
<tr>
<td>Clear</td>
<td>Early</td>
<td>THE BLACK BOOTS WERE BARELY USED.</td>
</tr>
<tr>
<td></td>
<td>Late</td>
<td>MATTHEW WANTED THE BLACK BOOTS.</td>
</tr>
<tr>
<td>In-focus</td>
<td>Early</td>
<td>The BLUE boots were barely used.</td>
</tr>
<tr>
<td></td>
<td>Late</td>
<td>Billy wanted the BLUE boots.</td>
</tr>
</tbody>
</table>

2.3. Procedure

Lip-jaw movement was measured by tracking changes in mouth shape. To do this, 3 blue dots were placed at the vermillion border of the upper and lower lips and 1 at each corner of the mouth (top panel, Figure 1). Speakers were
seated against a blue backdrop and in front of a video camera and music stand, which displayed the stimuli. The experimenter familiarized child speakers to the sentences by reading each of the cards out loud and with the child. All speakers quickly learned the sentences, which were similar in form (see, e.g., Table 1). After the initial familiarization phase, the experimenter stood to the side of the video camera. She interacted with the speaker throughout the experiment to help them render fluent sentence productions in the desired style.

Speaking style was blocked by condition. Sentences were first produced in a default conversational style, then in a clear speech style, and finally with the target adjective in focus. All speakers immediately and fluently replied, “What did you say?” to which the speaker would appropriately reply, “No, Billy wanted the BLUE boots.” Quick, quiet oral readings of each sentence were given during the task itself if a child needed help to fluently read/speak a sentence. Sentences were repeated if a first rendition was halting or non-fluent. Audio-video recording using a SONY DCR-PC101 camera captured lip-jaw movement. Higher-quality audio was recorded separately to a Marantz PMD660 using a Shure ULXS4 standard wireless receiver and a lavaliere microphone, located in a fixed position a few inches from the speaker’s mouth.

2.4. Measurement

Acoustic measurements were made in Praat (Boersma & Weenink, 2011). Utterances were displayed as an oscillogram and spectrogram. Target word durations were extracted based on audio-visual inspection of the waveform for cues to the stop closures that delimited word onsets and offsets. The duration of the initial C(C)V sequence was also taken, with vowel offsets easily identified by amplitude and spectral changes associated with closure for the post-vocalic consonant, all of which were obstruct consonants. Formant measures were extracted at the midpoint of F2 during its steadiest portion. Once acquired, formant values were converted from Hertz to Bark (z) using the formula proposed in Traunmüller (1990). Formant values were then normalized across speakers using a Bark difference metric (e.g., Syrdal and Gopal, 1986); specifically, z3 – z1 for information regarding degree of vocal tract stricture, and z3 – z2 for information regarding tongue advancement.

Mouth shape was extracted by exporting acquired color video as an image sequence at 30 frames per second. Barbosa and Vatikiotis-Bateson’s (2006) algorithm was then used to track the position of the blue dots (i.e., markers) placed around the speaker’s mouth. Position was defined according to the vertical and horizontal dimensions of the frame, measured in pixels. Movement was defined by the change in position of the markers from frame to frame. The analyses presented here focus only on opening during target word production, and thus on the difference between the medial upper and lower lip markers in the vertical (y) dimension of the 2D space (bottom panel, Figure 1). The frames associated with target word production were identified with reference to word-onset bilabial articulations: target word onset was defined as the first frame in which full bilabial closure was achieved, and target word offset as the frame preceding the next event of full bilabial closure (top panel, Figure 1).

2.5. Analyses

Linear mixed effects modeling was used to determine the effect of speaking condition (default, clear, in-focus) and age (child vs. adult) on the measures we use to assess hyper-articulation and its scope during target word production. Target word position (early vs. late) and vowel (/i/, /e/, /æ/, or /u/), when relevant, were included as additional fixed effects in the analyses. Speaker and word, when relevant, were treated as random effects. Measures of hyper-articulation were vowel space size and maximum opening. Vowel space size was calculated as the perimeter of the quadrilateral defined by the 4 target vowels in the normalized formant space. Measures of scope were the relative duration of the initial C(C)V sequence to overall target word duration and the relative duration of the opening movement (time to maximum opening) expressed as a proportion of the open-close cycle (or total movement, if more than one cycle) associated with target word production. All results are reported with estimated denominator degrees of freedom rounded to the nearest whole number.

3. Results

3.1. Hyper-articulation

The first set of analyses investigated the effects of condition and age group on vowel contrastiveness to test the hypothesis that clear and in-focus productions are both realized as hyper-articulated speech. Consistent with this hypothesis, the analysis of perimeter values indicated that the vowel space was larger for clear and in-focus productions of the target words than for default productions [F(2,70) = 8.97, p < .001], as shown in Figure 2. The analysis also indicated a significant effect of position within the phrase [F(1,70) = 5.87, p = .018] such that the vowel space perimeter was slightly larger when target words modified the object noun than when they modified the subject noun. There was no significant effect of age group on the perimeter values calculated in normalized F1 × F2 space. In addition, there were no significant interactions between any of the fixed effects, indicating a consistent effect of condition regardless of the speakers’ age or target word position within the sentence.
To investigate the extent to which individual vowels were articulated differently as a function of condition, analyses were also conducted on the normed F1 and F2 values associated with each vowel. These analyses indicated a significant effect of condition on the normalized F1 values for /i/ [F(2, 148) = 3.62, p = .029], and on the normalized F2 values for /a/ [F(2, 144) = 7.27, p = .001] and /u/ [F(2, 149) = 11.87, p < .001]. There was also a significant effect of age on the normalized F2 values for /i/ [F(1, 14) = 15.27, p = .002]: children produced /i/ with higher F2 values (bark distance from F3 was smaller) than adults. Post hoc tests revealed no significant differences between clear and in-focus productions of /i/, /a/, or /u/.

The analyses on maximum opening produced similar results to those on vowel quality. Consistent with the hypothesis of hyper-articulation, clear and in-focus productions of the target words resulted in greater maximum opening than default productions [F(2, 529) = 103.26, p < .001]. Not surprisingly, maximum opening also varied systematically with vowel [F(3, 101) = 131.36, p < .001]. The interaction between condition and vowel was also significant [F(6, 530) = 5.05, p < .001]. Analyses within each vowel nonetheless indicated that production varied systematically with condition regardless of the vowel in the target word (/i/, F(2, 140) = 37.28, p < .001; /æ/, F(2, 130) = 37.50, p < .001; /a/, F(2, 126) = 20.96, p < .001; /u/, F(2, 139) = 14.80, p < .001]. Similarly, post hoc comparisons indicated larger opening values for clear and in-focus productions than for default productions. This was true for all target words except those with the high back vowel, where only clear speech productions were associated with significantly more opening than default speech productions.

The condition by age interaction on maximum mouth opening was also significant [F(2, 529) = 6.98, p = .001], even though the simple effect of age was not. When the analysis was split by speakers’ age, production was still found to vary systematically with condition [child, F(2, 313) = 16.16, p < .001; adult, F(2, 320) = 91.80, p < .001]. Inspection of mean differences suggest that the interaction was due to the finding that children produced target words in the clear condition with somewhat greater opening values than those in the in-focus condition, and vice versa, for the adults (Figure 3). Variance in child productions was such, however, that the difference between clear and in-focus productions was only significant for the adults [mean difference = 2.57, p = .048].

3.2. Scope

The next set of analyses investigated the effects of condition and age group on the relative time devoted to articulation of the target word onset + stressed vowel sequence. The goal was to address the question of scope differences in clear and in-focus productions. The results were as follows.

The relative acoustic duration of the initial onset+vowel sequence in the target word varied systematically by condition [F(2, 607) = 28.45, p < .001] and, of course, by vowel [F(3, 120) = 50.14, p < .001]. The condition by vowel interaction was also significant [F(6, 607) = 3.39, p = .003], but within-vowel analyses nonetheless indicated that production varied systematically in spite of the interaction [i; F(2, 150) = 9.88, p < .001; /æ/; F(2, 155) = 6.27, p = .002; /a/; F(2, 150) = 14.70, p < .001; /u/; F(2, 152) = 5.55, p = .006].

Although the effect of age was not significant in the overall analysis, the interaction between condition and age was [F(2, 607) = 13.50, p < .003], as shown in Figure 4. Post hoc mean comparisons confirmed the differences evident in the figure: children produced longer onset+vowel sequences when the target words were under prosodic focus (mean difference from clear = .02, p = .017; mean difference from default = .03, p < .001) than when they were spoken in a clear or default speech style, but children did not differentiate the time to articulate onset+vowel sequences in the clear and default speech styles. Adults also produced the longest onset+vowel sequences in target words under prosodic focus (mean difference from clear = .05, p < .001; mean difference from default = .02, p = .004), but the clear speech onset+vowel sequences were shorter relative to the duration of the whole word than the default speech onset+vowel sequences (mean difference from default = .03, p < .001).

The differences between child and adult productions were even more striking in the analyses of the relative opening duration. Again, there were significant effects of condition [F(2, 539) = 4.05, p = .018] and vowel [F(3, 107) = 4.09, p = .009] on the duration of the opening phase, but no significant interaction between them. There was also a significant effect of age [F(1, 14) = 22.71, p < .001]: the opening and closing phases in adults’ productions of the target words were more...
was not effected by speaking condition). Also, peak displacement was greater in the clear speech and in-focus conditions compared to the default condition. None of these results varied with the speakers’ age, suggesting that 7-year-old children have acquired adult-like control over the parameter settings relevant for hyper-articulation.

The results on time spent in onset-vowel articulation were consistent with the hypothesis that it is the temporal scope of supraglottal articulatory changes that distinguishes clear speech from in-focus production. As expected, onset-vowel articulation was shorter relative to whole word articulation for clear speech productions compared to in-focus or default productions, consistent with a broader scope of change in clear speech relative to in-focus speech. However, unlike the results on hyper-articulation, the temporal results varied significantly with speakers’ age. Children spent more time overall on the onset-vowel sequence than adults, probably because they were less adept at the articulation of the complex onsets in the target word stimuli. More intriguingly, children’s emphasis on opening was observed regardless of speaking condition. This result could indicate less fine-grained temporal control over different modes of production in children. More generally, it suggests that the development of timing control is more protracted than control over parameters such as stiffness that underlie articulatory changes associated with hyper-articulation.

5. Acknowledgements

This research was supported by award #R01HD061458 from NICHD. The authors thank Faire Holliday for her significant help with data processing.

6. References


Distractor effects on response times in fricative production

Kevin D. Roon1,2, Eugen Klein3, Adamantios I. Gafos2,3
1CUNY Graduate Center, USA
2Haskins Laboratories, USA
3Universität Potsdam, Germany
kroon@gc.cuny.edu, euklein@uni-potsdam.de, gafos@uni-potsdam.de

Abstract

Studies have shown that response times in CV-syllable naming can be modulated by distractors that speakers hear while preparing responses, specifically by response-distractor congruency of articulator or voicing (Roon 2013). Roon and Gafos (2013) propose a dynamical model of phonological planning in which these RT modulations arise due to properties of a distractor exciting or inhibiting the ongoing response planning. The model makes predictions for types of response-distractor pairs beyond those tested to date (all stop-initial), including that similar RT modulations should be found for fricatives. We report results of an experiment where fricative distractors that were heterorganic from responses slowed down RTs, as did homorganic distractors mismatching in constriction location. These results validate the design of the model and offer evidence for its generality and usefulness in understanding the nature of representations and principles implicated in perceptuo-motor effects.

Keywords: phonological planning, dynamical modeling, perceptuo-motor effects, fricative production

1. Introduction

Roon (2013) presents experimental results from a response-distractor task in which participants repeatedly produced one of two nonsense CV syllables when prompted by a visual cue. After cue presentation, participants heard various distractors over headphones. Modulations in response times (RTs) are interpreted to reflect an influence of properties of the distractors on the production of the intended utterances, and are thus dubbed “perceptuo-motor effects” (Galantucci, Fowler, and Goldstein 2009). Two experiments reported in Roon (2013) show independent perceptuo-motor effects of articulator and voicing. RTs for trials where the distractor and response matched in voicing (but mismatched in articulator, e.g., ta-pa, Congruent condition of Figure 1A) or matched in articulator (but mismatched in voicing, e.g., ta-da, Congruent condition of Figure 1B) were faster than on trials where they mismatched in both voicing and articulator (e.g., ta-ba, both Incongruent conditions). Galantucci et al. (2009) used the same experimental task to elicit RT effects based on phoneme identity, but the experiments presented in Roon (2013) provide the first clear RT effects due to representations of a finer-grained level than phonemic identity (i.e., articulator and voicing). In addition to the congruency effects, there was an unexpected result in the experiments in Roon (2013). RTs were much slower in the Voicing experiment where the participants knew the articulator of their response but not its voicing (e.g., ta or da) compared to the Articulator experiment where they knew the voicing but not the articulator (e.g., ta or ka). This cross-experiment difference (52 ms) was independent of distractor type, including trials with no distractor (Figure 1A vs. 1B).

In Roon (2013) and Roon and Gafos (2013), a dynamical model of phonological planning (Figure 2) is proposed in which the specific production parameter values of an utterance are set. The model adopts the representational scheme of Articulatory Phonology (Brown and Goldstein 1986, et seq.) and defines one planning field for each tract variable that needs to be set (e.g., Tongue Tip Constriction Location), including one for voicing. The model accounts for both the congruency effects and the cross-experiment differences. The congruency effects are due to the parameter values of a perceived distractor serving obligatorily as input to the ongoing planning of an intended utterance. Compatible input values introduce excitation, while incompatible parameter values introduce inhibition. Since congruent responses and distractors in these experiments always differed along at least one parameter, RTs in the Congruent conditions were slower than a neutral condition where the distractor was a simple nonspeech tone. Inhibition can be either cross-field (e.g., the Lower Lip Constriction Location planning field inhibiting the Tongue Tip Constriction Location field, indicated by the double-headed arrows in Figure 2) or within-field (e.g., voiceless vs. voiced in the Voicing field). The cross-experiment RT differences arise in the model from the differences between trials where participants anticipate two possible responses that are defined as values within one planning field (e.g., voiced or voiceless) vs. trials involving two different planning fields (e.g., a lower-lip or tongue-tip voiceless stop). Due to the within-field dynamics (defined

Figure 1: Results from Roon (2013) showing effects of voicing (A) and articulator (B). Congruent Distractor condition was significantly different from Tone and Incongruent conditions in both experiments.

Figure 2: Expanded model of phonological planning proposed by Roon and Gafos (2013).
using Dynamic Field Theory, Erlhagen and Schöner 2002), the within-field anticipation (“Task Knowledge” in Figure 2) results in a more inhibited state of the planning field at the beginning of each trial compared to cross-field anticipation. This task knowledge refers to knowing that the response is always a tongue tip consonant vs. knowing that it is always a voiceless consonant. The different trial-initial states of the fields predict longer RTs in an experiment where voicing is unknown than when the articulator is unknown (see Roon 2013 for further details).

The model makes specific predictions about other response-distractor combinations. First, all responses and distractors used by Galantucci et al. (2009) and Roon (2013) were stop-initial syllables. Cross-field inhibition in the model (Figure 2) does not depend on constriction degree. Therefore, RTs should be longer than in the tone condition when the initial consonants of the distractors and responses match in voicing but differ in articulator (i.e., are heterorganic) for other manners of articulation, e.g., fricatives. Such response-distractor pairing is analogous to the Congruent condition in the Voicing experiment (Figure 1A), where the stop-initial responses and distractors matched in voicing but differed in articulator. Second, RTs should be longer than in the tone condition when distractors and responses match in voicing and articulator (i.e., are homorganic) but differ in constriction location, due to the within-field inhibition introduced by incompatible inputs. This response-distractor pairing is analogous to the Congruent condition from the Articulator experiment (Figure 1B), where responses and distractors matched in articulator but differed in voicing. Third, RTs in an experiment where the two potential responses on a trial involve two different planning fields should be shorter than in an experiment when the two responses involve one planning field, regardless of which field is involved. We present below results from an experiment that tested these three predictions.

2. Experiment

38 undergraduate students from New York University participated in the experiment. These were the same participants from the Voicing experiment reported in Roon (2013). All participants were monolingual native speakers of American English.

2.1. Design and procedure

Participants wore headphones while seated in front of a computer screen in a sound-attenuated booth at the Phonetics and Experimental Phonology Lab of New York University. In a practice block, participants were instructed that they would repeatedly see one of two cues (## or ==), and that their response should be fa for ## and fa for ==. They were told to respond as quickly as possible, but not so quickly that they made too many mistakes. They were also told they would hear various sounds over the headphones while performing the task, which they should ignore.

Table 1: Responses and distractors, with corresponding experimental conditions.

<table>
<thead>
<tr>
<th>Response</th>
<th>Distractor</th>
<th>sa</th>
<th>ha</th>
<th>tone</th>
<th>non</th>
</tr>
</thead>
<tbody>
<tr>
<td>fa</td>
<td>(Heterorganic)</td>
<td>Heterorganic</td>
<td>Tone</td>
<td>NoDistr</td>
<td></td>
</tr>
<tr>
<td>fa</td>
<td>Homorganic</td>
<td>Heterorganic</td>
<td>Tone</td>
<td>NoDistr</td>
<td></td>
</tr>
</tbody>
</table>

The response-distractor pairs are shown in Table 1. The time between the onset of the visual cue indicating the required response and the start of the audio distractor was either 100, 200, or 300 ms, resulting in three different Stimulus Onset Asynchronies (SOA). A short (15 ms) non-speech “marker” tone was played concurrently with the onset of the visual cue to mark that event in the acoustic record. Each response-distractor pair was presented 14 times at each of the three SOAs using ePrime 2.0 (Psychology Software Tools, Inc., Pittsburgh, PA), yielding 252 trials. Each response was also prompted 14 times with no distractor, adding 28 more trials, for a total of 280 trials per participant. 38 participants yielded 10640 trials in the experiment. 362 trials were excluded from analyses due to errors introduced by the computer during stimulus presentation.

The response fa with the distractor sa corresponds to the homorganic condition, since the fricative in both syllables involves the same primary articulator, that is, the tongue-tip articulator (but differ in constriction location, post-alveolar vs. alveolar). ha distractors are modeled as having a constriction degree of Critical for the Glottis planning field, which inhibits and is inhibited by the other articulator fields (here, Tongue Tip and Lower Lip). All other response-distractor pairs with speech distractors were heterorganic, as the response and distractor were both voiceless but never matched in primary articulator.

2.2. Data analyses

Two RTs were measured by hand for each trial based on inspection of the spectrogram and waveform: the time lag between the marker tone and the onset of discernible friction of the response, and the time lag between the marker tone and the first amplitude maximum indexing the onset of phonation for the vowel of the response.

Statistical analyses were performed by creating linear mixed-effects models with log RT as the dependent variable (however, all figures show RTs in milliseconds for more intuitive interpretation). Participant and Item (response syllable) were always included as random factors. Statistical models included fixed-effect predictors that are known to affect RTs (see Roon 2013): RT of trial_{j}, is a strong predictor of the RT for trial_{j}, an incorrect response on trial, predicts a slower RT on trial_{j+1}, and RTs on a trial tend to be faster when the required response is the same as the response on the previous trial. Participant-specific effects of trial number were included to account for longitudinal differences in RTs across participants. All of the above predictors were centered (Gelman and Hill 2007) to reduce collinearity. SOA was also included as a predictor, as RTs increase monotonically with SOA (Galantucci et al. 2009). Significance was determined by $|r| > 2$ for a given effect. All of these predictors had the anticipated effects and are not reported below.

2.3. Prediction 1: heterorganic effects for fricatives

To test the first prediction of the model, RTs were compared for trials where the distractor was either ha or the tone. The prediction is that RTs should be longer when the distractor was ha, given the cross-field inhibition introduced by the input to an incompatible articulator field (Glottis) by the distractor.

Determining the onset of friction from the acoustic record was much less reliable than identifying the onset of phonation. Therefore, RTs based on phonation onset were used. 699 trials on which the participant produced an incorrect response were excluded. 184 trials on which the response started earlier than 100 ms into the playing of the distractor were excluded, on the grounds that a distractor cannot have any impact on the planning of an utterance when there is
insufficient time to perceive the distractor. 368 trials were excluded due to the response starting later than 750 ms after the onset of the distractor, on the assumption that the participant was inattentive on that trial. If not already excluded, the first trial for each participant was excluded as there was no possible value for the RT of the previous trial (21 trials). Since the purpose of this test was to see specifically whether RTs on \textit{ha}-distractor trials were longer than on tone-distractor trials, trials with no distractor were excluded (888 trials). Trials with a \textit{sa} distractor were excluded (2689 of the remaining trials) because while \textit{sa} is incongruent in articulator with \textit{fa} responses, \textit{sa} shares articulator with \textit{fa} responses.

Figure 3A shows the RTs of the remaining 5429 trials by response within distractor. The mean RT for the \textit{ha} distractor condition was longer than for the tone distractor by 9 ms. A statistical model was created that included the fixed treatment effect of Condition (tone vs. \textit{ha}) in addition to the effects listed in section 2.2. The interaction of Condition and SOA was included, as was a mixed effect of participant-specific Condition to account for the possibility that participants may show varying sensitivity to the Condition effect.

RTs in the \textit{ha} distractor condition were significantly longer than in the Tone condition ($t = 2.29$). The interaction between Condition and SOA was not significant. This result supports the first prediction of the model.

### 2.4. Prediction 2: homorganic effects for fricatives

The second prediction of the model is that RTs for \textit{fa}-response trials should be slower in the \textit{sa}-distractor condition than in the tone-distractor condition. As in the analysis above, trials were excluded in this analysis for incorrect response, for starting too early or too late, for being the first trial of the experiment, or not having a distractor. In addition, trials with \textit{ha} distractors were excluded as were trials with \textit{fa} responses, as these were addressed in the preceding analysis. Figure 3B shows that for the remaining 2758 trials, the mean RT for trials with a \textit{sa} distractor was 8 ms longer than for trials with the tone distractor. Log RTs of those trials were analyzed in a statistical model having the same predictors as those in section 2.3, with the exception of the random effect of Item, which was not included since all responses were \textit{fa}. Results of the statistical model show that RTs in the \textit{sa}-distractor condition were not significantly slower than those in the tone-distractor condition ($t = 0.35$). The effect size in this \textit{sa}-distractor (homorganic) case was nearly the same as in the preceding \textit{ha}-distractor (heterorganic) case. However, there were only half as many trials, suggesting that there may have been a lack of sufficient statistical power. Nevertheless, these data are consistent with this second prediction of the model.

### 2.5. Prediction 3: cross-experiment differences

The last prediction of the model to be tested is that RTs should be faster in the present experiment than in the Voicing experiment from Roon (2013; Figure 1A above), irrespective of distractor conditions. In principle, the two sets of RTs could be compared to show within-participant effects, because the 38 participants are the same across the two experiments. However, their different sets of stimuli render a direct comparison of RTs across the two experiments difficult. Responses in the present experiment were fricative-initial CV syllables for which the onset of aperiodic energy offers a landmark corresponding to the achievement of the oral constriction (albeit somewhat variably). Responses in the Voicing experiment were stop-initial CV syllables (\textit{ta}, \textit{da}, \textit{ka}, or \textit{ga}). Their acoustics allow access to the timing of the release of the vocal tract closure for the stop, but not to the achievement of the closure.

It is therefore not appropriate to compare the RTs shown in Figure 1A with the RTs from the present experiment based on friction onset. The latter would be expected to start much sooner than the former since an earlier landmark is being measured (the achievement of oral constriction instead of its release). Conversely, comparing the RTs from the Voicing experiment with the RTs from the present experiment based on phonation onset is also not appropriate, since the phonation onsets of the responses in the Voicing experiment were later than the release bursts. Conclusive comparison between the two experiments is therefore unfortunately not possible. Nevertheless, two exploratory attempts are presented below.

One way to compare the RTs across the experiments is to estimate the likely closure duration of the stops in the Voicing experiment and shorten the RTs commensurately. Byrd (1993) reports stop closure durations for English based on an acoustic analysis of an extremely large corpus of spontaneous speech, which were: [\textit{t}] = 53 ms, [\textit{d}] = 52 ms, [\textit{k}] = 60 ms, and [\textit{g}] = 54 ms. The corresponding mean closure duration was then subtracted from the RT of each trial of the Voicing experiment as measured from the release burst. Figure 4A compares the mean RTs from the present experiment measured from the friction onset with the mean of the estimated, closure-adjusted RTs from the Voicing experiment. The latter are 39 ms shorter than those in the Voicing experiment, per the prediction of the model.

A second way to estimate a more appropriate RT comparison between the two experiments is to use an acoustic landmark common to each. One such landmark is the onset of phonation for the vowel. VOT was marked in all of the
responses for the Voicing experiment, so phonation onset is calculable simply by adding VOT to the RT reported for each response. However, adjusting RTs in this way must take into account inherent differences in naming latencies based on the initial consonant. Rastle, Croft, Harrington, and Coltheart (2005) used a delayed-naming task to determine intrinsic differences in simple naming latencies for all legal onsets in English, and measured two acoustic landmarks for each syllable from the onset of a go-signal tone: the onset of acoustic energy of any kind, and the acoustic start of the vowel. The RTs Rastle et al. (2005) found for the singleton consonants (before [ə] and [a] only) are schematized in Figure 5. The shortest time to phonation was 331 ms for [d]-initial responses (marked by the red line). Rightmost (red) numbers in Figure 5 indicate the difference between phonation onset times of the vowel in that consonant context compared to the baseline of [d]. With these considerations in mind, RTs for all trials for both the Voicing present experiments were recalculated by subtracting the corresponding phonation onset lag from each trial.

![Diagram](image_url)

Figure 5: CV naming latencies (ms) by initial consonant reported by Rastle et al. (2005). Boxes start at the average acoustic onset for each consonant (leftmost number) and end at phonation onset. Numbers inside boxes show the average frication duration of fricatives and VOT of stops. Rightmost number indicates phonation onset lag vs. [d].

Figure 4B compares the mean RT across all of the trials of the present experiment with all trials from the Voicing experiment. By this measure, RTs in the present experiment are again shorter than in the Voicing experiment, by 25 ms. While inconclusive given the reliance on data from other studies, both attempts to compare RTs across the two experiments seem to support the prediction of the model.

3. Discussion and conclusion

The results from the experiment reported here are consistent with previous results and support the three predictions made by the model proposed in Roon and Gafos (2013) to varying degrees. First, compared to RTs on trials where speakers hear a short non-speech tone, RTs for fricative-initial responses are significantly longer when speakers hear a distractor that matches their intended response in voicing and manner but differs in articulator. This new result shows that the effect of articulator incongruency found by Roon (2013) for stop-initial utterances applies to fricative-initial utterances as well. Second, a mismatch in constriction location between a response and distractor that otherwise match in voicing and articulator (i.e., fa-sa) resulted in slower RTs than on xa-response trials with a non-speech tone distractor, though this difference was not significant. The former effect derives from the model’s cross-field inhibition between the Tongue Tip and Glottis planning fields, while the latter derives from within-field inhibition introduced by incompatible inputs to the Tongue Tip planning field.

The third prediction of the model—that RTs should be shorter in an experiment where participants know the voicing of the response than in one where they know the articulator—could not be tested directly with the acoustic data collected in this experiment. However, RTs adjusted based on previous findings suggest that this prediction may be correct and that it would be worthwhile to investigate it further using a more appropriate experimental method.

The present experimental results provide support for several predictions of the model proposed by Roon and Gafos (2013), providing further evidence for its generality and its usefulness in understanding the representations and principles implicated in perceptuo-motor effects. The results from this study and those from Roon (2013) also have implications for models of speech production, which include little (e.g., Dell 1988) or no (e.g., Roelofs 1999) role for representations at a more fine-grained level than that of the phoneme. The model of Roon and Gafos (2013) crucially relies on the influence of the properties of articulator, voicing, and constriction location of perceived stimuli on the process of speech planning. It is difficult to see how these distractors could modulate production RTs if these levels of representation are assumed not to have a role in speech production.

4. Acknowledgments

KDR gratefully acknowledges support from NSF Grant BCS-0951831, and from NIH Grant DC-002717 to Haskins Laboratories and City University of New York. AIG gratefully acknowledges support from ERC AdG Grant 249440. The authors thank Nikolao Polits for assistance with data labeling.

5. References


The Acoustics of Constriction in a Vocal Tract Model Using 2D Digital Waveguide Modelling

Anocha Rugchatjaroen and David M. Howard

Audio Lab, Department of Electronics, University of York
York, UK

ar718@york.ac.uk, david.howard@york.ac.uk

Abstract

Typically in articulatory speech synthesis, the 3-D shape of a vocal tract for a particular speech sound has been established, for example, by magnetic resonance imaging (MRI), and this is used to model the acoustic output from the tract using numerical methods that operate in either 1, 2 or 3 dimensions. The dimensionality strongly affects the overall computation complexity, which has a direct bearing on the quality of the synthesized speech output. The computational cost of 2-D Digital waveguide modelling makes it a practical technique for real-time synthesis in an average PC at full (20 kHz) audio bandwidth. Thus, a 2-D Digital Waveguide Mesh (DWM) is proposed for this work, which is also commonly used in room acoustic modelling. The constrictions under consideration here include the full vocal tract closure associated with plosives (in English these are the consonants in 'baa', 'pa', 'do', 'to', 'go' and 'coo'); all have an sudden release of acoustic energy when the constriction is released that is known as a 'burst'. The centre frequency of the burst that relates to the vocal tract shape during the plosive closure is analysed as it is an important acoustic cue for consonant perception. In this work, all tract shapes are extracted from MRI recorded data.

Keywords: Fricative Synthesis, 2-D Digital Waveguide Mesh

1. Introduction

In producing human speech, the vocal tract consists of the pharynx, mouth and nose which encompass velum, hard palate, soft palate, tongue, teeth, jaw and lips and their muscles and cartilages. All work to shape the tract and resonate a sequence of different sounds powered by airflow from the lungs. Linguists classify individual sounds based on the position, or place, of the articulation and also the articulation gesture, known as the manner of articulation. Three of these manners involve constrictions in the tract; fricative, plosive and affricate. In fricatives there is a hissing noise due to a turbulent air-flow at the constriction, while in a plosive there is a burst noise as built-up air pressure behind a closure in the tract is suddenly released. An affricate is a concatenation between a plosive and a fricative. The main characteristic of fricative is then the frequency band of hissing noise, whose acoustic properties relate to the volumes and shapes of the cavities in front and behind the constriction; the larger/longer the frontal area, the lower is the resonance, the larger the rear cavity the lower is its anti-resonance. For a plosive, the main acoustic characteristics are the centre frequency of burst, locus frequency and formant transitions.

In previous source-filter modelling research for fricatives, Shadle investigated the physics of friction in the fricative in her research in 1985 (Shadle, The acoustics of fricative consonants, 1985). Two types of friction production for the fricative are named Obstacle and No-obstacle in her report after the obstruction in the tract to jet air. To model them, she used dipole source in her transmission-line model for the obstacle case and just source-tract interaction for no-obstacle case. Her results show significant accuracy for the Obstacle cases which contains /f/, /v/, /s/ and /z/ via spectra comparison with recorded natural speech. In 1989, Badin (Badin, 1989) also reported his measured spectra from his acoustic success source-filter modelling of the vocal tract for /s/ and /f/ and more analysis in friction source using Intra Oral air Pressure (IOP) and Area exponents which is different from Shadle’s which used velocity. In later years, Narayanan and Alwan claimed that source types have to be broken down into monopoles, dipoles and quadrupoles and injected in different places such as dipole at the teeth and monopole at the constriction exit (Jackson, 2000). In 2000, Jackson used inverse filtered coloured white noise from Shadle’s regression curve (Shadle, The acoustics of fricative consonants, 1985) as friction-noise source for his fricative study (Jackson, 2000). In this work, white noise is the only noise source used.

The Digital Waveguide Mesh has typically been used in the room acoustic modelling (Damian Murphy, 2007). A smaller space like a vocal tract, which is about 5 cm wide and 16.5-18 cm long for adults, can be modelled in 1-D, 2-D and 3-D. Mullen, et al. (Jack Mullen, Acoustical Simulation of the human vocal tract using the 1-D and 2-D digital waveguide software model, 2004) has published their results using 1-D and 2-D DWM in 2004. It demonstrated the inherent accuracy of the synthesized vowel formants and bandwidths produced by the waveguide. A few years later, Mullen successfully used 2-D DWM with raised-cosine impedance control to dynamically change the tract shape (Jack Mullen, Real-Time Dynamic Articulations in the 2D Waveguide Mesh Vocal Tract Model, 2007). Recently, Speed, et al., published the characteristic of his 3-D DWM cylindrical vocal tract model, which was able to produce natural sounds in the sub-10 kHz band (Speed, Howard, & Murphy, 2013). The latter required a very dense mesh requiring massive computational resources and these are lessened by using a lower dimensionality (2 rather than 3) and/or a less dense mesh. In this article, we use a 2-D mesh with the advantage of the availability and flexibility of existing boundary modelling and the raised-cosine impedance control.
Mullen’s 2D-DWM vocal tract simulation uses the boundary conditions and impedance alignment to control the tract shape. The cosine function works as a bell-shape distribution function across the tract to align impedances into each vertical waveguide string. His results showed that acoustic discontinuities in the output from a dynamically changing tract shape are removed. Shaping a part of the tract into a very small hole or a closure represents the constriction formed in this work. In this article, we evaluate the constricted tract by using injected source, white noise, which was injected at different places appropriate to the manner of articulation and show the accuracy of the acoustic result. Spectrograms of the synthesized sounds are compared to those of the recorded speech.

Section 2 introduces the system setting. Section 3 describes the experimental results and the conclusion can be found in Section 4.

2. Digital waveguide mesh (DWM) setting

Here, we use a rectilinear mesh for ease of modelling and visualisation. Excitation source pressure passes through the mesh from one to another end, glottis to lips. At each node junction in mesh, the pressure that passes out from the node is equal to junction-pressure minus pressure-in as in equation 1.

\[ p_k^- = p_j^+ - \frac{2\sum_{i=1}^{N} y_i p_i^+}{\sum_{i=1}^{N} y_i} - p_k^+ \]  

For example, at node \( j \), junction-pressure is calculated by dividing a summation of all pressure passing through the node weighted by admittance \( Y \) by all admittance. Then pressure-out at branch \( k \) that is attached to node \( j \) is equal to junction pressure minus pressure-in from the branch \( k \).

In Mullen’s 2D DWM, meshing from glottis to lips is aligned and controlled by cosine impedance function. Equation 2 shows the impedance function for a node at coordinate \( (x, y) \) where \( w \) is the width of the tract (vertical length) where \( x \) indexes the junction node from glottis to lips and \( y \) indexes the junction node along each cross-section.

\[ Z(x,y) = Z_x - (Z_x - Z_{\text{min}}) \left[ 1 + \cos(2\pi\frac{x}{w} - \frac{1}{2}) \right] \]  

The reflection coefficient at boundaries for voiced and voiceless fricative is caused by different termination at the glottis. When the glottis is held open for voiceless fricative, the reflection coefficient at glottis is set to -0.9. For voiced fricative, the glottis vibrates periodically to generate voice source. The model assumes that the glottis is a rigid end. Therefore, reflection coefficient here for this case is set to 0.97. This causes different resonances. Moreover, as we did not focus on source generation but on the filtering in the vocal tract, the naturalness of our results is poor for this manner of articulation.

Extra noise source injection is located using equation 3. Downstream friction source is assumed to be at the very next junction counted from the constriction place. It is injected at junction \( J_{\text{inj}} \):

\[ f_{\text{inj}} = \text{cell}(l_b/(f_s/\sqrt{c})) \]  

Where \( f_s \) is the sampling frequency, \( c \) is the speed of sound, and \( l_b \) is the distance of the constriction from glottis.

3. Experimental results

Figure 1 shows spectrograms of bursts in plosive simulations. It shows that the resonance spread almost over all frequencies for /p/, about 6 kHz for /t/ and 4.5 kHz for /k/. This is in the expected range indicated by (Howard & Damian, 2008) for the alveolar (/t/), where it is expected to be above 4 kHz, and lower for /k/, but not between 1.5 kHz to 4 kHz. However, the results confirm the discussion of vocal tract filter functions in (Johnson, 2008). He mentioned that the spectrum of burst from labial stops has no formant peaks and energy is spread diffusely, and that of alveolar has higher frequency peaks than that of palatal because of the shorter front cavity.

![Figure 1: Spectrogram of the burst from different places of articulation /p/ (left), /t/ (mid) and /k/ (right).](image)

For fricative results, the accuracy of acoustic spectrograms of the synthesized sounds are compared to those of the recorded speech. An example of the similarity of frequency range that resonated in the frontal area of the constricted tract is shown in Figure 2.

![Figure 2: Similarity of the frequency range in synthesized /S/ and recorded /S/.](image)
The comparison shows similarity in frequency response from our propose model to those from real speech. In our work, we did not block the injected noise source from scattering back through the back cavity, while in some theory they believe that the friction source is mainly propagated to the front only due to more pressure pushing in from the back as discussed in (Jackson, 2000).

For more evident, a comparison between resonance frequency from our system with the ideal from Stevens, 1989 mentioned in Johnson, 2008 (Johnson, 2008), is also shown in Figure 3.

![Test resonator comparing to Stevens, 1989](image)

**Figure 3:** Fricative resonance spectra from the proposed system evaluated using the same setting of vocal tract length in Steven, 1989.

The comparison graph shows the same frequency response as appeared in Stevens’. A constriction was set each time at different distances from the mouth: 0.0 cm, 1.5 cm, 2.2 cm, 3.5 cm and 5.9 cm. The response was a peak at 6.5 kHz from constriction at 1.5 cm, a peak at 4 kHz from constriction at 2.2 cm, a peak at 2.65 kHz from constriction at 3.5 cm and two peaks at 1.8 kHz and 5 kHz from constriction at 5.9 cm.

Another impulse response test from a constriction was set after real tract shape from (Speed, Howard, & Murphy, 2013). Figure 4 shows the responses when the tract shape was set from /f/, /θ/, /s/, /S/ and /h/ scanned data.

![Fricative spectra from the propose system using tract shape from recorded MRI data in (Speed, Howard, & Murphy, 2013)](image)

**Figure 4:** Fricative spectra from the propose system using tract shape from recorded MRI data in (Speed, Howard, & Murphy, 2013)

Comparing the results in Figure 4 to the theoretical ones in (Howard & Damian, 2008), the alveolar /s/ in the figure exhibits a peak of energy at around 5.7 kHz and palatal /S/ at around 3.7 kHz when Howard mentioned that the alveolar sound should be at around 4 kHz and the palatal at above 2.5 kHz.

4. Conclusion

The spectrograms show the similarities in spectrograms for dental and alveolar fricatives, which are the two of places of articulation that clearly have a front cavity, and show a wide frequency range similarity of noise for bilabial. This supports the theory that 2D-DWM can simulate the acoustic pressure of turbulent air-flow in the vocal tract. For plosives, the results also show appropriately positioned burst frequencies as the position of the closure changes in the tract. Therefore, this work shows that 2D-DWM can be used to model the air turbulent by setting a constriction in the vocal tract that produces acoustic characteristics that are similar to that of real speech. The use of a 2-D mesh appears to be sufficient for such modelling. Note that in this article, we only focus on the acoustic output resulting from the constriction (an analysis of the formant loci and formant transitions are not included).

5. Acknowledgement

I would like to thank all participants who spent their valuable time in participating my listening test and Dr. Aglaia Foteinou for her kind support.

6. References


Can articulatory gestures be learned by ear? Lip movements by adults imitating a front rounded vowel

Haruka Saito
Waseda University, Japan
qitengyao@gmail.com

Abstract
This study investigated whether Japanese speakers, whose language lacks a front rounded vowel, can infer the rounding gesture when imitating such a vowel on the basis of acoustic information alone. Two measurements were taken from videotaped images: lip aperture (horizontal and vertical) and protrusion (topmost and corner displacement). Results show that the degree of lip aperture was consistently reduced compared to Japanese /i/; while the degree of protrusion were speaker dependent even among most successful imitations. This suggests that (1) these latter articulatory aspects of rounding (protrusion) is more difficult to infer from the acoustics, and (2) auditory-only input leads speakers to adopt idiosyncratic production strategies for non-native sounds.

Keywords: lip rounding, videotaping, second language learning

1. Introduction
A large body of research has shown that speech perception is multimodal: what we hear is based not only on what our ears pick up but also on other perceptual and learned information. One influential source of such information is vision: for instance, incongruent visual images can alter the perception of phonemes (McGurk and McDonald, 1976), and being able to see the speaker’s face enhances phoneme identification in noise (Sumby and Pollack, 1954).

In terms of speech production, however, most proposed models include only audio (and somatosensory) input as a source for learning and refining production, rarely reflecting visual input in an adequate way (Guenther and Vladusich, 2012; Hickok et al., 2011). This situation is also true for second language (L2) learning models (Flege, 1995; Best and Tyler, 2007; but see Hazan et al., 2006), where the production errors of L2 speakers are attributed mainly to their lack of auditory discrimination ability. In these models, there is an implicit or explicit assumption that speakers can “translate” auditory information into articulatory gestures in their speech production system. In other words, speakers are assumed to be able to learn articulatory gestures by ear.

The purpose of this study is to investigate whether this assumption is true in the case of adult speakers who learn to produce a non-native rounded vowel.

1.1. Audio and visual cues of rounded vowels
Since labial rounding is the most visible articulatory gesture in vowels, it is natural to think that perception of roundedness would be degraded by a lack of visual cues. Traumuller and Ohnstrom (2007) demonstrated that the error rate for identification of Swedish rounded vowels by native listeners was as high as 8% among participants sensitive to visual cues in an auditory-only condition, which was much higher than in an audiovisual condition (0.3%). The authors concluded that the auditory cues for roundedness were less reliable than those for other features, like openness.

This gives rise to the question of how deprivation of visual cues for rounded vowels affects their subsequent production. Ménard et al. (2009) found that congenitally blind adult speakers of French produced significantly less acoustic contrast (smaller distances) between vowel phonemes contrasting in terms of roundedness than sighted speakers do. More recently, Ménard et al. (2013) used video and ultrasound to confirm that blind people of French produce a lesser degree of lip protrusion compared to sighted people while making greater compensatory use of the tongue. These studies on blind people suggest that deprivation of visual cues leads speakers to adopt different production strategies, and in turn that the fine-tuning of labial rounding gestures needs something more (such as visual or tactile input) than acoustics only.

1.2. Perception and production of L2 speakers
Given that auditory-only input reduces the use of the lips in subsequent output, as discussed above, this situation might be even more pronounced for L2 speakers because of the nature of their L2 auditory perception. For example, the perceptual assimilation model (PAM: Best and Tyler, 2007) posits that when listeners encounter a non-native sound, they tend to interpret it as (assimilate it to) the perceptually nearest phonological category (or on of the nearest categories) available in their own language (L1). According to this model, if an L2 rounded vowel is assimilated to an unrounded vowel in the L1, it will likely be very difficult for the assimilating listener to infer rounding gestures from the L2 auditory information alone.

Further, the speech learning model (SLM; Flege, 1995) posits that convergence or divergence between L1 and L2 sounds in a speaker’s perceptual phonological space will eventually result in non-nativelike production on that speaker’s part. However, it is still unclear (and has rarely been investigated) how perceptual convergence/divergence affect the actual articulatory aspects of L2 production.

In short, given that fine-grained lip-rounding gestures seem difficult to infer from the acoustics alone even for native speakers and that L2 speakers’ auditory perception is often affected by their L1 phonology, it is intriguing to investigate whether and to what extent L2 speakers can infer the articulatory aspects of rounding “by ear.” This will help us understand how L2 speakers adopt their production strategies to a non-native sound. In the present study, Japanese speakers, whose language lacks a front rounded vowel, were asked to imitate such a vowel, and their lip movements were observed.
2. Method

2.1. Materials

The target vowel was Mandarin high front rounded vowel /y/. Prior to the experiment, five female native speakers of Mandarin were recorded reading five Mandarin vowels (/y/, /i/, /a/, /a../, /u/) aloud ten times, bearing the high level tone of Mandarin. The two tokens with the best sound quality per speaker were chosen; in all, 50 productions (five vowels x five speakers x two tokens) were used as stimuli in both the perceptual similarity task and the imitation task (however, only the data for /y/ are analyzed in this paper). The five native Mandarin speakers (and one extra speaker) also served as a control group for the articulatory analysis described below; videotaping and recording conditions were the same as those for the experimental participants (see procedure).

2.2. Participants

The participants were 25 Japanese speakers (ages 18–36) who had no prior knowledge of Mandarin or any other foreign language with a front rounded vowel. Their knowledge of foreign languages was limited to English, and, for some of them, Spanish (the bulk of the participants were recruited from beginning Spanish classes at Waseda University in Tokyo). All participants reported no history of medical diagnosis of a hearing disorder.

2.3. Procedure

The experiment consisted of three sections: (1) recording of Japanese vowels (as baselines), (2) the perceptual similarity task, and (3) the imitation task.

For the recording section, participants sat on a chair with a headrest (a dentist’s chair, but set upright) and were asked to avoid, as much as possible, moving their heads during recording. With small stickers (diameter 4 mm) on their lips, participants read the five Japanese vowels (/i/, /e/, /a../, /o/, /u/) aloud, 10 times each. Participants were encouraged to pronounce the sounds ‘long and fully’. Only the first five tokens of /i/ for each speaker were analyzed in this paper; thus, only two of the stickers, the one just under the nose and the one on the left edge of the upper lip, are relevant here. Participants were instructed to move their lips back to neutral position before and after each production. Front and left views of participants’ lips were videotaped by two digital video-recorders (Canon iVIS HF MS1) at a rate of 30 fps. The acoustic signal was recorded by a laptop computer with a tabletop condenser microphone (Blue Yeti) at a 22050 Hz sampling rate with 16-bit resolution.

For the perceptual similarity task, using the multiple forced choice experiment function (ExperimentMFC) of Praat version 5.3.68 (Boersma and Weenink, 2005), participants listened to five Mandarin vowels (/y/, /i/, /a../, /x/, /u/) through headphones, and were asked to identify each of them in terms of the nearest of the five vowels of Japanese (/i/, /e/, /a../, /o/, /u/) and to rate their similarity with the Japanese vowels on a seven-point scale (Strange et al., 2004). Stimuli were presented in random order and repeated six times each. Thus, each participant listened to and categorized 60 productions (five speakers x two tokens x six repetitions) per vowel. Only the responses for /y/ are analyzed in this paper.

Finally, in the imitation task, participants again listened to the Mandarin vowels in random order, and imitated each vowel 20 times immediately after hearing it. Videotaping/recording conditions and instructions were the same as for the previous recording section. What was important to the experiment was that in this imitation task and the previous perceptual task, participants were only provided with the acoustics for the vowel sounds, not with any kind of visual, orthographic, or other information for the Mandarin. They were not told that the sounds they were listening to were Mandarin, either. In this paper, first 10 tokens of imitated /y/ were measured.

2.4. Measurement

2.4.1. Articulatory measurement

Vowels were measured at the stable points for lip posture, which were determined by looking at the video-recordings (using Avidemux 2.6.1) and stable points for formants by looking at spectrograms for the audio-recordings. Decisions on where to locate these thresholds were not very difficult, because the vowel productions were prolonged (usually above 600 ms) and stable for most tokens. After determining the stable points, front and side views of productions at those points were saved as BMP image files.

To observe the degree of lip aperture, two measurements, namely those of horizontal and vertical interlabial lengths, were taken from the frontal images (Mayr, 2010) (Fig. 1a). Vertical length was measured at the midpoint of the upper and lower lips.

Figure 1: Measurements of (a) lip aperture (b) protrusion.

Protrusion was measured as the degree of displacement from the speaker’s natural lip position. Two measurements were taken, namely, topmost displacement and corner displacement (Fig. 1b). The procedure for measuring topmost displacement was as follows. (1) Save (again as a BMP file) the side view of the natural lip position just before the participant starts to produce each vowel. (2) Overlay the images of natural and open lip posture for each production (using Photoshop Elements 12), moving and rotating layers manually in order to correct for head movement. The stickers under the nose and on the nasal bridge were used as references when overlapping. (3) Draw a reference line between the sticker under the nose and the small dent in the chin (the labiomental crease or hollow) on the images of natural lip position. (4) Draw a tangent line to the forward curve of the upper lip in natural position, parallel to the reference line. (5) Draw a perpendicular line from the leftmost pixel of the lip in open posture to the tangent line. To measure corner displacement, the last two procedures were different. (4a) Draw a line passing through the sticker at the corner of the upper lip in natural position, parallel to the reference line. (5a) Draw a perpendicular line from the same sticker to the reference line. The lengths of lines (5) and (5a) were taken as the index of topmost and corner displacement. If the displacement was forward (protruded), the length was considered to take a positive value, and if backward (spread), a negative value.

All the measurements were taken in pixels using ImageJ version 1.47 (NIH) and subsequently converted into millimeters using the diameter of the stickers as a reference.
2.4.2. Native listeners’ judgments

To evaluate how well participants imitated the vowel /y/ acoustically, four native listeners of Mandarin (not the same people as the controls) listened in random order to vowels produced by the Japanese participants and by the native controls, identified the vowels in terms of the five vowel phonemes of Mandarin, and rated their canonicity on a seven-point scale.

3. Results

3.1. Perceptual similarity Task

As shown in Table 1, of the participants, eighteen classified Mandarin /v/ as (closest to) Japanese /i/ (range 82%–100%; average 98%), four as Japanese /u/ (range 80%–100%; average 92%), and another three participants as “both” /i/ and /u/. Since I wanted to focus on analyses by participants who would auditorily assimilate an L2 rounded vowel to an L1 unrounded vowel (the majority of participants in this study, as shown by the results just above), the data from the 18 participants who chose /i/ were retained and those of the other 7 participants excluded from the subsequent articulatory analysis. The average similarity score of 3.5 assigned to Mandarin /y/ in relation to Japanese /i/ was significantly low compared to the average score of 6.0 for Mandarin /u/ as Japanese /i/ by the same participants (not shown in Table 1).

Table 1: Perceptual assimilation of /y/ to Japanese vowels.

<table>
<thead>
<tr>
<th>Assimilated Category</th>
<th>Number of Participants</th>
<th>Average %</th>
<th>Average similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>18</td>
<td>98%</td>
<td>55</td>
</tr>
<tr>
<td>u</td>
<td>4</td>
<td>92%</td>
<td>43</td>
</tr>
<tr>
<td>i/u</td>
<td>3</td>
<td>48%</td>
<td>3.2</td>
</tr>
</tbody>
</table>

3.2. Articulatory analysis

3.2.1. Horizontal and vertical lip aperture

The upper panel of Figure 2 shows the average horizontal length of the interlabial space for Japanese /i/ and imitated /y/ by the Japanese speakers and Mandarin /y/ by Mandarin speakers. Japanese /i/ was produced with the largest horizontal aperture and Mandarin /y/ with the smallest. Imitated /y/ fell between the two—more contracted than /i/ and less than native Mandarin /y/. A wilcoxon signed-rank test was performed using SPSS 21 and confirmed that there were significant differences between Japanese /i/ and imitated /y/. Vertical length shown in the lower panel of Figure 2 presented a similar tendency. A wilcoxon signed-rank test also revealed significant differences between Japanese /i/ and imitated /y/. However, note that variance in vertical length of imitated /y/ was much larger than horizontal length.

3.2.2. Protrusion

The left panel of Figure 3 visualizes the degree of topmost displacement for /y/ produced by each speaker. Five of six Mandarin speakers showed large displacement forward (5.8 mm on average); 15 of 18 Japanese speakers also showed forward movement, but the degree of displacement was smaller (1–2 mm on average) than native speakers. However, closer inspection of the image files suggests that the degree of topmost displacement might be affected by vertical lip aperture or deformation of the upper lip itself, since the data often does not fit well with the impressionistic degree of protrusion. Thus, degree of corner displacement (Figure 3, right panel) might be a better index of protrusion than topmost displacement is. While Mandarin speakers clearly pushed the corners of their upper lips forward (Figure 4a), Japanese speakers showed either forward displacement (eight speakers) or backward displacement (10 speakers) to varying extents. Figure 4 shows examples of side views of lip posture, variously reflecting forward corner displacement coupled with large vertical aperture (b), forward displacement with small vertical aperture (c), backward displacement with large vertical aperture (d), and backward displacement with small vertical aperture (c).

3.2.3. Analysis of successful imitations

Native listeners’ judgments revealed that imitated /y/ produced by Japanese speakers was heard as the intended vowel 41% of the time on average; and even when it was classified correctly, native listeners often found it not so accurate (similarity score: 3.5 on average). However, some tokens (13 of 180) were more successful, and were classified as /y/ above 75% of the time with a similarity score of above 5.0 (this threshold was determined based on the lowest score of native speakers’ production). The articulatory data for these 13 tokens were examined to see whether there were any common features among these successful imitations. Figure 5 plots and vertical length of interlabial space for those 13 successful tokens. As shown in the figure, there were large variations in these two parameters: the degree of corner displacement varied across positive (forward) and negative (backward) values, and vertical lengths ranged from large (17 mm) to very small (2 mm). In short, no common feature was found among either the successful or the unsuccessful tokens in terms of vertical aperture and protrusion.
with or without vertical contraction and protrusion), they succeeded equally in producing /y/-like sounds. This can be interpreted as an example of “motor equivalence” (Perkell et al., 1993), though this interpretation should be treated with caution, because L2 speakers often do not have as precise an acoustic target to achieve as native speakers do. In future research, it will be interesting to investigate what kind of additional information (for example, visual input, auditory input gleaned from more precise auditory discrimination), if any, can reduce the variety of learners’ production strategies. In sum, the basic gestural character of lip rounding—the contraction of the interlabial space—is relatively easy to infer from acoustic information, but other aspects of rounding seem difficult to learn by ear alone. Nevertheless, speakers adopt different production strategies to try to reach the acoustic target.

5. References


4. Discussion

The perceptual similarity task revealed that the majority of Japanese speakers in this study assimilated Mandarin rounded /y/ to the Japanese unrounded phoneme /i/, while at the same time their low similarity scores indicated that they were sensible to the phonetic differences between /y/ and /i/. Did they then infer rounding gestures from those differences? The results show that the participants reduced the degree of horizontal and vertical lip aperture in their productions of /y/ as compared to their productions of Japanese /i/, which indicated that they could infer at least this one aspect of the rounding gestures (although vertical aperture showed greater variance). However, as shown in the examples in Figure 4, the degrees of protrusion for /y/ varied greatly across speakers. Some speakers even exhibited backward displacement of the corners of the lips (4d, 4e). This suggests that this articulatory aspects of rounding are more difficult to infer from the acoustics. It is unclear whether this is because of the participants’ L1 background (Japanese is known for its “compression”-type rounding) or perhaps due to some universal characteristic(s) of rounding. However, considering that visually impaired speakers of French produce a lesser degree of protrusion (Ménard et al., 2013) and that the NURSE vowel in South Wales English apparently has rounding gestures but in fact its F2 and F3 are not reduced acoustically (Mayr, 2010), it is more probable that the difficulty in inferring from the acoustics, at least in terms of protrusion, is language universal. Moreover, the fact that the successful imitations also varied in their articulatory parameters indicates that auditory-only input leads L2 speakers to adopt idiosyncratic production strategies for L2 sounds. Mandarin speakers’ /y/ showed vertical contraction with protrusion (4a), but even when the Japanese speakers imitated it with other types of gestures (i.e., either

Figure 4: Examples of side views of lip posture.

Figure 5: Articulatory parameters of successful imitations.
Adaptive coding of orofacial and speech actions in motor-somatosensory spaces

Marc Sato¹, Coriandre Vilain¹, Laurent Lamalle², Krystyna Grabski³

¹GIPSA-LAB, Département Parole & Cognition, CNRS & Grenoble Université, Grenoble, France
²Inserm US 17 / UMS IRMaGE, Université Grenoble-Alpes et CHU de Grenoble / CNRS UMS 3552, Unité IRM 3T Recherche, Grenoble, France
³Montreal Neurological Institute, McGill University, Montreal, Canada

marc.sato@gipsa-lab.inpg.fr, coriandre.vilain@gipsa-lab.inpg.fr, Laurent.Lamalle@ujf-grenoble.fr, krystyna.m.grabski@gmail.com

Abstract

To test whether motor simulation might rely on sensory-motor adaptive coding common with those for motor execution, we used a repetition suppression (RS) paradigm while measuring neural activity with sparse sampling fMRI during repeated overt and covert orofacial and speech actions. RS refers to the phenomenon that repeated stimuli motor acts lead to decreased activity in specific neural populations and is associated with enhanced adaptive learning related to the repeated action. Common suppressed neural responses were observed in motor and posterior parietal regions in the achievement of both repeated overt and covert orofacial and speech actions, including the left premotor cortex, the superior parietal cortex and adjacent intraparietal sulcus. Interestingly, reduced activity of the auditory cortex was observed during overt but not covert speech production, a finding likely reflecting a motor rather an auditory imagery strategy by the participants. These results suggest online state coding of both orofacial and speech actions in somatosensory and motor spaces with and without motor behavior and sensory feedback.

Keywords: Motor control, speech production, motor imagery, efference copy, internal model, IMRI, repetition-suppression

1. Introduction

Phonemic perceptuo-motor goals that define successful speech motor acts are thought to be gradually learned by the central nervous system during language acquisition, and stored in the form of an internal forward model. Once mature, the internal forward model can simulate the perceptual consequences of planned motor commands by internal motor-to-sensory simulation and efference copies. These internal motor-to-sensory predictions, generated prior to the actual motor execution and sensory feedback, can assist in online sensory-motor state estimation and corrective motor adjustments in case of discrepancy with actual sensory feedback (e.g., Guenther, 2006; Hickok et al., 2011; Houde and Nagarajan, 2011; Guenther and Vladusich, 2012; Hickok, 2012; Tian and Poeppel, 2012).

In line with this hypothesis, recent brain imaging studies have provided evidence for online motor-to-sensory control mechanisms. Reduced responses of the auditory cortex have been repeatedly reported during self-produced overt speech, compared to utterances recorded and replayed to the subjects (e.g., Numminen and Curio, 1999; Houde et al., 2002; Ford and Mathalon, 2004; Heinks-Maldonado et al., 2006; Christoffels et al., 2007, 2011). Conversely, increased activity during overt speech production has been observed in auditory and somatosensory cortices with altered sensory feedback (e.g., Christoffels et al., 2011; Tourville et al., 2008; Golfinopoulos et al., 2011). The modulation of sensory activity in speech production is thought to be driven by efference copies that assist in online sensory feedback control and sensory-motor state estimation, by tuning sensory phonemic targets to normal sensory feedback or, conversely, by increasing sensitivity to perturbed feedback.

Remarkably, auditory activity and reduced neural responses compared to speech perception has been shown to also occur during covert speech production (e.g., Numminen and Curio, 1999; Tian and Poeppel, 2010, 2013). Since imagery process is known to depend on internal sensory-motor simulation (for reviews, see Jeannerod, 1994, 2001), these results further suggest the existence of internally-generated motor-to-auditory predictions even in the absence of overt motor behavior and sensory feedback.

In keeping with these later findings, we here examined whether these internally-generated motor-to-sensory predictions might also occur during orofacial and speech motor imagery, in the absence of overt motor behavior and sensory feedback monitoring.

In order to test whether motor simulation might also partly rely on motor-to-sensory state estimation common with those for motor execution, we used a repetition-suppression (RS), or adaptation, paradigm while measuring neural activity with sparse sampling fMRI during overt and covert repeated orofacial and speech actions. RS refers to the phenomenon that repeated motor act or stimulus presentation leads to a reduction in BOLD signal and is associated with enhanced adaptive learning and increased processing efficiency (Grill-Spector et al., 2006; Friston, 2012; Henson, 2012). A convergent neurocomputationial interpretation, based on sensory-motor adaptive control, proposes that RS arises from sensory-motor adaptive learning and reduced prediction errors for online state estimation and motor correction (Friston, 2012). From that view, RS is thought to reflect a combination of attention and predictive mechanisms.

In order to examine whether overt production and motor simulation might both induce RS in relation to auditory and somatosensory speech and non-speech adaptive coding, a factorial design was used with the production mode (overt action vs. covert action), the task (silent orofacial action, audible orofacial action, syllable production) and the stimulus repetition (first production, second production) as experimental factors. To this aim, participants were asked to overtly or covertly produce silent orofacial actions (lip protrusion, tongue retraction), audible orofacial actions (kiss, tongue click) and syllables (/pa/, /ha/) in trains of two consecutive trials.

367
2. Methods

2.1. Participants

Twelve healthy adults, native French speakers, participated in the study after giving their informed consent. All were right-handed, had normal or corrected-to-normal vision and reported no history of motor, speaking or hearing disorders.

2.2. Experimental Procedure

Three production tasks were performed overtly in a first functional run and then covertly in a subsequent functional run. The motor tasks consisted of a silent orofacial movement (a lip protrusion or a tongue retraction performed without phonation), an audible orofacial movement (a kiss or a tongue click) or a syllable production (/pa/ or /ta/). Crucially, each identical motor task (e.g., a kiss) was performed in sets of two consecutive trials in order to investigate RS. In addition, a resting condition, without any movement, served as baseline.

For covert actions, the only indication given to the participants was to mentally produce each action without any movement (motor imagery). Importantly, they received no indication regarding possible auditory, visual and/or somatosensory sensations (sensory imagery).

All motor tasks were briefly practiced overtly and covertly just before entering into the scanner. In addition, overt actions were performed in the first functional scan to further help producing covert actions in the second functional scan.

2.3. Data Acquisition

Magnetic resonance images were acquired with a 3T whole-body MR scanner (Philips Achieva TX). Participants were laid in the scanner with head movements minimized with a standard birdcage 32-channel head coil and foam cushions.

Functional images were obtained in two consecutive functional runs using a T2*-weighted, echo-planar imaging (EPI) sequence with whole-brain coverage (TR = 8s, acquisition time = 3s). Each functional scan comprised fifty-three axial slices parallel to the anteroposterior commissural plane acquired in non-interleaved order (3x3mm² in plane resolution with a slice thickness of 3mm without gap). A high-resolution T1-weighted whole-brain structural image was acquired for each participant after the last functional run (MP-RAGE, sagittal volume of 256x224x176mm³ with a 1mm isotropic resolution).

In order to avoid movement artefacts and to reduce acoustic noise, a sparse sampling acquisition was used. This acquisition technique is based on neuropsychological properties of the slowly rising hemodynamic response, which is estimated to occur with a 4–6 s delay in case of orofacial and speech movements (e.g., Grabski et al., 2012a, 2012b, 2013). In the present study, functional scanning therefore occurred only during a fraction of the TR, alternating with silent interscanning periods, where participants produced the required motor task. The time interval between the visual instruction onset and the midpoint of the following functional scan acquisition was of 5s.

In each functional run, the motor tasks were performed in two consecutive trials in a pseudorandom sequence. This RS structure allows measuring changes in BOLD signal for repeated compared to novel performed actions. Altogether, 252 functional scans were therefore acquired (2 runs x (3 conditions x 2 repetitions) + (1 resting baseline)) x 18 trials). In addition, three "dummy" scans at the beginning of each run were added to allow for equilibration of the MRI signal and were removed from the analyses.

2.4. Data Analyses

Data were analyzed using the SPM8 software package (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). Brain activated regions were labeled using the SPM Anatomy toolbox (Eickhoff et al., 2005) and, if a brain region was not assigned or not specified in the SPM Anatomy toolbox, using the Talairach Daemon software (Lancaster et al., 2000).

Data preprocessing: The first three volumes (‘dummy’ scans) were discarded. For each participant, the functional series were first realigned by estimating the six movement parameters of a rigid-body transformation in order to control for head movements between scans. After segmentation of the T1 structural image and coregistration to the mean functional image, all functional images were spatially normalized into standard stereotaxic space of the Montreal Neurological Institute (MNI) using segmentation parameters of the T1 structural image. All functional images were then smoothed using a 8mm full-width at half maximum Gaussian kernel.

Individual and group analyses: Neural activations related to the motor tasks were analyzed using a General Linear Model (Friston et al., 1995), including 12 regressors of interest (one for each repetition of each motor task in both covert and overt conditions) and the six realignment parameters as regressors of non-interest, with the silent trials forming an implicit baseline. The BOLD response for each event was modeled using a single-bin finite impulse response (FIR) basis function spanning the time of acquisition (3s). Before estimation, a high-pass filtering with a cutoff period of 128 s was applied. Beta weights associated with the modeled FIR responses were then computed to fit the observed BOLD signal time course in each voxel for each condition. Individual statistical maps were calculated for each repetition of each task in both covert and overt conditions with the related baseline and subsequently used for group statistics.

At the group-level, a repeated measures analysis of variance was performed, with the production mode (2 levels: overt mode, covert mode), the task (3 levels: silent orofacial movement, audible orofacial movement, syllable production) and the stimulus repetition (2 levels: RS1, RS2) as within-subject factors and the subjects treated as a random factor. Main effects (bidirectional F contrasts), interactions (bidirectional F contrasts) and conjunctions (unidirectional T contrasts) were calculated with a significance level set at $p = .05$, family-wise-error (FWE) corrected at the voxel level with a cluster extent of at least 30 voxels.

3. Results

3.1. Task and production mode

Common activity between silent orofacial actions, audible orofacial actions and syllable productions was observed in a set of largely overlapping brain areas, including the sensory-motor and premotor cortices, the supplementary motor area, the inferior parietal cortex and adjacent parietal operculum, the insular cortex, the basal ganglia and the cerebellum. In addition, silent and audible orofacial actions induced stronger activity than speech actions in several motor areas while, conversely, auditory neural activity was evident during audible orofacial actions and syllable production. These results appear fully consistent with previous brain-imaging studies on
orofacial and speech motor control, with the above-mentioned brain areas classically assigned to motor preparation, execution and regulation loops (e.g., Riecker et al., 2005; Bohland and Guenther, 2006; Grabski et al., 2012, 2013).

Complementing these findings, overlapping activity between the overt and covert modes was observed in the ventral and dorsal part of the left premotor cortex, the posterior part of the left inferior frontal gyrus, the dorsal part of the right premotor cortex, the left insular cortex and the supplementary motor area. The overt mode however induced greater BOLD response in brain areas involved in motor preparation, coordination and execution. These results confirm, and extend to orofacial actions, that mental simulation and actual execution of an action both involve motor preparation and coordination processes (e.g., Jeannerod, 1994, 2001).

Importantly, no auditory activity was observed in the covert mode, a result likely reflecting a motor rather an auditory imagery strategy, with participants instructed to mentally produce each action without indication regarding possible auditory, visual and/or somatosensory sensations.

3.2. RS

Figure 1: Effect of the repetition. Top: Brain regions showing overlapping activity between the two repetitions (conjunctive, unidirectional T contrast); Middle: Brain region showing significant change in activity between the repetitions (main effect, bidirectional F contrast) and related contrast estimates; Bottom: Brain region showing significant change in activity between the repetitions in the overt mode and in the covert mode (bidirectional F contrast).

In both the overt and covert conditions, repeated motor acts led to decreased activity in the left premotor cortex, the superior parietal cortex and adjacent intraparietal sulcus, the left inferior frontal gyrus (pars triangularis) and adjacent prefrontal cortex, the left insular cortex, the supplementary motor area, the middle and posterior cingulate cortices and the left fusiform gyrus (see Figure 1).

Importantly, no RS and auditory activity was observed in the covert mode, even for audible orofacial and speech actions. As previously mentioned, this finding likely reflects a motor rather an auditory imagery strategy here used by the participants.

4. Discussion and conclusion

In the present study, common suppressed neural responses were observed in motor and associative somatosensory regions in the achievement of repeated overt and covert orofacial and speech actions. RS in premotor and posterior parietal cortices has been repeatedly observed in fMRI-adaptation studies on transitive and intransitive manual and orofacial actions (e.g., Dinse et al., 2007; Hamilton and Grafton, 2009; Grabski et al., 2012b). Based on a hierarchical organization of action into distinct levels of control that represents increasingly abstract aspects of the performed behavior (for a recent review, see Grafton and Hamilton, 2007), these studies argue for a role of these premotor and posterior parietal regions in predictive sensory-motor control and action goal coding.

While action goal coding and forward sensory-motor control processes have been primarily studied in the context of upper limb movements, recent models of speech production also postulate that speech motor control is based on a running internal estimate of the dynamic state of orofacial effectors (based on position and velocity estimates) and its effect on sensory outputs in relation to the intended phonemic goal (e.g., Guenther, 2006; Hickok et al., 2011; Houde and Nagarajan, 2011; Guenther and Vladusich, 2012; Hickok, 2012; Tian and Poeppel, 2012). In these models, it is assumed that an internal forward model of the vocal tract, that captures the relationships between speech motor commands and their sensory consequences, is gradually learned by the central nervous system during language acquisition. Once mature and based on an estimate of the dynamic state of the actual vocal tract, the internal forward model can provide accurate predictions of sensory outputs of planned motor commands and, in relation to actual sensory feedback, can make ongoing adjustments to the articulatory trajectory being generated.

By providing evidence for neural adaptive changes through repeated overt silent and audible orofacial actions as well as syllables in the premotor cortex, the superior parietal cortex and the transverse and superior temporal gyri, our results appear in line with these models and an internal state estimate of orofacial speech effectors in somatosensory-motor spaces in relation to the intended action goal.

Importantly, somatosensory-motor adaptive changes were also observed in the same premotor and posterior parietal regions for imagined actions. This finding suggests that somatosensory state estimation processes for speech and non-speech orofacial actions occur even in the absence of overt motor behavior and sensory feedback. This fits well with the view that mental simulation and actual execution of an action share similar motor preparation and coordination processes and that motor imagery depends on internal sensory-motor simulation (for reviews, see Jeannerod, 1994, 2001). This result also suggests that efference copies are automatically sent to the sensory systems even when corrective motor adjustments related to the intended proprioceptive targets are not needed (Tian and Poeppel, 2010, 2012, 2013).

No RS and auditory activity was however observed in the covert mode, even for audible orofacial and speech actions. As
previously mentioned, this finding likely reflects a motor rather than an auditory imagery strategy here used by the participants. From that view, a recent magnetoencephalographic study demonstrates early RS in the auditory cortex in an auditory imagery task (with participants instructed to imagine hearing a cued syllable), whereas early repetition enhancement was found during a motor imagery task (with participants instructed to imagine saying a syllable without any movements; Tian and Poeppel, 2013).

Finally, RS observed in premotor, somatosensory and auditory regions during overt speech production but only in premotor and somatosensory regions during covert speech production also indirectly suggest the existence of distinct internal monitoring. Whether motor, auditory and somatosensory speech signals are integrated in the human brain remains to be further explored.

5. References

Hamilton AF, Grafton ST (2009) Repetition suppression for performed hand gestures revealed by fMRI. Human Brain Mapping 30(9):2898–906
Houde JF, Nagarajan SS (2011) Speech production as state feedback control. Frontiers in Human Neuroscience 5:82
Prosodic adaptation in game interaction between speakers of two Italian varieties

Michelina Savino
Dept. of Education, Psychology, Communication – University of Bari, ITALY
michelina.savino@uniba.it

Abstract

In Bari Italian (BI) and Lecce Italian (LI), unbiased yes-no questions are prototypically realised with two completely different tunes: \(L+H^*L-L\%\) in BI, \((H+)\L^*L-H\%\) in LI, this implying that the functional rise for marking questioning is assigned to the nuclear syllable in BI, and to the boundary in LI. This paper explores if and to what extent prosodic adaptation between BI and LI participants in game interaction involving asking questions is conditioned by their native variety-specific intonation grammar rules. Results show that BI speakers are those who adapt their question intonation to LI partners’ contour, and that this is achieved either a) by having recourse to the allophonic variant of their native rise-fall pattern which has a rising boundary \((L+H^*L-L\%)\), in order to “sound similar” – at least to some extent – to the rising LI question contour yet still preserving the functional rise on the nuclear syllable, or b) by copying the whole \((H+)\L^*L-H\%\) question contour of their LI partners. These results seem to indicate that prosodic adaptation could be – but is not necessarily – conditioned by constraints imposed by the variety-specific intonation grammar rules of the speakers. However, these observations are preliminary and need further research, also in relation to different possible factors influencing the direction of the adaptation.

Keywords: prosodic adaptation, spontaneous imitation, question intonation, Italian varieties

1. Introduction

There is a growing body of research literature providing evidence that people involved in a conversation tend to make their speech more similar to that of their conversational partner. Even though it has been claimed that such adaptation (also referred to as alignment, entrainment, accommodation, convergence) basically occurs via an automatic process of imitation (at different linguistic levels, Pickering & Garrod 2004, Garrod & Pickering 2009), a considerable number of studies have demonstrated that this process is not entirely automatic, but it is conditioned by social, interpersonal, and linguistic factors (for example Pardo 2006, Pardo et al. 2012, Babel 2012, Kim et al. 2011), the latter including evidence that imitation is constrained by the phonological grammar of speakers’ native language (see for example Nielsen 2011). This specific aspect in relation to prosodic adaptation (involving specifically intonation) has received little or no attention so far, and it is the one addressed in the present paper. In particular, we want to explore if and to what extent prosodic alignment between speakers of two Italian regional varieties is influenced by their native variety-specific intonation grammar rules. This is the case of the Italian varieties spoken in Bari and in Lecce, both located in a south-eastern region of Italy (Apulia). In these two Italian accents, unbiased yes-no questions are prototypically realised with two completely different tunes (Savino 2012): Bari Italian (henceforth BI) has a rising nuclear accent followed by a low boundary, phonologically described with the sequence \(L+H^*L-L\%\), whereas Lecce Italian (henceforth LI) is characterised by a terminal rise preceded by an accentual low or falling target, described with the sequence \((H+)\L^*L-H\%\) (see also Stella & Gili Fivela 2009). This implies that the functional rise for marking questioning is located differently in the two varieties: on the nuclear accent in BI, on the boundary in LI.

A yes-no question contour with a rising boundary can also be found in BI tonal inventory, but it does not bear any linguistic function. This rise-fall-rise contour is just an allophonic variant of the prototypical rising-falling pattern, being realised by adding an extra terminal rise after the functional accentual rise. This contour is described with the sequence: \(L+H^*L-H\%\). This intonation pattern can be occasionally encountered in spontaneous speech in relation to some paralinguistic functions like speaker attitude (Savino 2000; 2012), and it is systematically observed in read aloud yes-no questions, where the F0 excursion of the terminal rise is wide (Grice et al. 1997, Refice et al. 1997).

As to non-prototypical yes-no question contours with functional rise in non-terminal position in LI tonal inventory, the current stage of research offers only a very preliminary and fragmented picture showing a rising-falling or a rising-falling-rising contours as possible alternatives (Stella & Gili Fivela 2009, Savino 2012.), but detailed accounts also in relation to specific pragmatic meanings are not currently available. The aim of this paper is to explore if and to what extent prosodic alignment in BI and LI speakers when asking questions is conditioned by constraints imposed by the intonation grammar rule of the two Italian varieties as to the assignment of the phonological marker for questioning: a rise on the nuclear syllable in BI, a rise on the boundary in LI. If prosodic alignment would occur, different degrees of adaptation can be hypothesised:

1) a “strong” one, implying speakers to be able to copy the prototypical yes-no question F0 pattern of their conversational partner (which would entail overriding the native variety prosodic rule for marking questioning);

2) a “less strong” one, where speakers would resort to non-prototypical question contours in their native tonal inventory, whose features could at least partially resemble the native F0 pattern of their conversational partner.

2. Methodology

In order to elicit a considerable amount of genuine unbiased yes-no questions, five BI-LI pairs of speakers were asked to play the popular game called “Guess who?”. In this game, each participant receives a board with a set of pictures drawn on it, i.e. portraits of female/male characters along with their first name written at the bottom. The game starts with each player selecting a card from a separate pile of cards containing the same pictures drawn on the board, and it consists in being the
first to guess which card the partner has selected, by asking exclusively yes-no questions on characters’ features in order to eliminate candidates, until only one is left. In order to vary the lexical content of yes-no questions, in our experimental sessions participants were given five different boards, each provided with a different set of character types, like players of a number of sports, or objects and tools of everyday usage. When one of the two players had first gained 3 scores in a game round played with one of the boards, participants were requested to start a new game with another board. The winner was declared the one who had accumulated the highest score at the very end of the whole game session.

Participants sat at a desk and wore each an AKG C520 condenser microphone headset connected to a Marantz PMD 661 digital recorder. A cardboard screen situated between the two desks inhibited eye contact during the interaction, in order to both ensure that players could not see each others’ cards, and to maximise the effects of vocal alignment with respect to those related to visual behaviour (Rutter & Stephenson 1977). A silent experimenter and her assistant were always present during the session but without interfering during the whole game. Mean duration of the game sessions was 35.0 minutes (stdv=8.9).

2.1. Participants
Ten informants (five from Bari and five from Lecce dialectal areas) participated in the game sessions. They were all female students, aged 22-24, born and living in the target dialectal areas. Also, they were not familiar with each other (that is, they had never met, or were used to see each other before participating in the game sessions). These parameters were controlled since gender (e.g. Pardo 2006) and familiarity (e.g. Truong & Heylen 2012) seem to play a role in speech alignment. Students were given one exam credit as a reward for participating in the experiment.

2.2. Corpus
A total amount of 1010 yes-no questions were collected from the five recording sessions. They were all orthographically and intonationally annotated using Praat software tool (Boersma and Weenink, 2001). During a pre-processing step, a number of utterances were excluded from further analysis since they were identified as yes-no questions biased towards a positive answer (i.e. confirmation-seeking yes-no questions), intonationally characterised by a falling contour ((H+L*L-L%) in both varieties (see for example Grice & Savino 2003 for BI, and Stella & Gili Fivela 2009 for LI). These confirmatory yes-no questions typically occurred at the final stage of a game round, when one of the players asked the crucial question about the identity of the character/object to be guessed and she was confident as to the correctness of her guess. Other cases of utterances produced with accompanying disfluency, laughing, or other kind of signal noise generating possible ambiguities in intonation labelling were also discarded.

As results of this pre-processing selection, 983 utterances were finally included in the analysis, whose distribution across game sessions and BI/LI participants is shown in the rightmost column of Table 1. Intonation analysis was carried out within the autosegmental-metrical framework, on the basis of previous intonational accounts of the two Italian varieties under investigation.

3. Results
Table 1 shows the distribution (in percentage) of contour type occurrences across BI and LI speakers in each of the game interactions (sessions are identified by participants’ initial names (e.g. SR=game session run by participants S and R). Note that, as in previous works, also here the contours with either a low or a falling nuclear accent followed by a rising boundary are pooled together in one category, namely (H+L*L-H%), since they are phonologically accounted as the same tune. In all interactions, an overall trend can be noted: all BI speakers produced the smallest amount of yes-no questions with their prototypical L+H*L-L% contour (an example of which is shown in Figure 1, produced by BI speaker in FN session), whereas all LI participants were consistent in realising almost all their questions with their native prototypical (H+L*L-H%) pattern (an example of which is given in Figure 4, produced by LI speaker in FN game session), excluding LI speaker in the SN session, whose case will be discussed later in this section. This outcome gives an indication as to the direction of the prosodic adaptation (BI towards LI), since LI speakers appear to be much more resistant to diverge from their native prototypical interrogative F0 contour with respect to their BI interlocutors.

Table 1: Distribution of contour type occurrences across Bari and Lecce speakers in each game interaction. Sessions are identified by participants’ initial names (e.g. SR=game session run by participants S and R). The rightmost column shows the distribution (absolute values) of all yes-no questions analysed.

<table>
<thead>
<tr>
<th></th>
<th>L+H*</th>
<th>L-H%</th>
<th>(H+L)*</th>
<th>L+H</th>
<th>L-H%</th>
<th>L+H*</th>
<th>L-H%</th>
<th>L+H</th>
<th>L-H%</th>
<th>TOT quest.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>207</td>
</tr>
<tr>
<td>Bari</td>
<td>4%</td>
<td>63%</td>
<td>33%</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecce</td>
<td>2%</td>
<td>98%</td>
<td>4%</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD</td>
<td></td>
<td></td>
<td></td>
<td>219</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bari</td>
<td>17%</td>
<td>19%</td>
<td>64%</td>
<td>107</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecce</td>
<td>3%</td>
<td>93%</td>
<td>4%</td>
<td>112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN</td>
<td></td>
<td></td>
<td></td>
<td>231</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bari</td>
<td>2%</td>
<td>25%</td>
<td>73%</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecce</td>
<td>1%</td>
<td>99%</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AV</td>
<td></td>
<td></td>
<td></td>
<td>144</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bari</td>
<td>15%</td>
<td>68%</td>
<td>17%</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecce</td>
<td>5%</td>
<td>95%</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td></td>
<td></td>
<td></td>
<td>182</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bari</td>
<td>15%</td>
<td>85%</td>
<td>94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecce</td>
<td>56%</td>
<td>44%</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As to BI speakers, they realised the great majority of their yes-no questions by either adding an extra rising boundary (H%) after the BI typical L+H*L- nuclear pattern, and/or by reproducing the whole (H+L*L-H%) contour of their LI interlocutors. The latter case occurred predominantly for two BI participants (game sessions LD and FN), in 64% and 73% of their produced yes-no questions, respectively. The former was found as the most recurrent option in the remaining three BI speakers (game sessions SR, AV, and SN) as they added a rising boundary after the native L+H*L- pattern in 63%, 68% and 85% of their questions, respectively. However, even though to different extents, all BI speaker but one (the one in
session SN) made use of both strategies for accommodating intonationally to their LI game partner. Figures 2-3 show one of these cases, i.e. the F0 contour of yes-no questions produced by the same BI speaker during the FN game interaction, one with the L+H*L-H% pattern (Figure 2), and the other realised as a copy of the typical (H+)*L-L-H% LI contour (Figure 3). An example of the (H+)*L-L-H% pattern produced by the LI participant in the same FN interaction is offered in Figure 4. Interestingly, we also observed that the reproduction of the prototypical LI question tune by BI participants was often accompanied by the repetition also of the lexical content of the question. In the examples shown in the figures, the question ‘è un uomo?’ (is it man?) was first asked by the LI speaker (Figure 4), and then copied by the BI partner (Figure 3) in the immediately subsequent turn.

Moreover, we measured the pitch excursion size of the rising terminals realised by BI speakers in both L+H*L-H% and (H+)*L-L-H% contours (onset of the rise was manually marked in a second step labelling), and we found it was consistently similar in the two contour types. Mean values are shown in Table 2. A two-tail paired t-test indicates that pitch excursion mean values are significantly higher in (H+)*L-L-H% than in L+H*L-H% for BI speaker in FN interaction (\(t(106)=4.93, p<0.001\)), and not significantly different in the two contour types for the remaining three BI participants (SR, SD, AV sessions). For BI speaker in SN interaction, comparison was not possible because she did not produce any (H+)*L-L-H% pattern, yet pitch excursion of the terminal rise in her L+H*L-H% questions is comparable in size to that of the other BI speakers.

As to the LI participants, it can be only registered that the very negligible number of questions produced with a rising nuclear accents were always followed by a rising terminal. As a side observation, LI speaker in SN game session produced a considerable amount of yes-no questions (44%) characterised by a rise-fall movement with peak alignment much later than in BI rise-fall for questions, here provisionally described with the sequence L*+H L-L%. At the current stage of knowledge, it is not clear whether the use of such a contour...
is associated to some specific pragmatic meaning (speaker attitude?), or if it is just peculiar of the place (within the Lecce dialectal area) this speaker comes from.

Table 2: F0 excursion size (mean values, in semitones) of the L-H% rising boundary produced by Bari Italian speakers in questions with L+H*-L-H% and with ((H+)*L*L-H%) contours (stdv values in brackets).

<table>
<thead>
<tr>
<th></th>
<th>L+H*-L-H%</th>
<th>(H+)<em>L</em>L-H%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR Bari</td>
<td>3.3 (1.1)</td>
<td>3.8 (0.9)</td>
</tr>
<tr>
<td>LD Bari</td>
<td>3.2 (1.2)</td>
<td>3.5 (1.5)</td>
</tr>
<tr>
<td>FN Bari</td>
<td>4.4 (1.6)</td>
<td>6.2 (1.6)</td>
</tr>
<tr>
<td>AV Bari</td>
<td>3.4 (1.4)</td>
<td>3.8 (1.0)</td>
</tr>
<tr>
<td>SN Bari</td>
<td>3.0 (1.1)</td>
<td></td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

Results of this explorative study can provide some preliminary indications as to prosodic adaptation between speakers of two Italian varieties whose intonational grammars differ in the assignment of the functional rise for marking questioning: on the nuclear syllable in BI, on the boundary in LI.

Results show that BI speakers are those who adapt intonationally to their LI partners, and this is achieved by two different modalities/degrees:

1) by resorting to the allophonic variant of their native prototypical rising-falling contour, characterised by a rising instead of a falling boundary. In this way, BI speakers can to some extent “sound similar” to their LI partners’ who produce their questions with a (falling)rising contour, but they can still preserve the functional rise on the nuclear syllable. This adaptation modality is observed, to different extents, in all BI participants. Interestingly, this strategy has been previously reported in BI speakers when reading questions aloud, i.e. when they assume they have to adapt their speech to a more prestigious, standard-like variety (which has a (H+)*L*L-H% contour in questions, as in LI). This is associated with producing a wide terminal rise in questions (see Savino 2012 for a detailed discussion).

2) by copying the LI prototypical (H+)*L*L-H% contour, thus violating their native intonation grammar rule for marking questioning. This modality is predominant in two out of five BI speakers.

These outcomes seem to indicate that prosodic adaptation could be – but is not necessarily – conditioned by constraints imposed by the intonation grammar rules of the speakers’ native dialect. Because of this, the direction of the prosodic adaptation (BI towards LI) can be only partly explained as conditioned by the availability of the rise-fall-rise allophonic variant of the prototypical rise-fall in BI tonal inventory, since some BI speakers were also able to copy the typical (H+)*L*L-H% F0 pattern of LI questions.

An alternative possible explanation of the direction of prosodic adaptation as observed in our data might lie in the influence of social factors. Even though Bari and Lecce are located in the same region, speakers of the two varieties are perceived as belonging to socio-linguistically and culturally very different groups, and such a cultural identity is perceived as much stronger in Lecce than in Bari communities. This might account for the strong resistance by LI participants in our game interactions to deviate from their native prototypical F0 contour. However, this parameter was not controlled in our exploratory study, and this issue is also worth being further investigated in future research.

5. Acknowledgements

I would like to thank my research assistants Grazia Materia and Federica Scarlino for invaluable help in preparing the materials for the experiment and collecting the data. Thanks to Grazia Materia also for assistance during the early stage of data pre-processing. I am also indebted to Mario Refice for support in Praat scripting, and to two anonymous reviewers for insightful comments and suggestions. All errors are of course mine.

6. References


Sensory-motor interactions in speech perception, production and imitation: behavioral evidence from close shadowing, perceptuo-motor phonemic organization and imitative changes.

Lucie Scarbe1, Denis Beautemps1, Jean-Luc Schwartz1, Marc Sato1

1 Gipsa-lab,Département Parole & Cognition, CNRS & Grenoble Université, Grenoble, France
lucie.scarbe@gipsa-lab.grenoble-inp.fr, denis.beautemps@gipsa-lab.grenoble-inp.fr, jean-luc.schwartz@gipsa-lab.grenoble-inp.fr, marc.sato@gipsa-lab.grenoble-inp.fr

Abstract
Speech communication can be viewed as an interactive process involving a functional coupling between sensory and motor systems. In the present study, we combined three classical experimental paradigms to further test perceptuo-motor interactions in both speech perception and production. In a first close shadowing experiment, auditory and audio-visual syllable identification led to faster oral than manual responses. In a second experiment, participants were asked to produce and to listen to French vowels, varying from height feature, in order to test perceptuo-motor phonemic organization and idiosyncrasies. In a third experiment, online imitative changes on the fundamental frequency in relation to acoustic vowel targets were observed in a non-interactive situation of communication both unintentional and voluntary imitative production tasks. Altogether our results appear exquisitely in line with a functional coupling between action and perception speech systems and provide further evidence for a sensory-motor nature of speech representations.

Keywords: speech perception, speech production, sensory motor interaction

1. Introduction
An old and classical debate in the speech communication domain concerns the possible motor implication in speech perception and, more generally, the auditory versus motor nature of the speech code. Auditory theories assume that speech perceptual processing and categorization are based on acoustic cues and auditory representations (Stevens and Blumstein 1978, 1979; Lindblom et al. 1988, 1990). Conversely, the motor theory of speech perception (Liberman et al., 1985) and its direct realist variant (Fowler et al., 1986) claim that there is a crucial role of the motor system in speech perception. More recently, a number of perceptuo-motor theories attempted various kinds of syntheses of arguments by tenants of both auditory and motor theories, proposing that implicit motor knowledge and motor representations are used in relationship with auditory representations and processes to elaborate phonetic decisions (Skipper et al., 2007; Schwartz et al., 2012).

Various experimental settings enable to test and study the relationship between speech perception and action. Let us describe three of them which provide the basis for the present work. First, close-shadowing provides a natural paradigm for testing perceptuo-motor links. Indeed, Porter et al. (1984) and later Fowler et al. (2003) observed very fast reaction times when participants had to shadow a syllable as quickly as possible. Compared to manual responses, oral speech responses were also found quicker than manual ones (Galantucci et al., 2006). This difference was interpreted by the theoretical assumption that perceiving speech is perceiving gestures, and that gesture perception directly controls speech response and makes it faster.

Another way to prove the evidence of a perceptuo-motor linkage is to directly test the existence of a common perceptual and motor phonemic organization. From that view, Bell-Berti (1979) showed that differences between subjects in the perception of the [i] versus [I] contrast in American English seemed to be linked to differences in the articulatory implementation of this contrast. Menard et al. (in press) further showed similar idiosyncrasies in both vowel production and perception, a result suggesting a link between perceptual and motor phonemic prototypes in the human brain.

Finally, the ability to converge and to imitate a listener also attests of a perceptuo-motor coupling. Recently, online unintentional and voluntary imitative changes in relevant acoustic features of vowel targets were observed during speech production in a non-interactive situation of communication (e.g., Garnier et al., 2013; Sato et al., 2013). These results were explained by the possibility that speech production continuously draws on perceptuo-motor learning from the external speech environment and prior listener’s sensory-motor knowledge.

In the present study, we further tested sensory-motor interaction in these three paradigms. In a close shadowing experiment (Experiment A), we compared reaction times to auditory and audio-visual speech stimuli from manual and oral responses. We expected to find faster reaction times to oral compared to manual responses, and to audiovisual compared to auditory stimuli. The second experiment (Experiment B) tested perceptuo-motor phonemic organization in vowel production and perception. Our aim was to possibly determine a common phonemic organization in vowel perception and production as well as to test subtle perceptuo-motor idiosyncrasies between participants. Finally, the third experiment (Experiment C) concerned phonetic convergence and voluntary imitative changes in relation to acoustic vowel targets.

2. Methods
2.1. Participants
Three groups of respectively fifteen, twenty-seven and sixteen healthy adults, native French adults, participated in Experiments A, B and C. All participants had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders.
2.2. Stimuli

2.2.1. Experiment A

Multiple utterances of /apa/, /ata/ and /aka/ sequences were individually produced by a male native French speaker (who did not participate in the experiments) in a sound-attenuated room. The corpus was audio-visually recorded with the objective to obtain 4 different occurrences of /apa/, /ata/ and /aka/ with various durations of the initial /a/ vowel (i.e., 0.5s, 1s, 1.5s and 2s) so as to obtain 12 distinct stimuli.

2.2.2. Experiment B

Thirteen acoustic stimuli were used for the vowel perception task of Experiment B. Those stimuli were synthesized from VLAM (Variable Linear Articulatory Model), an articulatory-to-acoustic model of the vocal tract based on Maeda’s adult model (Boe and Maeda, 1997; Boe, 1999). Using VLAM, we generated thirteen stimuli distributed regularly within the maximal adult vowel space from high to low front unrounded vowels.

2.2.3. Experiment C

A vowel database was created from /e/, /œ/ and /o/ French vowels produced by two male and female speakers. From these stimuli, f0 was artificially shifted by steps of ±5Hz (from 80Hz to 180Hz for the male vowels, and from 150 to 350Hz for the female vowels) using the PSOLA module integrated in Praat software (Boersma and Weenink, 2013).

2.3. Experimental procedure

The three experiments were carried out in a sound-proof room. Participants sat in front of a computer monitor at a distance of approximately 50 cm. The acoustic stimuli were presented at a comfortable sound level through a loudspeaker, with the same sound level set for all participants. The Presentation software (Neurobehavioral Systems, Albany, CA) was used to control the stimulus presentation during all experiments, and to record key responses in Experiment A and B (see below). All participants’ productions were recorded for off-line analyses.

2.3.1. Experiment A

The experiment consisted of two categorization tasks: close-shadowing in one case, where the responses were provided orally by repeating as quickly as possible the presented speech sequence; manual decision in the other case, where the responses were provided manually, by pressing as quickly as possible the appropriate key. The stimuli to categorize consisted in /apa/, /ata/ and /aka/ sequences. For each task (oral vs. manual response) and each modality (auditory vs. audiovisual), 16 occurrences of /apa/, /ata/ and /aka/sequences were presented in a fully randomized sequence of 48 trials. The order of task and modality of presentation was fully counterbalanced across participants.

2.3.2. Experiment B

This experiment consisted of two perception and production tasks, counterbalanced across participants. For the production task, participants were asked to produce fifteen repetitions of the 10 oral French vowels /ɛ, ɛ, œ, e, o, e, o, e, a, œ/ according to a visual orthographic target. Target vowels were presented in a fully randomized order. For the perception task, participants had to manually categorize acoustic stimuli among the four front unrounded French vowels /i, e, e, a/. Each stimulus was presented ten times in a fully randomized order.

2.3.3. Experiment C

Experiment C consisted in three vowel production tasks. First participants had to individually produce /e/, /œ/ and /o/ vowels, according to a visual orthographic target. This allows the experimenter to measure participant’s f0. In the subsequent task, participants were asked to produce the three vowels according to an acoustic target. Importantly, no instruction to “repeat” or to “imitate” the acoustic targets was given to the participants. Finally, the third task was the same as the second task except that participants were explicitly asked to imitate the acoustic targets. The only indication given to participants was to imitate the voice characteristics of the perceived speaker. Acoustic target for each participant were 27 stimuli selected from the vowel database, with the 9 quantified f0 frequencies varying from -20% to +20% by steps of 5% around his/her own pitch, as measured in the first task.

2.4. Data analysis

All acoustic analyses of participants’ productions were performed using Praat software (Boersma and Weenink, 2013).

2.4.1. Experiment A

The proportion of correct responses was determined for each participant and each condition, together with reactions times (RTs) for correct responses. RT in the oral task was estimated from the burst onset of the stop consonant to categorize to the burst onset of the oral response.

2.4.2. Experiment B

For the production task, the mean F1 frequency for /i, e, e, a/ was computed for each participant. In all this study, frequencies are estimated in bark, thanks to the formula proposed by Schroeder et al. (1979) for the perception task, the mean F1 frequency of all stimuli categorized respectively as /i, e, e, o, e, o, o, a, a/ was determined for each participant. For both the perception and production tasks, mean normalized bark values for /e/ and /œ/ with regard to their distance from /a/ and /i/ was then calculated. Correlation scores between production and perception was finally determined for all participants.

2.4.3. Experiment C

In all tasks of Experiment C, we measured f0 for each produced vowel. In the second and third tasks, correlation analyses between f0 values in the perceived and produced vowels were performed for each participant.

3. Results

3.1. Experiment A - see Figure 1

3.1.1. Reaction times

RTs were entered into an ANOVA with three factors: modality (auditory, audiovisual), response (speech, key) and syllable (/pal, /f, /ka/). Although no significant difference between the auditory and audiovisual stimuli was observed, RTs were shorter for speech responses (240 ms) than for key responses (462 ms) ($F(1,14)=81.8; p<0.001$). Interestingly, an interaction between the three factors was found ($F(2,28)=4.6; p=0.01$). While, for speech responses, RTs for /pal/ did not differ between the auditory (196ms) and audiovisual (208ms) modalities, for key responses an audiovisual advantage was
observed (audiovisual stimuli: 415 ms, audio stimuli: 442 ms). For /tæ/, manual RTs were longer for audiovisual (506 ms) than for auditory (465 ms). For /kæ/, no differences were found between the modalities and tasks.

3.1.2. Perceptual recognition

As for RTs, the percentage of correct responses were entered into an ANOVA with three factors: modality (auditory, audiovisual), response (speech, key) and syllable (/pæ/, /tæ/, /kæ/). No difference was observed between auditory (95%) and audiovisual (94%) stimuli. However, participants made significantly fewer errors for key (97%) than for speech responses (93%) (F(1,14)=13; p<0.002), and fewer errors for /pæ/ (98%) than for /tæ/ and /kæ/ syllables (93%) (F(2,28)=6.8; p<0.004). In addition, a significant interaction between the modalities and syllables was also observed (F(2,28)=5.6; p<0.01). For /tæ/ and /kæ/, more correct responses were observed for key (97% and 97%) than for speech (90% and 89%) responses. For /pæ/, no difference was observed.

3.2. Experiment B - see Figure 2

In the production task, the mean F1 values for /ɪ/, /ɛ/, /ɛ/ and /æ/ in barks were respectively 3.1 (range: 2.6-3.6), 4.4 (range: 3.4-4.5), 5.9 (range: 4.4-7.1) and 7.3 (range: 6.2-8.5). Idiosyncrasies were weak for /ɛ/ (normalized distance from /ɪ/ between 19 and 46 bark but with a small standard deviation at .35 and .88 bark with a standard deviation at .15). In the perception task, the mean F1 values for /ɪ/, /ɛ/, /ɛ/ and /æ/ in barks were respectively 2.8 (range: 2.6-3.9), 4.2 (range: 3.8-4.5), 5.5 (5.3-5.7) and 6.8 (range: 6.6-7.0). Variability in perception was extremely small, showing that no idiosyncrasies were found between participants. From these results, a quasi perfect correlation of acoustic values between produced and perceived vowels is observed (with a mean slope for all participants of .93, range: 0.7-1.3).

3.3. Experiment C - see Figure 3

In Experiment C, imitative changes were observed in both tasks, though stronger in voluntary imitation. Slope coefficients differed significantly from zero in both the production (t(15)=6.2; p<0.001) and imitation (t(15)=19.2; p<0.001) tasks. In addition, slope coefficients were higher in the imitation (0.83) compared to the production (0.44) tasks (t(15)=5.6; p<0.001). Similarly, correlation coefficients differed significantly from zero in both the production (r=0.93) compared to the production (r=0.63) tasks (t(15)=4.2; p<0.001).

4. Discussion and Conclusion

4.1. Experiment A

Overall, as in the studies by Fowler et al. (2004) and Porter et al. (1984), orofacial responses were much quicker than manual ones. While no differences were found between auditory and audiovisual modalities in the close shadowing task, quicker response times were however observed in the audiovisual modality for the manual categorization task for bilabial consonants, likely due to the visible anticipatory gesture. The fact that this visual gain was not seen in the orofacial modality is probably due to a floor effect considering the small response time in close shadowing. Although these results do not provide global evidence for faster response times in the audio-visual modality in the close shadowing task, they appear compatible with a sensory-motor framework in which there is a functional connection between action and perception systems.
4.2. Experiment B
Our results for the production task appear partly coherent with those found by Ménard and Schwartz (in press). One important difference, however, is that though our study displays idiosyncrasies in production more or less in line with their study, we did not find almost any idiosyncrasy in the perception task. This difference is likely due to the different experimental factors used in these two studies. While we only tested adults, Ménard and Schwartz tested two groups of 4 and 5 years old children and one group of adults. Moreover, the stimuli used in the perception task for the adults were not the same as ours (with a larger number and type of stimuli, and a more variable distribution in the acoustic space). Given the larger variability of the stimuli used by Ménard and Schwartz (in press), idiosyncrasies are more likely to emerge. Importantly, in line with the maximal dispersion theory of Lindblom (1972) and with a perceptuo-motor coupling of vowel perception and production (Schwartz et al., 2012), we found a near to perfect acoustic equidistance between the centers of vocalic targets both in the production and perception tasks (see Figure 3).

4.3. Experiment C
As in Garnier et al. (2013) and Sato et al. (2013), we found a quasi perfect imitation of vowel targets on F0 in the voluntary imitation task, as well as clear evidence for phonetic convergence in the production task. This latter result suggests that participants tend to converge towards an acoustic speech target even if they don’t imitate consciously. Altogether, these results are perfectly compatible with a perceptuo-motor linkage in speech production and perception.

4.4. General discussion
Taken together, the three experiments largely confirmed previous results and strongly suggest a functional perceptuo-motor coupling of speech perception and production systems. They provide further evidence for a sensory-motor nature of speech representations. This series of coupled experimental paradigms for studying the relationship between perceptual and motor processes will now serve as a platform for assessing the recovery of this relationships in hearing impaired subjects after cochlear implantation.

5. Acknowledgements
This work was supported by the French National Research Agency (ANR) through funding for the PlasMody project (Plasticity and Multimodality in Oral Communication for the Deaf).

6. References
Measuring Reaction Times: Vocalisation vs. Articulation
Sonja Schaeffler, James M. Scobbie, Felix Schaeffler

CASL, Queen Margaret University, Edinburgh, UK
sschaeffler@qmu.ac.uk, jscobbie@qmu.ac.uk, fschaeffler@qmu.ac.uk

Abstract
There is a sizeable delay between any formulation of an intention to speak and the audible vocalisation that results. Silent articulatory movements in preparation for audible speech comprise a proportion of this phase of speech production. The extensive literature on Reaction Time (RT) is based on the delay between a stimulus and the acoustic onset to speech that is elicited, ignoring the preceding silent elements of speech production in what is an utterance-initial position. We used a standard Snodgrass and Vanderwart picture-naming task to elicit speech in a standard Reaction Time protocol, but recorded the behaviour of two typical speakers with audio plus Ultrasound Tongue Imaging (201 frames per second) and de-interlaced NTSC video of the mouth and lips (60fps). On average, Acoustic Reaction Time occurred between 120 to 180 ms later than a clearly observable articulatory movement, with no consistent advantage for lip or tongue-based measures.

Keywords: Ultrasound Tongue Imaging, video, tongue, lips, Reaction Time, speech preparation

1. Introduction
Like all organisms, humans have to react to a wide range of external stimuli in an appropriate way, within an appropriate time. A huge range of scientific studies related to such abilities have been undertaken, many specifically looking at Reaction Time, which is operationalised as the time it takes from presentation of a stimulus to an observable response. A very common observable response in psycholinguistic experiments is verbal feedback, and here, Reaction Time is usually determined via “voice key”, a device which is triggered automatically as soon as the sound pressure reaches a pre-defined level. However, what is detectable as the onset of acoustic output is only one stage in the speech production process. Before anything becomes audible, the articulators (i.e. the tongue, the lips, the jaw, etc.) have already moved into place. This movement shows that the motor plan for the response has been put into action and is an earlier observable response than audible speech.

Models of speech production such as WEAVER++ (Roelofs 1997; Levelt et al 1999) or GODIVA (Bohland et al. 2012) are readily applied to single word productions, though as Mooshammer et al. (2012) point out, the models lack a great deal of detail and “apart from the effects of learning/practice very little is understood about the assembly of motor (gestural) plans for an utterance” (p. 375). We can assume in a single word Reaction Time task, the necessary sequential stages include extraction of phonologically encoded items from the lexicon, followed by the construction of a speech motor-plan for execution. The measurable acoustic response follows in turn, but silent movements of the articulators themselves are also clearly detectable by human observers, preceding audible output, and these can be quantified, given the appropriate instrumentation. (We should note that for longer phrases in connected speech, these processes overlap or occur in parallel, which is a more complex problem that need not concern us here.) Some instrumentation such as EMA has been used previously (Mooshammer et al. 2012) to look at this silent phase in production, though not directly to measure Reaction Time, but EMA is relatively expensive to use and is not generally available. Other instruments such as high-speed video or Ultrasound Tongue Imaging are much more readily available and with appropriate automated analysis methods (cf. Palo et al. 2014) could be a competitor to audio recordings plus voice key.

This paper is a pilot study aiming to investigate how much of a discrepancy there is between acoustic and articulatory onset of speech across a variety of tokens with different phonemic targets in a standard Reaction Time task. We obtained high-quality acoustic recordings to determine when verbal response becomes first audible, and used ultrasound recordings of the tongue, and video recordings of the lips, to observe when verbal response is first initiated by the articulators.

2. Method

2.1. Participants
We report data from two female native speakers of Scottish varieties of English (aged between 20 and 35). The participants reported no visual or hearing impairments.

2.2. Material
Verbal responses were collected in a standard picture naming task, with target items drawn from the well-tested and frequently used Snodgrass-and-Vanderwart picture inventory (1980). The picture inventory consists of 260 targets.

2.3. Tongue and lip recordings: instrumentation
All articulatory and acoustic recordings were obtained simultaneously and synchronised using AAA software from Articulate Instruments Ltd (2012). The participants were fitted with a purpose-built headset to ensure stabilisation of the ultrasound probe (Articulate Instruments Ltd 2008; Scobbie et al. 2008). Attached to the helmet was a small Audio Technica AT803b microphone for high-quality acoustic recordings, plus a NTSC micro-camera to capture recordings of the speakers’ lips.

Ultrasound recordings were obtained at a rate of 201 frames per second from a SonixRP system. Video was captured then deinterlaced to an effective rate of 59.95 fps. Recordings started 1.5 seconds before prompt presentation so that the whole speech production process was captured.

2.4. Synchronisation
Tests of synchronisation of audio to ultrasound, and of audio to video were made that suggest a <5ms error. Our audio-
ultrasound synchronisation uses a square-wave sync signal and each scan is stored as a unique frame with no overlap. Synchronisation at 201 fps is therefore within a 5 ms window. The video capture of the mouth and lips was synchronised using an Articulate Instruments ‘bright-up’ electronic clapperboard that records a sync pulse on camera and microphone outputs. Video frame rate was set at 19.97 fps. For both image types, we assume no extra delay in data transfer, or due to image creation (in contrast to video ultrasound, cf. Wrench et al. 2006).

3. Token selection, annotation and analysis

3.1. Tokens for analysis

Of the 260 targets presented, 19 elicited erroneous responses in Speaker 1 and were excluded, leaving 241 tokens for analysis. Speaker 2 could not tolerate the headset for the entire length of the experiment and stopped after recording of 156 tokens. Of the recorded tokens, 4 elicited erroneous responses, leaving 152 tokens for analysis for Speaker 2. Targets that elicited British English variants instead of the American English target (e.g. ‘waistcoat’ instead of ‘vest’) were not excluded.

3.2. Annotations

3.2.1. Labels

For each of the tokens (241 for Speaker 1; 152 for Speaker 2) we annotated a) the onset of target-related lip movement b) the onset of target-related tongue movement and c) the acoustic onset.

3.2.2. Articulatory annotation criteria

To be target-related, the relevant articulator had to be seen to move in a smooth and consistent manner towards one of the initial articulatory targets. For example a word like “barn” requires initial labial closure, and this might be detectable from an open mouth position or from a closed mouth that opens, then closes again for the /b/. We annotated only the unambiguous lip movement towards closure in each case. In addition, the tongue has to move to the appropriate position for the /a/. This might be from a neutral schwa-like position or from palatal contact. Again, only a single, contiguous unambiguous holistic tongue movement towards an articulatory target was annotated.

3.2.3. Acoustic annotation criteria

The acoustic onset was determined by visual inspection of the speech signal for a rapid and unambiguous increase in acoustic intensity and/or spectral cues to the onset of vocalisation. Figure 1 illustrates the different sources of data available for annotation in AAA.

3.3. Annotation confidence ratings

Given our hand-labelled holistic approach, we gave each articulatory annotation a confidence rating from 1 = very unsure to 5 = absolutely sure. Tokens that received a low rating included e.g. instances where it was difficult to determine the boundary between onset of target-related articulation and some directly preceding, non-linguistic pre-speech behaviour (e.g. tongue clicks, pursing of lips). For the purpose of exploring the suitability of our newly developed method for obtaining Articulatory Reaction Time data, we only used tokens for further analysis with annotations that for both lip and tongue onset received either a ‘4’ or ‘5’ confidence rating. This left us with 132 tokens (55%) for Speaker 1 and 82 tokens for Speaker 2 (54%).

3.4. Phonetic properties of targets

All tokens were coded in analysis for the phonetic properties of the actually uttered target word, i.e. ‘Onset Voicing’ (voiced / voiceless), ‘Onset Manner’ (vowel or glide / stop or affricate), ‘Onset Place’ (labial / glide / lingual) and ‘Number of Syllables’ (monosyllabic / polysyllabic).

It is important to note that the tokens that had received lower confidence ratings of ‘1’, ‘2’ or ‘3’ did not fall into any specific phonetic categories. An analysis of Speaker 1’s excluded 109 tokens comprised of 56 stops/affricates, 36 fricatives and 17 vowels/glides with all places of articulation represented, compared to 67 stops/affricates, 35 fricatives and 30 vowels/glides in the confidently rated 132 tokens. This suggests that there was no place or manner of articulation that was inherently more difficult to determine articulatory onsets for.
4. Results

4.1. Acoustic Reaction Times

Across all remaining tokens the untrimmed Acoustic Reaction Time was on average 897 ms (SD 289 ms) for Speaker 1 and 624 ms (SD 205 ms) for Speaker 2. To exclude outliers that were caused by word finding difficulties or measurement artefacts we trimmed the sample by excluding all tokens with Acoustic Reaction Times that were higher or lower than the mean ±2SD (Speaker 1: n = 7; Speaker 2: n = 5). After removal of outliers the mean Acoustic Reaction Time was 851 ms (SD 251 ms) for Speaker 1 and 586 ms (SD 127 ms) for Speaker 2. That means overall Speaker 2 exhibited a much faster Reaction Time than Speaker 1.

4.2. Articulatory Reaction Times

Lip Reaction Time was on average 677 ms (SD 237 ms) for Speaker 1 and 466 ms (SD 127 ms) for Speaker 2. Tongue Reaction Time was on average 670 ms (SD 237 ms) for Speaker 1 and 442 ms (SD 137 ms) for Speaker 2. Observed Articulatory Reaction Time thus occurred approximately 175-180 ms ahead of Acoustic Reaction Time for Speaker 1. For Speaker 2 observed Articulatory Reaction Time occurred approximately 120-145 ms ahead of Acoustic Reaction Time. In other words, silent articulation comprised 20% - 25% of the Reaction Time as computed from the acoustic onset to speech (cf. Figure 2).

Within tokens, the mean difference between Lip and Tongue Reaction Time was 7 ms (SD 80 ms) for Speaker 1, which means there was no significant difference between the detection of Reaction Time using lip vs. tongue motion. For Speaker 2, the mean difference between Lip and Tongue Reaction Time within tokens was 24 ms, so Speaker 2’s Lip Reaction Time was significantly slower than her Tongue Reaction Time (paired samples t-test, t(123)=3.672, p<.01).

4.3. Phonemic target types

Since each target lexeme was coded for the primary place, manner, and voicing of its initial consonant or vowel, we were able to test whether these phonemic targets gave rise to differently timed initiations of lip or tongue movement. There was no pattern that would suggest that one articulatory measure, lip or tongue, provided a more advantageous route to uncovering the Reaction Time for a certain phonemic target type. ‘Onset Place’, ‘Onset Manner’ and ‘Number of Syllables’ showed no significant effects on Reaction Time measurements.

Only ‘Onset Voicing’ showed a trend, with the trend going in opposite directions for the two speakers. For Speaker 1 Onset Voicing only had an effect on Acoustic Reaction Time (independent samples t-test, t(123)=1.756, p=.082), with Acoustic Reaction Times being on average 67 ms shorter for voiceless onsets than for voiced onsets. For Speaker 2 Onset Voicing showed a trend only for Tongue Reaction Time (t(71.8)=1.769, p=.081), with Tongue Reaction Times being on average 50 ms longer for voiceless onsets than for voiced onsets (cf. Figure 3).

This overall lack of phonemic bias is particularly interesting in light of findings that Reaction Times derived via the commonly used voice key are in fact quite susceptible to differences in onset type. Kessler et al. (2002) report striking phonetic biases, caused by phonemes’ intrinsic differences in sound pressure level. Surveying data from four large-scale studies, they demonstrate that voiceless, posterior, and obstruct consonants set off voice keys later than others.

5. Discussion and conclusion

Overall, the findings demonstrate that articulatory measurements capture Verbal Reaction Time reliably at a much earlier time point than acoustic measurements, and while this is a small data set with only two speakers it raises interesting points related to inter-individual speaker differences. These are typical speakers, but also very different speakers.

Speaker 1 had not just fuller lips than Speaker 2, but also a tendency to move lips more, and maybe with a wider range. This might explain why Speaker 2 exhibited what seemed a slower Lip Reaction Time compared to her Tongue Reaction Time: We might not have been able to detect very early onsets of lip movement as easily as in Speaker 1, and lip movement might also just not be an as prominent feature in articulation in Speaker 2 as it is in Speaker 1. To estimate the impact of inter-individual (habitual or anatomical) differences on articulatory measures a much larger sample of speakers is required.

Another interesting point is the substantial difference in overall Reaction Time: Speaker 2 was overall much faster in
her verbal response than Speaker 1. Speaker 2’s advantage of Articulatory over Acoustic Reaction Time was smaller but actually in proportion with her overall ‘compressed’ run-up to audible speech: For both speakers the duration of the Articulatory Reaction Time amounted to only around 75 - 80% of the Acoustic Reaction Time.

This suggests that there is a robust advantage of articulatory measurements of Reaction Time to measurements based on acoustic data – no matter how quickly speakers are able to get their articulators into motion. A systematic quantification and qualification of the articulatory data advantage could potentially inform a re-evaluation of well established Reaction Time protocols based on analysis of facial video or Ultrasound Tongue Imaging.

6. References


Making the magnitude of auditory feedback errors predictable influences both vocal and event-related responses.

Nichole E. Scheerer & Jeffery A. Jones

Psychology Department and Laurier Centre for Cognitive Neuroscience, Wilfrid Laurier University, Waterloo, Ontario, N2L 3C5, Canada

nikkischeerer@gmail.com, jjones@wlu.ca

Abstract

 Fluent speech production relies on the combined effort of a feedback control system driven by sensory feedback, and a feedforward control system driven by internal models. However, the factors that dictate the relative weighting of these feedback and feedforward control systems are unclear. In this event-related potential (ERP) study, participants produced vocalizations while exposed to blocks of frequency altered feedback (FAF) perturbations that were either predictable in magnitude (consistently either 50 or 100 cents), or unpredictable in magnitude (50 and 100-cent perturbations varying randomly within each vocalization). Investigation of vocal and P1-N1-P2 ERP responses revealed decreases in the magnitude of vocal responses, smaller N1-amplitudes, and shorter vocal, P1 and N1 response latencies following predictable FAF perturbation magnitudes. These results suggest that after repeated exposure to predictable FAF perturbations, auditory feedback is deemed unreliable, and weighting of the feedforward system increases.

Keywords: frequency altered feedback, predictability, feedforward control, feedback control, ERP, P1, N1, P2.

1. Introduction

Computational models of speech motor control (e.g., the directions into velocities of the articulators model; Guenther, 2006) and experimental evidence (e.g., Jones and Keough, 2008) suggest that fluent speech production relies on the combination of a feedback control system, driven by sensory feedback, and a feedforward control system, driven by internal models. The feedforward control system is necessary for maintaining fluent speech, as strict reliance on the feedback control system would induce disfluencies during speech production (Civier et al., 2010; Guenther, 2006). On the other hand, the feedback control system is crucial for detecting and correcting speech production errors, and for providing feedback to update the mapping of the internal models (Civier et al., 2010). However, it is currently unclear what factors dictate the relative weighting of feedback and feedforward input for speech motor control.

Previous studies have demonstrated that vocal and ERP responses to FAF are modulated by the prediction created by the internal model (Heinks-Maldonado et al., 2005; Houde et al., 2002), as well as the predictability of experimentally induced manipulations of the direction (Korzyukov et al., 2012) and latency (Burnett et al., 2008; Chen et al., 2012) of feedback perturbations. Together, these studies suggest that being able to predict the magnitude of experimentally induced feedback errors influences both behavioural and neural responses to these errors. For this reason, in this experiment participants produced vocalizations while exposed to blocks of FAF perturbations that were either predictable in magnitude (consistently either 50 or 100-cents), or unpredictable in magnitude (50 and 100-cent perturbations varying randomly within each vocalization). Since both behavioural and neural responses to FAF have shown signs of habituation following the presentation of repetitive predictable stimuli, we expected that when the magnitude of the feedback perturbations were predictable, vocal and ERP responses would be smaller as the predictable nature of the stimuli would result in increased weighting of the feedforward control system. On the other hand, when feedback perturbation magnitudes varied randomly, the unpredictable nature of the sensory feedback would drive an increase in the weighting of the feedback control system, allowing for rapid correction of auditory feedback errors. Accordingly, we expected that predictable feedback perturbation magnitudes would elicit smaller vocal and ERP responses relative to the randomly varying feedback perturbation magnitudes, reflecting habituation as a result of the repetitive and predictable nature of the stimuli. If as expected vocal and neural responses differ as a function of stimulus predictability, these results will suggest that the weighting of the feedback and feedforward speech motor control systems is influenced by the predictability of the magnitude of auditory feedback errors.

2. Methods

2.1. Participants

Thirty-six participants between the ages of 18 and 25 years (mean=19.33, SD=1.85; 27 female) participated in this study. Participants were right-handed native Canadian English speakers and did not speak a tonal language. Informed consent was obtained from all participants according to the ethical policies at Wilfrid Laurier University. All participants received course credit or financial compensation for their participation in this study.

2.2. Procedure

Participants vocalized the vowel sound /a/, 120 times over four blocks, while exposed to unaltered and FAF. Each experimental block contained 30 trials where the participants’ auditory feedback was perturbed either 50 or 100 cents downwards, four times per vocalization, or left unaltered. Each perturbation had a fixed duration of 200 ms and occurred with an inter-stimulus interval of 700-900 ms, resulting in vocalizations that were approximately five seconds in length. The four experimental blocks were divided into two conditions: the non-random condition, where perturbation magnitudes remained consistent at 50 or 100-cents within each vocalization, and the random condition, where perturbation magnitudes were different combinations of 50 and 100-cents
within each vocalization (see Figure 1). The 50-cent non-random block contained 25 trials where the participant’s voice was perturbed 50-cents downwards four times per vocalization, and five trials where the participant’s voice was left unaltered, but sampled four times per vocalization. This resulted in 100 50-cent non-random trials, and 20 unaltered non-random trials. Similarly, the 100-cent non-random block contained 25 trials where the participant’s voice was perturbed 100-cents downwards four times per vocalization, and five trials where the participant’s voice was left unaltered, but sampled four times per vocalization. This resulted in 100, 100-cent non-random trials, and 20 unaltered non-random trials. The two random blocks both contained 24 trials where the participant’s voice was perturbed 50 and 100-cents downwards, in different combinations, for a total of four perturbations per vocalization, while the other six vocalizations were left unaltered. A latin-square was used to determine all possible combinations of four 50 or 100-cent perturbations for the random condition. This resulted in 16 different combinations of 50 and 100. Each combination of 50 and 100 occurred three times per experimental session, and was pseudo-randomly presented throughout the two random experimental blocks. This resulted in 96 50-cent random trials, 96 100-cent random trials, and 48 unaltered random trials, split between the two blocks of random trials. FAF and unaltered trials were pseudo-randomly presented within each block. The block order for all participants was: random, non-random, random, non-random. However, the presentation order of the non-random blocks was counterbalanced across participants.

2.4. Behavioural Recording and Analysis

The unaltered voice signal was segmented into separate vocalizations and F0 values were calculated for each vocalization using the SWIPE’ algorithm (Camacho & Harris, 2008). Each vocalization was then segmented based on the onset of the four perturbations. F0 values for each of the four perturbed segments were normalized to the baseline period, which was the portion of the segment 200 ms prior to the onset of the perturbation, by converting Hertz values to cents using the following formula:

\[ \text{cents} = 100(12 \log_2 \frac{F}{B}) \]  

(1)

In the formula, \( F \) is the F0 value in Hertz and \( B \) is the mean frequency of the baseline period.

Cents values were calculated for the 200 ms prior to the perturbation (the baseline period), and 1000 ms after the perturbation. An averaged F0 trace was constructed for each perturbation magnitude, 0 (unaltered), 50, and 100, in each condition (random and non-random) for each participant. Vocal responses were quantified by examining the response magnitude and latency. The magnitude of compensation was determined by finding the point at which the participant’s averaged F0 trace deviated maximally from the baseline mean, and the latency was calculated as the time at which this maximal deviation occurred.

2.5. ERP Recording and Analysis

EEG signals were recorded from 64 scalp electrodes and referenced online to the vertex (Cz) electrode. Data were band-pass filtered (1-30Hz) and digitized (12-bit precision) at 1000 samples per second. Electrode impedances were maintained below 50 kΩms throughout the duration of the experiment. After data acquisition, EEG voltage values were re-referenced to the average voltage across all electrode sites. The data were then epoched into segments from 100 ms before the onset of the perturbation to 500 ms after perturbation onset. Data were analyzed offline for movement artifacts and any segment with voltage values exceeding 55 µV of the moving average over an 80 ms span were rejected. In addition, a visual inspection of all data was completed to ensure artifacts were being adequately detected. Eight subjects were eliminated from further analyses, as they had less than 50% of their trials retained across all perturbation magnitudes. Across all other participants, on average, between 84-92% of trials were retained for each perturbation magnitude.

Eight central electrodes were included in the analysis: CP1, C1, FC1, FCz, Cz, FC2, C2, and CP2. These electrodes were then grouped into left (CP1, C1, FC1), medial (FCz and Cz), and right (FC2, C2, and CP2). These electrodes were chosen based on visual inspection of the regions demonstrating the most robust P1-N1-P2 components, as well as previous research suggesting that fronto-medial and centro-frontal regions are the maximal generators of these components in response to FAF (Chen et al., 2012).

For each participant, averaged waveforms were created for the unaltered and the FAF conditions for each electrode. Grand averaged waveforms were created for all conditions by averaging the data from all participants for each electrode.
followed by baseline correction. For all average files for each participant, the maximum amplitude and latency were calculated for the ERP components of the P1-N1-P2 complex. Based on visual inspection of the latency of the most prominent ERP peaks, these components were extracted at time windows from 50-100ms, 100-200ms, and 200-300ms, respectively.

2.6. Statistical Analyses

Preliminary statistical analyses were conducted to investigate the influence of electrode site on ERP amplitudes and latencies. Since ERP responses were not found to differ as a function of electrode site (all comparisons p>.05), electrode site was not included as a factor in subsequent analyses. In addition, behavioral and ERP responses in the 0 cent (unaltered) condition were not found to vary as a function of experimental condition (unpredictable vs. predictable), thus all control responses were collapsed into a single category. In order to investigate the effect of randomly varying the perturbation magnitude within a vocalization, five experimental categories were examined: 0 cent (unaltered), 50-cent (random), 50-cent (non-random), 100-cent (random), and 100-cent (non-random).

To examine the influence of randomly varying perturbation magnitudes within a vocalization, a 5 (experimental condition) x 2 (block order) repeated-measures analysis of variance (RM-ANOVA) was conducted on the amplitudes of the P1-N1-P2 complex using SPSS (v. 19.0). A 4 (experimental condition) x 2 (block order) RM-ANOVA was also conducted on the latencies of the P1-N1-P2 complex as well as the vocal response latencies. The unaltered (0 cent) condition was not analyzed with regards to ERP and vocal latencies as stimuli were not presented during the unaltered trials, thus data were randomly sampled with no true reference, rendering latency information meaningless.

For all RM-ANOVAs, post-hoc least significant difference (LSD) tests were conducted to examine differences in the recorded responses as a function of the experimental condition. The Greenhouse-Geisser (Greenhouse & Geisser, 1959) correction was used in instances where Mauchley’s Assumption of Sphericity was violated. However, original degrees of freedom are reported for ease of interpretation.

3. Results

3.1. Vocal Response Magnitudes and Latencies

The RM-ANOVA conducted to investigate the effect of randomly varying perturbation magnitudes within a vocalization on vocal response magnitude showed a significant main effect of experimental condition (F(4,104)=13.036, p<.001; see Figure 2). There was also a main effect of block order (F(1,26)=13.113, p=.001), as overall participants had larger vocal responses when the 50-cent non-random block occurred before the 100-cent non-random block. Pairwise comparisons indicated that all FAF conditions elicited significantly larger vocal response magnitudes than the unaltered condition (p<.05). In addition, the 50-cent non-random condition elicited significantly smaller responses than all other FAF conditions (p<.05), while the 50-cent random condition elicited smaller responses compared to the 100-cent random condition (p<.05), but not the 100-cent non-random condition (p>.05). Lastly, the 100-cent random condition elicited significantly larger responses than all other conditions (p<.01).

Figure 2: Averaged vocal response magnitudes in cents for all participants as a function of perturbation category and block order.

The RM-ANOVA investigating the influence of experimental condition on vocal response latencies showed a main effect of experimental condition (F(3,78)=2.985, p<.05). Pairwise comparisons indicate that responses were significantly faster in response to non-random 50-cent perturbations, relative to 100-cent perturbations in both conditions (p<.05).

3.2. ERP Responses

3.2.1. P1 Amplitudes and Latencies

A two-way RM-ANOVA was conducted to investigate the influence of experimental condition and block order on P1 amplitudes. The interaction between experimental condition and block order was significant (F(4,104)=2.978, p<.05). This effect appears to be driven by smaller amplitudes in the 50-cent conditions relative to the 100-cent conditions, when the 50-cent non-random condition was presented first, while amplitudes were smaller for the 100-cent conditions relative to the 50-cent conditions, when the 100-cent non-random condition was presented first (see Figure 3). The two-way RM-ANOVA investigating the influence of experimental condition and block order on P1 latency resulted in a main effect of experimental condition (F(3,78)=5.342, p<.01). Pairwise comparisons investigating latency differences across the experimental conditions indicated that the 50-cent random perturbations resulted in slower P1s relative to the other FAF conditions (p<.05).

3.2.2. N1 Amplitudes and Latencies

A two-way RM-ANOVA was conducted to investigate the influence of experimental condition and block order on N1 amplitudes. A significant main effect of experimental condition (F(4,104)=7.180, p<.001) was found (see Figure 3). Pairwise comparisons indicated that all FAF conditions elicited significantly larger N1 amplitudes (absolute value) than the unaltered condition. In addition, the 100-cent random condition elicited significantly larger N1 amplitudes (absolute value) than all other conditions (p<.05).

The two-way RM-ANOVA investigating the influence of experimental condition and block order on N1 latency found a main effect of experimental condition (F(3,78)=5.091, p<.01), as well as block order (F(1,26)=7.702, p<.01). Pairwise comparisons indicate that N1 latencies were longer in the 100-cent random condition relative to the 100-cent non-random condition, however, they were shorter than the 50-cent random condition. In addition, the 50-cent random condition had longer N1 latencies than all other conditions. Post-hoc comparisons investigating block order revealed that
participants who were exposed to the 50-cent non-random condition prior to the 100-cent non-random condition had larger N1 latencies in all experimental conditions relative to the participants who were exposed to the 100-cent non-random condition first.

3.2.3. P2 Amplitudes and Latencies

A two-way RM-ANOVA was conducted to investigate the influence of experimental condition and block order on P2 amplitudes. The experimental condition by block order interaction was significant (F(4,104)=2.928, p<.05; see Figure 3). This effect appears to be driven by larger P2 amplitudes in the 100-cent conditions, relative to the 50-cent conditions, but only when the 50-cent non-random condition is presented prior to the 100-cent non-random condition.

The two-way RM-ANOVA investigating the influence of experimental condition and block order on P2 latency failed to find any significant effects.

4. Discussion

In this study, speakers were exposed to FAF perturbations that were either predictable in magnitude, or unpredictable in magnitude. Behavioural and neurological responses to these FAF perturbations were examined in order to investigate whether being able to predict the magnitude of brief (temporally unpredictable) FAF perturbations altered responses to these perceived speech production errors.

Decreases in the magnitude of vocal responses, smaller N1 amplitudes, and shorter vocal, P1 and N1 response latencies following non-randomly varying FAF perturbation magnitudes, supports the notion that experimentally induced predictability can modulate responses to FAF. We suggest that increasing the predictability of the magnitude of FAF perturbations makes these perturbations easier to distinguish from speaker-generated vocal variability. After repeated exposure to these predictable FAF perturbations, auditory feedback is deemed unreliable, and weighting of the feedforward system increases. As a result of the increased weighting of the feedforward control system, deviant auditory feedback is less salient, resulting in smaller responses, both behaviourally and neurologically. Modifying the weighting of the feedback and feedforward control systems in different contexts is physiologically advantageous. Increasing the weighting of the feedback system is advantageous in situations where the information from auditory feedback is reliable and can be used to update the mapping of the internal model. For example, throughout development auditory feedback is required to maintain the mapping of the internal model as growth related changes to the articulators, vocal folds, musculature, and lung capacity occur (Civier et al., 2010; Guenther, 2006). Even in adulthood, increased weighting of the feedback control system can help to rapidly update the mapping of the internal model following the acquisition of dental appliances, or as aging related changes occur to the articulators, vocal folds, musculature, and lung capacity. On the other hand, increased weighting of the feedforward system can also be physiologically advantageous. As development halts, the mapping of the internal models should remain relatively stable. For this reason, the additional information provided by auditory feedback becomes redundant. Increasing the weighting of the feedforward control system not only increases the fluidity of speech, but also reduces susceptibility to externally generated noise, and frees up attentional resources for the processing of potentially important stimuli, rather than predictable auditory errors (Heinks-Maldonado et al., 2005).

5. References


Civier, O., S.M. Tasko, F.H. Guenther (2010). “Overreliance on auditory feedback may lead to sound/syllable repetitions: Simulations of stuttering and fluency-inducing conditions with a neural model of speech production”. In: Journal of fluency disorders, 35, pp. 246-279.


Representation of German binomials: Evidence from speech production

Jana Schlegel, Sophie Egger & Bettina Braun

Department of Linguistics, University of Konstanz, Germany

{Jana.2.Schlegel, Sophie.Egger, Bettina.Braun}@uni-konstanz.de

Abstract

Binomials (e.g., German Ebbe und Flut; ‘ebb and flow’) are a common phenomenon of many languages, but little is known about how they are stored, produced and processed. We tested the production of German nominal binomials and compared their onset latency to the onset latency of forms in which one part of the binomial was replaced by an alternative constituent that was phonologically similar to the original (e.g., Brat as an alternative to Flut). Studies on multi-word phrases and idioms suggest that such frequently occurring expressions are accessed faster and produced with shorter duration than infrequent forms, suggesting that they are accessed as a single unit, rather than word-by-word. We hypothesized single storage for frequent binomials and expected shorter onset latencies and constituent durations for original binomial forms than for alternatives. This is what our results show. They hence lend support that advocate single representations for frequent constructions.

Keywords: German, binomial, onset latency, production, constituent duration, association strength, lexical frequency

1. Introduction

Binomials represent a linguistic phenomenon that is present in many languages of the world (e.g., Khatibzadeh & Sameri, 2013; Malkiel, 1959). They are part of everyday language and occur for example in book or film titles, poems and product names and are applied by the advertising industry in order to arouse interest (H.-G. Müller, 2009). But despite the cross-linguistic pervasiveness, most of the previous studies have focused on English binomials (e.g., Cooper & Ross, 1975; Pinker & Birdsong, 1979). For German, the first extensive work, which listed up to 1300 binomials, has only been published in 2009 by H.-G. Müller (2009). Furthermore, the main focus of previous studies was laid on the factors that determine the order of the constituents of a binomial (e.g., Cooper & Ross, 1975; Malkiel, 1959; G. Müller, 1997; Lenz, 2002). On the other hand, we still know very little about how binomials are stored, processed and produced by language users. In the current paper we aim at filling this gap by presenting two production experiments that investigated the mental representation of German binomials (e.g., Ebbe und Flut, ‘ebb and flow’).

Recent studies on multi-word phrases and idioms suggest that frequently occurring idioms and phrases are stored and accessed as one unit (e.g., Spranger, Levelt, & Kempen, 2006; Janssen & Barber, 2012). Furthermore these analyses have shown that such high-frequent multi-word units are easier to access leading to shorter onset latencies and constituent durations (e.g., Bybee, 2010; Fenk-Oczlon, 2001; Janssen & Barber, 2012; Tabossi, 2009). Due to the frequent occurrence of binomials, we hypothesize that binomials are also stored and accessed as one entity in comparison to non-binominal coordinations that are structurally similar but do not frequently occur together. To test this hypothesis, we investigated whether common German nominal binomials (e.g., Ebbe und Flut, ‘ebb and flow’) are initiated earlier and with shorter constituent durations than alternative forms in which one of the constituents is replaced by an alternative phrase (e.g., Brat (brood) as an alternative to Flut (flow)). Since more strongly associated binomials may form stronger exemplars and therefore affect onset latencies and constituent durations, we collected association strength measures, which have been argued to be a viable indication of how frequently the constituents co-occur in daily language use (e.g., Haskins, Yonelinas, Quamme, & Ranganath, 2008; Jenkins & Russell, 1952; Tanaka-Ishii, & Terada, 2011). We predict that more strongly associated binomials are affected more by replacing a constituent than binomials with a weaker association.

2. Experiment 1

In Experiment 1, the second constituents of the original binomials were replaced by words that were phonologically similar but that do not occur frequently with the first constituent. This allowed us to compare the latency for initiating the producing of the two constituents in the original and the alternative form with an identical first constituent across conditions (e.g., Ebbe und Flut vs. Ebbe und Brat). Furthermore, the duration of the first constituent was compared across conditions.

2.1. Methods

2.1.1. Materials

The original binomials were selected on the basis of a web based association experiment with 35 native speakers of German (17 female, average age 25 years, SD = 2.5). We presented the first constituent of 33 well described binomials that frequently occur in German (taken from Müller, 2009 and Hofmeister 2001, 2010) followed by the conjunction and (e.g., Ebbe und…: E: ebb and…) and asked participants to indicate the first word that came to their mind. From these 33 binomials we selected the 10 items with the highest association strength (on average 94%, SD = 4.3) and the 10 items with the lowest association strength (on average 34%, SD = 6.6). Afterwards the second constituent of the binomials was replaced by a phonologically similar noun. The replacement had the same syllable number and stress pattern, a similar phonotactic structure as the original constituent of the binomial (e.g., Flut vs Brut; see Table 1 in Appendix) and was matched to the original second constituent in lexical frequency according to dlexDB (Geyken, Hanneworth, & Kliegl, 2012). The original second constituents had 120 occurrences per million (o.p.m), the replacements 53 o.p.m. This difference was not significant (t(19) = 1.3, p × 0.2).

Apart from these 20 experimental nominal constructions, we selected six practice coordinations with proper names (e.g., Andreas and Pia).
2.1.2. Participants

Twelve monolingual German speakers (10 female) participated voluntarily (average: 25 years, SD: 3.6); they were unaware of the purpose of the experiment. None of them had taken part in the association experiment.

2.1.3. Procedure

The experimental list contained all original binomial and alternative coordinations (within-subject design). The list started with the six practice trials. The order of the other trials was pseudo-randomized with the constraint that the two versions of a given binomial (its original and alternative) were separated by at least 5 other trials. For one half of the binomials, the original was presented first, for the other half, the alternative was presented first.

Participants were seated in a soundproof cabin wearing headphones with an integrated microphone, which was used to record their productions during the session. In order to avoid read speech, we first presented the second part of the binomial, followed by the first part. Participants had the task to assemble the intended form in the reverse (and hence correct) order. Each trial started with a fixation cross, which appeared at the centre of the screen for 250ms. After a pause of 2s (showing a blank screen), the second constituent of the binomial (e.g., Flut) or the replacement (e.g., Brut) was presented for 350ms in black Arial 42font on white background at the centre of the screen. After another pause of 2s, the first constituent together with an ampersand appeared (e.g., Ebbe &) in the same font centred on screen. Together with the visual onset of the second constituent, a beep of 10ms duration was played to the left channel of an M-Audio Microtrack II recorder. Participants were instructed to remember the second part of the binomial and to produce the binomial in the correct order as quickly as possible. Their productions were recorded on the right channel of the recorder (44.1kHz, 16Bit). The ampersand was used to unambiguously mark the first constituent of the nominal construction.

2.2. Results

The recordings were manually annotated at the lexical word level, using broadband spectrograms and standard segmentation criteria (Turk, Nakai & Suguhara 2006). In particular, we manually measured participants' onset latencies relative to the onset of the visual presentation as well as the duration of the first constituent. Onset latencies and the duration of the first constituent of the productions were statistically analysed using linear mixed effects regression models with coordination type (original binomial vs. alternative) and association strength of the original binomial (strong vs. weak) as fixed factors and participants and items, as crossed random factors allowing for random adjustments of intercepts and slopes (Barr, Levy, Scheepers, & Tily, 2013). We additionally included trial number and attempt (first or second encounter of a coordination) as control predictors. $P$-values were calculated by comparing a model with a given factor (or interaction) to a model that lacked that factor (or interaction), all else being equal (using the anova-function in R). Results for onset latencies showed a significant main effect of coordination type ($\beta = -0.035$, $SE = 0.009$, $p < 0.005$), but no effect of association strength and no interaction (both $p$-values $> 0.2$). Similarly, attempt and trial number did not have an effect (all $p$-values $> 0.3$). Original binomials were initiated on average 35ms earlier than the alternative forms (420ms vs. 385ms, see left-hand bars of Figure 1).

For the duration of the first constituent, the model showed a significant effect of coordination type as well ($\beta = -0.009$, $SE = 0.004$, $p = 0.02$). The first constituents of original binomials were on average 9ms shorter than the same constituents in the alternative constructions (252ms vs. 243ms, see Figure 1).

In order to exclude the alternative possibility that the results are caused by semantic priming from the second constituent of the binomial (which was shown on screen first) on the first constituent, we additionally collected the backwards association strength, i.e. the association strength between the second constituent of the binomial (original constituent vs. replacement) and the original first constituent. We tested a different set of 96 participants in a web-based association experiment (58 female, average 26 years, SD = 2.8). Participants were visually presented with the second constituent (original or replaced, manipulated within-subjects), followed by the conjunction and (e.g., Flut und ...) and had to type in the first word that came to their mind. We calculated the backwards association strength between the presented second constituent (e.g., Flut or Brut) and the original first constituent (e.g., Ebbe). The average backward association strength ranged between 0% and 92%. It was on average 44% for originals and 0% for alternatives (t(19)) = 2.7, $p < 0.05$). We then selected the 10 binomials, for which the original had the lowest backwards association strengths (on average 25.4%, $SD = 25.3$) and their alternatives and ran the model again. Importantly, we also see an effect of coordination type on onset latencies and on the duration of the first constituent for this subset ($\beta = -0.03$, $SE = 0.001$, $p < 0.05$ and $\beta = -0.02$, $SE = 0.007$, $p < 0.005$, respectively). The effect sizes are comparable to those of the complete data set.

2.3. Discussion

The results of the first experiment show a significant effect of coordination type on the onset latency of original binomials and on the duration of the first constituent. This finding is in line with our hypothesis that binomials are stored as one entity in the mental lexicon and are therefore accessed faster and produced with a shorter duration. Furthermore, our results suggest that these findings also hold for those binomials that have a weak associative connection. Note that our analyses did not show an effect of attempt (the first or second encounter of a coordination), which suggest that the current within-subjects design is a useful method to study the representation of

![Figure 1: Average onset latency and duration of the first constituent, split by coordination type (alternative vs. original), as calculated by the statistical model. Whiskers show standard errors.](image-url)
binomial constructions. In order to corroborate our findings we tested whether the second constituents also have shorter durations in original binomials compared to alternatives.

3. Experiment 2

Experiment 2 investigated whether coordination type also effects the duration of the second constituent in a binomial. Therefore we created alternative coordinations in which the first constituent was replaced by an alternative constituent.

3.1. Methods

3.1.1. Materials

For each of the 20 binomials of Experiment 1, we created 20 novel alternatives by replacing the first constituent (e.g., *Ebbe und Flut* had the alternative: *Treppe und Flut*). As in Experiment 1, the original constituents and the replacements had the same syllable structure, stress pattern and a similar phonotactic form and did not differ in lexical frequency (189 o.p.m for originals compared to 115 o.p.m for alternatives, a difference that was no significant: \( t(19) = 1.7, p < 0.3 \)).

3.1.2. Participants

Twelve monolingual German native speakers (11 female) participated voluntarily (average: 24 years, SD = 2.8). None of them took part in any of the experiments reported above; they were not informed on the purpose of the experiment.

3.1.3. Procedure

The procedure, the experimental lists, and the recording setting were identical to Experiment 1.

3.2. Results

The recordings were manually coded at the lexical word level using the same criteria as in Experiment 1. The duration of the second constituent was analysed using a linear mixed effects regression model with coordination type (original binomial vs. alternative) as fixed factor and participants and items as crossed random factors allowing for random adjustments of intercepts and slopes (Barr et al., 2013). Results showed a significant main effect of coordination type (\( B = -0.016, SE = 0.005, p = 0.0004 \)).

![Figure 2: Average duration of the second constituent split by coordination type (alternative vs. original), based on the statistical model. Whiskers show standard errors.](image)

The second constituent of the original binomials was on average 16ms shorter than the second constituent in alternatives (354ms vs. 338ms, see Figure 2). Similarly to Experiment 1, attempt (first or second presentation of a binomial) and trial number did not have significant effects on the duration of the second constituent (both \( p \)-values > 0.1).

3.3. Discussion

Experiment 2 showed that participants produced original binomials with a shorter second constituent compared to coordinations in which the first constituent was replaced by an alternative noun. The current findings hence corroborate the findings reported in Experiment 1 and lend further support to the interpretation that common binomials are stored and accessed as one unit, which leads to shorter durations in comparison to non-binomial coordinations.

4. General Discussion

We presented two production experiments that probed the representation of binomials by means of onset latencies and constituent durations. We showed shorter onset latencies for the initiations of German binomials, as compared to coordinations in which one of the constituents was replaced by a structurally similar noun. Furthermore, the two constituents of an original binomial were produced with shorter durations compared to the alternative coordinations. Together, these results are in line with usage-based accounts (e.g., Bybee, 2010; Tomasello, 2003) and certain other proposals regarding the representation of frequent multi-word expressions (e.g., Sprenger et al., 2006), which predict shorter onset latencies and constituent durations for high frequency words and constructions (e.g., Fenk-Oczlon, 2001; Tabossi et al., 2009; Janssen & Barber, 2012). Note that our results cannot be explained by the lexical frequencies of the individual constituents of the binominal alone, but are the result of the coordination of the two constituents into a frequent binomial.

Note that the two constituents of a binominal are in many instances also semantically related (e.g., Cooper & Ross, 1975). Therefore, an alternative interpretation for our results is that it is the semantic relationship between the two constituents in a binominal that leads to shorter onset latencies and shorter constituent durations (e.g., Swinney et al. 1979). A semantic interpretation was also provided by Jolsvai, McCauley, and Christiansen (2013) who found no frequency effects for the onset latencies of multi-word phrases. However, our results show that shorter onset latencies and constituent durations also occur in a subset of the data, in which the semantic association between the two constituents is low. A further argument against a pure semantic priming account is that the association strength between the two constituents in a binominal (as assessed in a free association task) was not a significant predictor for onset latencies. Therefore, we are confident that our results point to the storage of German binomials as a single unit (like assumed for other, frequently occurring phrases (e.g., Bybee, 2010; Tabossi et al., 2009), rather than to a semantic priming account. However, future studies will have to manipulate the semantic relationship of the constituents in a binominal more explicitly to exclude an explanation that is purely based on semantic priming.
5. Appendix

Table 1: Original binomials with respective alternative constituent, split for strong (top half) and weak (bottom half) association, English translations in italics

<table>
<thead>
<tr>
<th>Original binomial</th>
<th>Alternative to constituent 1</th>
<th>Alternative to constituent 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebbe &amp; Flut /ebb &amp; flow</td>
<td>Trepp/stairs</td>
<td>Bröt/brood</td>
</tr>
<tr>
<td>Tag &amp; Nacht /day &amp; night</td>
<td>Prag/Prague</td>
<td>Pacht/lease</td>
</tr>
<tr>
<td>Pech &amp; Schwefel /bitumen &amp; sulfur</td>
<td>Blech/plate</td>
<td>Frevel/iniquity</td>
</tr>
<tr>
<td>Blitz &amp; Dommer /lightning &amp; thunder</td>
<td>Sitz/seat</td>
<td>Sommer sonner</td>
</tr>
<tr>
<td>Obst &amp; Gemüse /fruits &amp; vegetables</td>
<td>Probst/provoist</td>
<td>Kombüse/galley</td>
</tr>
<tr>
<td>Leib &amp; Seele /body &amp; soul</td>
<td>Weib/broad</td>
<td>Kehler/throat</td>
</tr>
<tr>
<td>Rat &amp; Tal /advice &amp; act</td>
<td>Staatsstaat</td>
<td>Staatsstaat</td>
</tr>
<tr>
<td>Hüle &amp; Fürle /sleeve &amp; wealth</td>
<td>Gülle/duary</td>
<td>Gülle/duary</td>
</tr>
<tr>
<td>Mann &amp; Frau /man &amp; woman</td>
<td>Banan/ban</td>
<td>Staubam</td>
</tr>
<tr>
<td>Berg &amp; Tal /mountain &amp; valley</td>
<td>Werk/factory</td>
<td>Stahl/steel</td>
</tr>
<tr>
<td>Haus &amp; Hof /home &amp; yard</td>
<td>Mause/mouse</td>
<td>Boot/boat</td>
</tr>
<tr>
<td>Feuer &amp; Flamme /fire &amp; flame</td>
<td>Steuer/dax</td>
<td>Tanne/ftir</td>
</tr>
<tr>
<td>Tür &amp; Angel /door &amp; hinge</td>
<td>Küter/kur</td>
<td>Mangel/lack</td>
</tr>
<tr>
<td>Luft &amp; Liebe /air &amp; love</td>
<td>Duft/duour</td>
<td>Fliege/fly</td>
</tr>
<tr>
<td>Land &amp; Leute /country &amp; people</td>
<td>Hand/hand</td>
<td>Meute/mob</td>
</tr>
<tr>
<td>Rand &amp; Band /edge &amp; strap</td>
<td>Wand/wall</td>
<td>Wand/wall</td>
</tr>
<tr>
<td>Not &amp; Elend /need &amp; misery</td>
<td>Tod/death</td>
<td>Gegend/region</td>
</tr>
<tr>
<td>Saft &amp; Kraft /juice &amp; strength</td>
<td>Haft/custody</td>
<td>Haft/custody</td>
</tr>
<tr>
<td>Sein &amp; Schein /being &amp; presence</td>
<td>Bein/leg</td>
<td>Bein/leg</td>
</tr>
<tr>
<td>Herz &amp; Nieren /heart &amp; kidneys</td>
<td>Schmerz/pain</td>
<td>Viren/viruses</td>
</tr>
</tbody>
</table>

6. References


Social Factors in Convergence of F1 and F2 in Spontaneous Speech

Antje Schweitzer and Natalie Lewandowski

Institute for Natural Language Processing, Stuttgart University, Germany
{antje.schweitzer,natalie.lewandowski}@ims.uni-stuttgart.de

Abstract

We present results on phonetic convergence of normalized F1 and F2 values in German spontaneous speech (46 dialogs, 20.8 hrs of speech). We are interested in the influence of social factors, specifically of mutual likeability and competence ratings, on convergence. To this end we fitted linear mixed models with speakers’ F1 and F2 values, using partners’ averaged values as well as the mutual social ratings as predictors. Our results show significant general convergence effects as well as significant effects of the interaction between partners’ F1 and F2 values and the social ratings on speakers’ productions of F1 and F2. This indicates that vowel formants are subject to phonetic convergence in spontaneous speech, and that social factors have an effect on the degree of convergence.

Keywords: phonetic convergence, vowel formants, spontaneous speech

1. Introduction

Phonetic convergence is the process of adapting one’s speech to an interlocutor. The opposite, i.e., the assumption of a speaking style that differs from that of the interlocutor, is called divergence. In both cases, the perception of the interlocutor’s speech affects a speaker’s current production targets. An issue related to convergence, sometimes even considered equivalent to convergence, is imitation, which occurs when speakers’ production targets are influenced by properties of stimuli that they have been exposed to before.

According to Communication Accommodation Theory (CAT, e.g. Giles and Smith 1979; Giles, Coupland, and Coupland 1991; Shepard, Giles, and Le Poire 2001; Giles and Ogay 2006), the adaptation seen in convergence or divergence is a dynamic process and affects not only speech, but communicative behavior in general (i.e. linguistic and phonetic features, but also paralinguistic aspects). CAT proposes that convergence decreases social distance between interlocutors and thus reflects a speaker’s (often unconscious) need for social integration or identification with the interlocutor’s social group (Giles, Coupland, and Coupland 1991). In contrast, divergence is caused by the need to distance oneself from the interlocutor’s group. Interlocutors may also converge to increase intelligibility and efficiency of communication (Triandis 1960; Natale 1975; Gallois et al. 1995). Thus, social factors and the communication setting are clearly important when investigating convergence.

However, most recent studies on phonetic convergence use rather controlled and limited speech material, often drawing on methodology that is typically used in imitation research, without real conversational interaction, or focus on only specific target words or phrases in conversations (Babel 2010; Abrego-Collier et al. 2011; Kim, Horton, and Bradlow 2011; Babel 2012; Pardo et al. 2012). Few recent studies on convergence use larger-scale fully annotated corpora such as the Columbia Games Corpus (Levitan and Hirschberg 2011). In our opinion, testing the reality of convergence “in the wild” by investigating convergence on such corpora is indeed overdue, but there is one possible drawback in using game task corpora to this end: Given that efficiency of communication is a prerequisite for successfully playing such a game task, the question arises whether convergence in a game corpus may be a consequence of the game concept instead of a natural phenomenon in conversation.

Furthermore, social factors are assumed to be central in convergence, but to our knowledge there are no corpora to date which take social aspects of the conversation into account while providing data from completely free, spontaneous conversations. To close this gap, we have created the German Conversations (GECO) database, which provides data on speakers’ mutual social assessment (in terms of likeability and competence), in addition to large-scale fully annotated recordings of high audio quality. In this paper, we investigate convergence of vowel formants in this corpus.

2. Speech data

GECO consists of spontaneous conversations between previously unacquainted female German speakers on topics of their choice. Most speakers were students between age 20 and 30. Each dialog lasted approx. 25 minutes. Participants wore AKG HSC271 head-sets with rubber foam windshields while talking to each other in a sound-attenuated booth. We recorded about one half of the dialogs in a unimodal (UM) condition, where speakers could not see each other, and the other half in a multimodal (MM) condition, where speakers could see each other through a transparent screen. There are 22 dialogs (approx. 10.3 hours of dialog) in the UM condition and 24 (approx. 10.5 hours) in the MM condition. Subjects were naïve to the research questions; in both conditions, they were told that the purpose of the study was to research how small talk between strangers works. They were provided with a list of potential topics to ease conversation, but were explicitly told that they were completely free to choose other topics as well. In fact, participants rarely consulted the list. The recordings were automatically annotated on the segment, syllable, word, and prosodic levels. The resulting corpus amounts to 20.8 hrs. of dialog, with approx. 250,000 words, 360,000 syllables, and 870,000 phones.

2.1. Social factors

As elaborated above, it is well accepted that the degree of accommodation (and its direction, i.e., divergence or convergence) is related to social factors (e.g. Giles and Smith 1979; 1

1The GECO corpus is freely available for non-commercial use at http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.en.html
3. Methodology

3.1. Data processing

We extracted F1 and F2 values for each non-reduced monophthong in our data, along with vowel identity, duration, word stress, word frequency, speaker ID, listener ID, F0 at vowel midpoint (as calculated by get_f0 from the ESPS software package). For calculating F1 and F2 we used Praat (Boersma and Weenink 2014) to extract the first two formants using the “Burg” method, allowing a maximum of five formants, an expected maximum of 5500 Hz, a window length of 25 ms, and a pre-emphasis of approx. 0.75, which indicates that its raw F1 is relatively high. The Asterisk in Fig. 1 would then have an F2’ of 0, which indicates a hypothetical example token located at the point indicated by the asterisk in Fig. 1 would then have an F1’ value of 2 as indicating that approx. 97.5% of all tokens of that vowel were produced with a lower F1 than this token, thus the token is located at the upper edge of the distribution in terms of F1.

In this way, F1’ and F2’ indicate each vowel token’s position relative to all speakers’ tokens of the same vowel. Figure 1 illustrates the normalization technique: it depicts the vowel space of all speakers in terms of F1 and F2. The boxes around each vowel indicate the region of one standard deviation above and below the mean. They can thus be interpreted as reference frames for normalization: Values falling at the edges of these boxes yield normalized values of +1 or -1. For instance, the yellow box highlights the region for all /a:/ vowels, and the red arrows then indicate the normalized axes for /a:/ vowels. A hypothetical example token located at the point indicated by the asterisk in Fig. 1 would then have an F2’ of 0, which indicates that its raw F2 is equal to the mean for all speakers, and an F1’ of approx. 0.75, which indicates that its raw F1 is relatively high compared to that of all other /a:/ tokens.

3.2. Normalization

As the formant values of course are vowel-specific, we scaled and centered (i.e. z-scored) all formant values using vowel-specific means and standard deviations. Note that while this may sound reminiscent of Lobanov’s (1971) speaker normalization procedure, our normalization technique is actually different: The aim in applying Lobanov’s technique is to express the formant values in terms of their location in a specific speaker’s vowel space. The aim of our technique is to express the formant values in terms of their location in the region that all recorded speakers used for this specific vowel, because we need the same reference frame for normalization in order to assess whether speakers used similar values. The parameters resulting from our transformation will be referred to as F1’ and F2’, respectively. Thus a value of 0 for F1’ for instance indicates that the respective vowel token was produced with an F1 that is exactly average across all speakers for this vowel, while a value of 2 indicates that the vowel token was higher than this average by two standard deviations. Bear in mind that for normally distributed data, only 2.5% of the values are more than 2 standard deviations higher than the mean. As our F1 and F2 values were approximately normally distributed, we can then interpret an F1’ value of 2 as indicating that approx. 97.5% of all tokens of that vowel were produced with a lower F1 than this token, thus the token is located at the upper edge of the distribution in terms of F1.

In this way, F1’ and F2’ indicate each vowel token’s position relative to all speakers’ tokens of the same vowel. Figure 1 illustrates the normalization technique: it depicts the vowel space of all speakers in terms of F1 and F2. The boxes around each vowel indicate the region of one standard deviation above and below the mean. They can thus be interpreted as reference frames for normalization: Values falling at the edges of these boxes yield normalized values of +1 or -1. For instance, the yellow box highlights the region for all /a:/ vowels, and the red arrows then indicate the normalized axes for /a:/ vowels. A hypothetical example token located at the point indicated by the asterisk in Fig. 1 would then have an F2’ of 0, which indicates that its raw F2 is equal to the mean for all speakers, and an F1’ of approx. 0.75, which indicates that its raw F1 is relatively high compared to that of all other /a:/ tokens.

3.3. Statistical analysis

Our aim is to find out whether speakers’ F1 and F2 values are influenced by their partners’ F1 and F2 values. Specifically, if there was a positive relationship (i.e. if speakers produce higher values when confronted with higher partners’ values),
this would indicate convergence. A negative relationship on the other hand would indicate divergence. To assess the relationship between partners’ and speakers’ F1’ and F2’ values, we performed two sets of linear mixed effects analyses using R (R Core Team 2013) and the lme4 package (Bates et al. 2014). The dependent variables were F1’ and F2’, respectively. If speakers converge to their partners, we would expect that partners’ F1’ and F2’ productions are significant predictors of speakers’ F1’ and F2’. As it is not yet clear how much context is needed for speakers to converge, i.e., how many vowels must have been perceived before speakers’ productions are affected, we do not want to make any assumptions as for exactly which of the partner’s preceding tokens affect each produced vowel. Therefore, while we predicted F1’ and F2’ for each vowel token for every speaker and every dialog, we averaged partners’ F1’ and F2’ values across that whole dialog. These averaged values were used as predictors. Thus F1’ and F2’ values of all vowel tokens of a speaker in a dialog were considered once individually as dependent variables, and then again indirectly when they contributed to the average F1’ or F2’, which then served as predictor variables for all vowel tokens of the other speaker in the same dialog. All F1’ and F2’ values as well as the averages were centered prior to fitting the models.

To control for random factors (for instance reduction effects due to stress, vowel duration, and word frequency, but also speaker-specific effects on vowel formants) we included intercepts for speaker, as well as by-vowel slopes for duration, stress, and word frequency. All random factors were justified, as confirmed by likelihood ratio tests for each factor, always comparing the model without the factor in question to the full model. This was done once at the beginning, including only partners’ F1’ or F2’ averages as fixed factors. We then iteratively added the social factors and their interactions as fixed effects to both models, always confirming that including the factor was justified by way of likelihood ratio tests of the model with the factor in question compared to the model without the factor in question.2 For the two winning models, we re-checked that all random effects were still justified for these richer models.

To assess the significance of the fixed effects in the winning models, we used the “Wald” method of the confint function provided by the lme4 package (Bates et al. 2014). This function allows approximation of confidence intervals based on the estimated local curvature of the likelihood surface. We chose a confidence level of 0.975 (Bonferroni correction for two tests, one for F1’, one for F2’). We regard effects as significant if the estimated confidence interval does not contain zero at this confidence level.

4. Results

The best model both for F1’ and F2’ was the model which included as predictors (i) partners’ average F1’ (or F2’) scores, (ii) the likeability score for the partner (iii) the competence score for the partner, and (iv) their interactions. Visual inspection of the residual plots of each winning model revealed no obvious deviations from normality and homoscedasticity.

Estimates for the coefficients in the two winning models are given in Table 1. They exhibit similar patterns. In both cases, we observed a general convergence effect, irrespective of the social ratings: we observed a positive coefficient for the main effect of partner’s score (lines labeled partner in Table 1). The effect was more pronounced (i.e. with a higher coefficient) in case of F1’ than in case of F2’, but the effect was significant at a level of 0.975 in both cases. This means that the default behavior across all dialogs was convergence.

In addition, there are interactions of likeability and competence with partners’ scores which are in opposition: there is a positive coefficient for the interaction between likeability and partners’ scores (lines labeled partner:likeability), i.e., the more a speaker liked her partner, the more “influence” the partner’s score had on the speaker’s productions, i.e. the general convergence effect described above is strengthened with higher likeability scores. The effect is significant at a level of 0.975. We find the opposite for the competence scores: for both F1’ and F2’ we observe negative coefficients for the interaction between competence and partners’ scores (lines labeled partner:competence), i.e., the more competent a speaker rated her partner, the lower the contribution of the partner’s score in predicting the speaker’s F1’ or F2’, i.e. the general convergence effect is weakened for higher competence scores. This effect was also significant at a level of 0.975 in both cases. In case of F1’, there was also a small but significant positive effect of the three-way interaction between partners’ average F1’ and the competence and likeability scores, while there was a negative effect in case of F2’ (lines labeled partner:likeab.:comp.).

It should be noted that we also observed main effects of the social ratings and their two-way interaction on speaker’s F1’ and F2’, irrespective of the partner’s score, which we found surprising. They indicate for instance that higher competence ratings for the partner lowered speakers’ F1’ and F2’ in general (lines labeled likeab.:competence). We currently have no explanation for these findings.

Table 1: Coefficients of the winning linear mixed models. Estimates for the coefficients are in the second column, upper and lower bounds of the corresponding confidence intervals are listed in the third and fourth columns. The last column indicates whether we consider the effect significant (*) or not significant (n.s.).

<table>
<thead>
<tr>
<th>F1’ results</th>
<th>Coefficient</th>
<th>estim.</th>
<th>upper</th>
<th>lower</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.439</td>
<td>-0.968</td>
<td>0.090</td>
<td>n.s.</td>
<td></td>
</tr>
<tr>
<td>partner</td>
<td>0.143</td>
<td>0.116</td>
<td>0.169</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>likeability</td>
<td>-0.005</td>
<td>-0.010</td>
<td>-0.001</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>competence</td>
<td>0.021</td>
<td>0.017</td>
<td>0.024</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>partner:likeability</td>
<td>0.043</td>
<td>0.029</td>
<td>0.056</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>partner:competence</td>
<td>-0.063</td>
<td>-0.077</td>
<td>-0.049</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>likeab.:competence</td>
<td>0.002</td>
<td>0.001</td>
<td>0.003</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>partner:likeab.:comp.</td>
<td>0.007</td>
<td>0.003</td>
<td>0.011</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F2’ results</th>
<th>Coefficient</th>
<th>estim.</th>
<th>upper</th>
<th>lower</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.082</td>
<td>-0.097</td>
<td>0.262</td>
<td>n.s.</td>
<td></td>
</tr>
<tr>
<td>partner</td>
<td>0.043</td>
<td>0.025</td>
<td>0.061</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>likeability</td>
<td>-0.006</td>
<td>-0.010</td>
<td>-0.002</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>competence</td>
<td>0.011</td>
<td>0.008</td>
<td>0.015</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>partner:likeability</td>
<td>0.014</td>
<td>0.003</td>
<td>0.025</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>partner:competence</td>
<td>-0.020</td>
<td>-0.031</td>
<td>-0.009</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>likeab.:competence</td>
<td>0.000</td>
<td>-0.001</td>
<td>0.001</td>
<td>n.s.</td>
<td></td>
</tr>
<tr>
<td>partner:likeab.:comp.</td>
<td>-0.003</td>
<td>-0.005</td>
<td>0.000</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>
5. Discussion
To summarize the results presented above, the main finding is that there is a general convergence effect both for F1’ and F2’, which is strengthened for increased likeability scores and weakened for increased competence scores. As mentioned above, our speakers’ mutual ratings were mostly positive. Also, speakers usually indicated that they found the dialogs pleasant. Thus, the general convergence effect may be a consequence of the homogeneity of the participant group in terms of gender, age, and occupation. Irrespective, our results confirm that convergence occurs naturally in fully spontaneous dialogs, and that it can be detected even using fully uncontrolled speech material.

On a more abstract level, the results clearly confirm the relevance of social factors in convergence: For both F1’ and F2’ including the social factors as fixed effects improved the fit of the models in all cases. In addition, the interactions between mutual social ratings and partner’s F1’ and F2’ scores nicely demonstrate that social factors affect the degree of convergence. This suggests that accounts of phonetic convergence should acknowledge social factors and speaks against a purely biological account of convergence.

Concerning the asymmetry of likeability and competence effects, we can currently only speculate on possible causes. Even though the two variables were correlated (Pearson’s r=0.70), there seem to be subtle differences between likeability and competence. We would argue that competence is a more competitive asset than likeability—people are more likely to compete with respect to competence than with respect to likeability. Some related evidence comes from investigating backchannel frequency in the GECO corpus using linear regression models (Schweitzer and Lewandowski 2012). We found that the more backchannels speakers produced, the more competent and likeable they found their partner, i.e. the effects of partners’ likeability and competence on speakers’ production of backchannels were symmetric. However, we also found that the more backchannels speakers produced, the less competent they tended to be rated themselves by their partners—there was a marginally significant (t(46)=-1.95, p=0.058, $\beta=-0.37$) negative relationship between how many backchannels speakers produced and how competent they were perceived by their partners. This effect was not present for likeability. Assuming that producing many backchannels is in a way similar to converging to the partner, both signaling appreciation in some way, it might be that the findings in the present paper are related to these earlier findings. It is possible that speakers are intuitively aware of the negative relationship between showing (maybe too much) appreciation and the impression of competence on the partner. Thus, when talking to a more competent partner, speakers might be inclined to be more subtle or careful in converging. As no adverse effect of showing appreciation on one’s impression of likeability needs to be feared, and as we would not expect that conversation partners are competing with respect to likeability in the first place, there is no need for speakers to reduce convergence when talking to more likeable partners. We hope to shed more light on the asymmetric effects of likeability and competence in the future.

6. Acknowledgements
This study is part of the project Phonetic Convergence in Spontaneous Speech within the SFB 732 funded by the German Research Foundation (DFG).

7. References
On the inter-dependence of tonal and vocalic production goals in Chinese

Jason A. Shaw1,4, Wei-rong Chen2, Michael I. Proctor3, Donald Derrick1, Elita Dakhoul1

1MARCS Institute, University of Western Sydney, Australia  
2National Tsing Hua University, Taiwan  
3Macquarie University, Australia  
4School of Humanities and Communication Arts, University of Western Sydney, Australia

J.Shaw@uws.edu.au g944710@oz.nthu.edu.tw michael.proctor@mq.edu.au D.Derrick@uws.edu.au

Abstract

We studied tone-vowel coproduction using Electromagnetic Articulography (EMA). Fleshpoints on the tongue and jaw were tracked while native Chinese speakers (n = 6) produced three vowels, /a/, /i/, /u/, combined with four Chinese tones. We found differences in tongue position across tones for /a/ and for /i/ but not for /u/. The low and rising tones patterned together in conditioning lower tongue blade (TB) position for /a/ and a higher TB position for /i/. This pattern suggests a degree of inter-dependence between tonal and vocalic targets. The effect of tone on TB height was mediated by jaw movement such that, even as TB sensor position varied across tones, the Euclidean distance between TB and jaw sensors within each vowel remained stable. Thus, for this set of Chinese vowels, there is a relational invariance between active articulators, tongue and jaw. When viewed in terms of this relation, vowel and tonal targets appear to be completely independent.

Keywords: tones, vowels, speech production models, Chinese, EMA, coarticulation

1. Introduction

Models of speech production generally assume that the glottal source and the supra-glottal vocal tract filter are independent (Fant 1960; Stevens 1998) – an assumption implicit in models of syllable structure in which vowel quality and tone are independent (e.g. Yip 2002; Duanmu 2007; Gao 2009). In contrast, traditional Chinese phonology divides the syllable into two non-decomposable parts: an ‘initial’ (shèngmŭ) and a ‘final’ (yuǎnmŭ) (e.g., Chao 1968). The ‘initial’ is the first consonant of a syllable. The ‘final’ includes the nuclear vowel, tone and optional coda into a single unit. More holistic supraphonic units, such as the finals of traditional Chinese phonology, are consistent as well with more contemporary exemplar-based models that posit word-specific phonetics or online abstraction over exemplars of various-sized units (e.g., Bybee 2003; Pierrehumbert 2001).

In this study, we seek to evaluate the status of vowels as units of speech production that are independent from tone. We expect to find, if vowels are independent from tones, consistent vowel targets across different tones modulo any effects of tone-vowel coarticulation. On the other hand, if units of speech production are larger, more holistic complexes, such as words or ‘finals’ we expect each tone-vowel combination may have unique spatial targets.

To address the question of tone-vowel independence, we conducted an EMA study of natural variation in tongue displacement across tones. Previous data suggest that tongue position varies to some degree with tone height (Erickson et al. 2004; Hoole & Hu 2004; Hu 2004). These studies report EMA data from one or two speakers with a limited number of contexts, tones and repetitions. More data is needed to evaluate the stability of vowel targets across tones.

2. Method

Six native speakers of Mandarin Chinese (3 male) participated. Each speaker produced multiple repetitions of three maximally-dispersed vowels (i̯/i/–/u/) in labial-initial syllables (/pV/) with each of the four Mandarin tones: 1 ‘high’, 2 ‘low-high’, 3 ‘low’, and 4 ‘high-low’. Each syllable was produced 12 times by each speaker, generating a corpus of 864 tokens (12 reps x 3 vowels x 4 tones x 6 speakers). Syllables were presented in Pinyin and randomized with fillers.

We used an NDI Wave electromagnetic articulograph system sampling at 100Hz to capture articulatory movement. The NDI wave supports 5D sensors and 6D sensors, which can be used for automatic head correction by a proprietary algorithm. We used only 5D sensors for this experiment, attached to the tongue tip (TT), blade (TB), dorsum (TD), lips, jaw, nasion and mastoids. Acoustic data were recorded simultaneously with a shotgun microphone. Head movements were corrected computationally after data collection with reference to the mastoid and nasion sensors. The post-processed data was rotated so that the origin of the spatial coordinates corresponds to the bite plane (front teeth). We analyzed the spatial location of the lingual and jaw sensors at the vowel target, using the findgest algorithm in MVIEWS (Tiede, 2010). Vowel targets were determined by a 20% threshold of peak velocity of the TD sensor in the opening movement of the vowel. All results below are based on measurements taken at this landmark.

3. Results

We report the results in three stages. First we report f0 across the tones and vowels. These results indicate consistent tone patterns across vowels. We then analyze tone-conditioned spatial variation found at the vowel target landmark, focusing on the mid-sagittal plane. After considering the vertical and longitudinal dimensions separately, we introduce a second order measure that integrates displacement in these dimensions relative to the jaw sensor.

3.1. F0 contours

Figure 1 shows the average f0 contour for each vowel across tones. F0 was sampled at regular fixed intervals based on 10 percent of total vowel duration for each tone-vowel combination. The raw f0 samples were converted to z-scores within speaker, an effective normalization procedure for tone (Rose 1987), before averaging. Figure 1 demonstrates that our speakers produced f0 trajectories across vowels that are highly consistent with previous findings for Mandarin Chinese (e.g., Howie 1976), including a small effect of intrinsic f0 on static
high (tone 1) and low (tone 3) tones (c.f., Shi and Zhang, 1987).

Figure 1: Average normalized f0, y-axis, plotted by time expressed as a percentages of total vowel duration, x-axis, for each combination of tone (T1-T4) and vowel (/i/-/a/-/u/) in the corpus.

3.2. Tongue and jaw sensors across tones

Figure 2 provides a summary of tongue and jaw position within the mid-sagittal plane across tones and vowels. Each point represents the average spatial position across speakers at the vowel target landmark. Tongue edges are represented as quadratic fits between Tongue Tip (TT), Tongue Blade (TB) and Tongue Dorsum (TD) sensors (within vowel and within tone). Similar tongue and jaw position across tone can be seen for /u/, small differences for /i/ and larger differences for /a/.

Before averaging values across speakers, sensor positions were normalized by z-score. Figure 2 shows average z-scores projected back onto mm units using the group mean and standard deviation for each sensor. This provides a visualization of effect size in mm units. All statistical analyses were conducted on normalized values. Separate repeated measures ANOVAs were conducted on z-scores for each sensor and vowel in the vertical and longitudinal (anterior-posterior) dimensions. Significant effects of tone on vowel position were found at the vertical sensor for both /i/ [F(3,15)=5.95, p < .01] and /a/ [F(3,15)=11.55, p < .001], at the TT sensor for /a/ [F(3,15)=6.87, p < .01], and at the Jaw sensor for /a/ [F(3,15)=4.58, p < .05]. The only significant effect of tone on longitudinal position was found for /a/ at the TT sensor [F(3,15)=4.67, p < .05]. These effects are highlighted with rectangular boxes in Figure 2. There were no significant effects of tone on /u/ in either dimension.

Effects of tone on vowel target position, summarized in Figure 2, were observed in two of the three vowel contexts examined: tones that start low (2 and 3) pattern together in their influence on tongue position. The results for /a/ production are consistent with the findings of Erikson et al. (2004), who observed that the tongue body is lower (and F1 higher) for /a/ produced with a low tone. The broader constellation of effects found for /a/ with tones that start low (lower jaw, lower TB, lower and more posterior TT) points to a common physiological explanation. Reduction of vocal fold tension for low tones can be achieved by lowering the larynx (Honda et al. 1999; Moisik et al. 2014), which could pull the jaw down (Honda 1995). Jaw movement in the opening phase of vowels involves both a rotational component, whereby the jaw rotates around a terminal hinge, the temporomandibial joint, and vertical and horizontal translations of that axis (Edwards and Harris, 1990). The pattern of effects for /a/ is as expected if the effect of tone on vowel position is mediated by the rotational component of jaw movement. Since the jaw lowers in an arc-like motion, greater lingual displacement as a function of jaw movement is expected for sensors distal to the temporomandibial joint. On this account, the relative stability of the TD sensor follows from its posterior position. Given a degree difference in jaw rotation, the magnitude of the effect on sensor displacement is proportional to the distance from the terminal hinge. Thus, the more anterior lingual sensors, TT and TB, show larger effects of tone, as expected if driven by rotational jaw movement. Moreover, for the most anterior lingual sensor, TT, lowering goes hand in hand with retraction. This also is expected if the arc-like motion of the jaw is driving the effect of tone on tongue position.

Accounting for the effect of tone on /i/ production is less straightforward. The only significant difference was found at the TB sensor in the vertical dimension. As with /a/, tones that start low, tone 2 and tone 3, pattern together. However, unlike for /a/, tones 2 and 3 influence /i/ in the opposite direction. For /i/, the TB was higher for tone 2 and tone 3 than for tones that start high, tone 1 and tone 4. Figure 3 zooms in on these differences comparing the effect for /i/ (right) with that found for /a/ (left). Point estimates are mean values of vertical displacement. Error bars represent 95% confidence intervals. Tones 2 and 3 pattern together but they influence /a/ and /i/ in different directions.

The physiological explanation we offered for the effect of tone 2 and 3 on lingual position for /a/ does not generalize straightforwardly to /i/. However, it may be the case that TB is raised for /i/ with low tones to keep the acoustics of /i/ stable across the four tones by countering mechanistic factors. In other words, the same pull of low tones on the jaw may receive lingual compensation for /i/ but not /a/. This may seem odd from the standpoint of articulatory-acoustic dynamics, where the formant values of /i/ are relatively robust to articulatory variation (Stevens, 1989), but the compensation account is reasonable given the vowel space of Mandarin Chinese. Mandarin has only one low monophthong, /a/, but is comparatively crowded in the upper regions of the vowel space, containing /i/, /i/, and /a/ monophthongs (Duanmu 2007). Lingual compensation for tone-induced pull on the jaw for /i/ and /a/ may be driven by increased articulatory precision required to maintain contrast between non-low vowels.

396
Alternatively, it is possible that low tone production with /i/ involves a different laryngeal mechanism than low tone production with /a/. Recent work on Chinese tones has established two mechanisms of laryngeal articulation engaged in low tone production, one which involves larynx lowering and one which involves larynx raising together with laryngeal constriction (Moisik et al., 2014). Of particular interest is that the stimuli in this study included only words containing the vowel /i/. It is possible that preferences for different mechanisms of low tone production vary with vocalic context, and that larynx raising (for low tones) is more likely in /i/ than in /a/. A complication in applying this mechanistic account to our data involves the fact that tones 2 and 3, the tones that start low, pattern together in their influence on TB height whereas Moisik et al. (2014) observe larynx raising only for tone 3.

3.3. TB to jaw distance across tones

To further investigate the relationship between jaw and TB movement, we computed the Euclidean distance between the TB sensor, where opposite effects of tone were found for /a/ and /i/, and the jaw sensor. Unlike the analysis reported above, this measure takes into account changes across tones in both vertical and longitudinal dimensions. Figure 4 summarizes the measurements across speakers. As expected, the TB sensor is farthest away from the jaw sensor for the /a/ target (42.5mm), followed by the /a/ target (40.9mm), and then the /i/ target (38.9mm), which is closest to the jaw. However, in contrast to the analyses of vertical TB displacement reported above, there is a negligible influence of tone on TB-to-jaw distance. A repeated measure ANOVA on TB-to-jaw distance with tone and vowel as independent factors showed a marginal effect of tone [F(3,15)=3.83, p = .058] but not tone [F(3,15) < 1] and no interaction between tone and vowel [F(3,15) < 1].

The null effect of tone on TB-to-jaw distance for /a/ can be expected, since neither the TB sensor nor the jaw sensor was individually influenced by tone. For /a/, we have already seen that both the jaw and the TB sensor were influenced by tone in the vertical dimension. We now see that these parallel movements maintain a fixed distance between TB and jaw sensors. This indicates that the magnitude of jaw sensor displacement is comparable to the magnitude of TB sensor displacement. The result reinforces our view expressed above in the discussion of /a/ that the effect of tone on vowel targets is mediated by jaw movement. The stable TB-to-Jaw distance for /i/ indicates that, here also, the significant effect of tone on vowel height can be attributed to the jaw. This was not apparent from analyses in 3.2 in part because the contribution of the jaw to TB position is divided across vertical and longitudinal dimensions. The TB-to-jaw distance incorporates these into a single measure, bringing out an invariance masked by single dimensional analyses.

4. Discussion

This study was designed to test whether vowels in Mandarin Chinese have production goals that are independent of tone, as is assumed by modern phonological accounts of Chinese and by most models of speech production, or, alternatively, whether vowels and tones form more integrated composite targets, as in the finals of classical Chinese phonology. There are known physiological linkages between vowel height and vocal fold tension which are mediated by extrinsic muscles acting on the larynx and hyoid bone (Honda 1995). These coarticulatory forces must be taken into account in considering the nature of vowel targets. If vowels and tones have independent production goals then we expect the influence of tone on vowel to fall out from these independent specifications. On the other hand, arbitrary variation, i.e., variation in vowel position that cannot be attributed to the coarticulatory influence of tone, provides support for more holistic speech targets, i.e., tone-vowel inter-dependence.

4.1. The case for tone-vowel inter-dependence

Significant effects of tone on lingual position measured at the vowel target were observed for two of the three vowels examined. This result may appear at first blush to support the hypothesis that tones and vowels are inter-dependent, as in the ‘finals’ of traditional Chinese phonology, more holistic accounts of lexical representation, such as exemplar theory (Pierrehumbert 2002), or other theories that advocate speech production units larger than the vowel, e.g., Fujimura’s (1986) icebergs. We offered a partial physiological explanation for why /a/ is lowered for tones that start low, tone 2 and 3. The effect of these tones on /i/ was in the opposite direction. For /i/, TB was higher for tones 2 and 3. We speculated on possible mechanistic (Moisik et al., 2014) and functional accounts for this pattern. However, with respect to the question of tone-vowel independence, maintenance of small but systematic differences in vowel target as a function of tone supports an integrated representational hypothesis. In the absence of a model that can account for why low tones lead to higher TB for /i/ and lower TB for /a/, the pattern appears arbitrary and, as such, supports the inter-dependence view of tone-vowel relations. TB height may vary across /i1/, /i2/, /i3/, and /i4/ in our data because each of these ‘finals’ are independent units of speech production or because /pi1/, /pi2/, /pi3/, and /pi4/ are all different words of Chinese.
4.2. The case for tone-vowel independence

In contrast to the case for tone-vowel inter-dependence exposed above, we believe that the data also can be viewed as unequivocally supporting tone-vowel independence. However, this interpretation of the data requires that we either focus on certain areas of the tongue, e.g., the stable TD sensor, or that we pursue an alternative expression of vowel targets.

Vowel targets are typically considered to be positions in space, corresponding, for example, to the static images of x-rays (e.g., Stevens and House, 1955). Although details of specific models vary, we take the standard view of vowel targets to involve the position of the tongue relative to the palate. This can be expressed in terms of constriction location and degree, as in Task Dynamics (Saltzman and Munhall, 1989) or as a fixed target in space with quantifiable dimensions (Guenther, 1995). Although there are important differences between these models of vowel targets, they have in common that the production goal is not expressed in terms of the relation between active articulators. Rather, the production goals are expressed independently of the coordinative structures that may achieve them. On this view, to see the data as supporting tone-vowel independence requires focusing on a specific portion of the tongue. Only the TD sensor remains stable across tones. On the view that the entire surface of the tongue contributes to the vowel target, our data instead indicates that vowel targets in Chinese vary with tones in ways that are not fully predictable from physiological constraints on coarticulation, at least not according to our current understanding.

Besides restricting our definition of vowel target to the TD sensor, there is an alternative expression of vowel targets that permits an unequivocal interpretation of the data in terms of tone-vowel independence. This alternative is that it is the relation between articulators that serves to dictate production goals for these vowels. Seen through the lens of a relative notion of vowel target, our data provide strong support for tone-vowel independence. The relationship between tongue position and jaw position remained stable across tones, even as sensor position varied, e.g., at the TB sensor for /a/ and /a/. As a consequence, the Euclidean distance between the TB sensor and the jaw differed for phonologically distinct vowels, /a/, /i/, and /u/, but remained constant across tones. It is therefore possible to characterize the three Chinese vowels in this study in terms of a single dimension, the relationship between lingual and jaw position. On this view, spatial variation in lingual position need not disturb achievement of vowel targets, as long as the jaw is free to co-vary in accordance with a vowel-specific tongue-jaw relation.

Of the two perspectives on the data that permit interpretations in terms of tone-vowel independence, it is not yet clear which (TB-to-jaw distance or TD position) is more important for vowel targets. We leave this question to future study.

5. Acknowledgements

We would like to thank our six Mandarin participants as well as Chong Han, Jia Ying, and Yuan Ma for help recruiting them. We are also grateful for preliminary discussions of the data with Allard Jongman, Joan Sereno, San Duanmu, Cathi Best and Denis Burnham. Research was funded by a grant from the MARCS Institute ‘Time course of tone perception and production’ to Jason A. Shaw, Michael Tyler, Michael Proctor, Donald Derrick and Chong Han.

6. References


Howie M (1976), Acoustical studies of Mandarin vowels and tones (No. 6). Cambridge University Press.


RT-MRI Based Dynamic Analysis of Vocal Tract Configurations: Preliminary Work Regarding Intra- and Inter-sound Variability

Samuel Silva¹, António Teixeira¹²

¹JEETA, University of Aveiro
²Dep. of Electronics, Telecommunications and Informatics, University of Aveiro

Abstract

The study of the dynamic aspects of speech production is relevant for different classes of sounds. To improve knowledge regarding this subject it is important to consider data from various repeated utterances, from different speakers, leading to the challenging scenario of processing and analysing a large amount of data in a common framework.

We propose a method for dynamic analysis of articulatory data jointly exploring multiple utterances of each sound using a quantitative analysis framework and covering multiple articulators. Instead of analysing each utterance and then inferring notable features, all utterances for a particular sound are considered simultaneously, possibly from more than one speaker, aiming towards a characterization of average dynamic aspects and their corresponding variation. Application examples are provided involving the analysis of vowel dynamics and inter-vowel comparison of European Portuguese nasal vowels data gathered using real-time Magnetic Resonance (RT-MRI).

Keywords: dynamic articulatory analysis, nasal vowel dynamics, RT-MRI

1. Introduction

The articulatory characterization of the dynamic aspects of different sounds is of the utmost importance. One notable example where these aspects are relevant, previously addressed by the authors and colleagues, is the production of nasal vowels (Oliveira et al. 2012). Furthermore, for articulatory synthesis it is important to characterize the dynamic production aspects considering the various articulators. In both cases, the complexity of speech production and its variability across speakers might make it difficult to perceive which targets and articulator movements are relevant for producing a specific sound.

The study of the dynamic aspects of speech production has traditionally been performed by studying the trajectories of notable landmarks (e.g. Shosted, Sutton, and Benmamoun 2012) or the occurrence and location of constriction points in the vocal tract (e.g. Hagedorn, Proctor, and Goldstein 2011). These studies provide valuable knowledge about speech production, but are not adequate to deal simultaneously with data from multiple speakers. Therefore, in the increasingly more common scenario of large sets of dynamic vocal tract data (e.g. resulting from US and RT-MRI studies (Teixeira et al. 2012)), it is important to move to an analysis framework that allows easy integration of data from different articulators, repeated utterances and multiple speakers.

In (Silva et al. 2013) the authors proposed a novel quantitative analysis framework that allows regional comparison of vocal tract profiles, encompassing the different regions under influence of the various articulators. Recently, in (Silva and Teixeira 2014), this quantitative framework has been expanded to support analysis comparing articulatory data between speakers or aggregating data from multiple speakers to provide overall characteristics of different sounds. These methods address the analysis of articulatory differences considering static representative configurations of the vocal tract for the different sounds.

A first proposal of a dynamic analysis framework was presented (Silva et al. 2013), supported by and extending quantitative vocal tract comparison methods and covering the analysis of multiple articulators. As remarked at that time, it was desirable that such dynamic analysis could be used to assess dynamic aspects of different sounds considering not only individual utterances, but all relevant utterances available. Following that goal, we propose here a dynamic analysis jointly exploring multiple utterances of each sound, produced by one or more speakers, using a quantitative analysis framework. Instead of analysing each utterance and then inferring notable features, all utterances are considered simultaneously, aiming towards a characterization of average dynamic aspects and their corresponding variability. Application examples are provided involving the analysis of intra-sound and inter-sound variability of European Portuguese (EP) oral and nasal vowels.

This article is organized as follows: Section 2 describes the main aspects of the proposed methods; Section 3 shows applications examples to illustrate their use; and Section 4 presents conclusions and ideas for future work.

2. Methods

In what follows the vocal tract comparison methods are detailed and the proposed dynamic analysis representation presented.

2.1. Vocal Tract Comparison

To provide context to the dynamic analysis, we briefly present the main features of the framework used to compare between vocal tract profiles (Silva et al. 2013; Silva and Teixeira 2014). Two important aspects are covered: the quantitative comparison method and the data it provides and the proposed visual representation.

2.1.1. Comparison Considering Multiple Articulators

Vocal tract comparison is performed comparing contour segments (e.g. tongue dorsum and velum) or comparing extracted distances between articulators (e.g., lip aperture). A comparison between two vocal tract profiles (Figure 1) results in seven normalized difference values (i.e., between zero and one, where...
one means no difference) for the tongue dorsum (TD), velum (VEL), tongue back (TB), tongue tip (TT), lip protrusion (LP), lip aperture (LA) and pharynx/larynx (Ph). While TD, VEL, TB and Ph segments are compared using the Pratt index (Pratt 2007), for the remaining articulators, distances are obtained between notable VT points and normalized.

For example, for the tongue tip (TT), position is compared by computing the distance between the tongue tips in both contours normalized by the longest distance from each tongue tip to the alveolar ridge (AR). 

$$TT = 1.0 - \frac{d_{TT}}{\max(d_A(TT \rightarrow AR), d_B(TT \rightarrow AR))}.$$ 

For a detailed description the reader is forwarded to (Silva et al. 2013).

2.1.2. Vocal Tract Comparison Data Representation

Representation of the data from one comparison between two vocal tract profiles if performed over the unitary circle in a radar-like representation. Coloured circular coronas for negligible difference (green, $> 0.75$), mild difference (yellow, $[0.5, 0.75]$) and strong difference (red, $< 0.5$) are also shown. All difference points are connected by line segments in an attempt to improve visualization (Silva et al. 2013). At the centre, an orange polygon depicts the standard deviation for each difference value.

Another important information when characterizing the difference between vocal tract profiles is how the difference occurred. For example, if there is a difference at the velum, between two vocal tract profiles, it is important to know in which direction did the velum move from the first to the second. To provide this information, for mild and strong differences, an arrow is presented with its origin at the parameter point, and oriented according to articulator movement. Figure 2 shows an example of a traditional vocal tract profile superposition for [a] and [i] and its representation using the proposed quantitative framework.

2.2. Analysis of Dynamic Aspects of the Vocal Tract

In what follows the main features of the proposed dynamic analysis methods are provided with an emphasis on matching multiple utterances, the different kinds of analysis possible and the adopted representation of the computed data.

2.2.1. Matching Multiple Utterances

Since multiple utterances of the same vowel presented different durations (annotated using Praat (Boersma and Weenink 2013)), before performing any kind of analysis taking advantage of multiple utterances to characterize the mean dynamic behaviour and dynamic intra-vowel variability, they had to be properly matched. Two methods have been tested: 1) time axis normalization, with each vowel utterance duration assumed to vary between 0 and 1; 2) Dynamic Time Warping (DTW) (Berndt and Clifford 1994) was applied to each pair of audio signals (for the same sound). Since no major difference has been detected between the end results for both methods, for the sake of simplicity time axis normalization was adopted.

Based on the speech signal matching, the vocal tract contours extracted from the RT-MRI image frames corresponding to each utterance are interpolated in order to allow comparison of corresponding contours between the utterances.

2.2.2. Dynamic Analysis of Vocal Tract

Considering the available data, three different kinds of dynamic analysis can be performed covering the different articulators:

- Intra-sound dynamics, i.e., how does the vocal tract vary from the beginning to the end of the production and how consistent it is across productions. For example, cross compare all utterances of [i].
- Inter-sound comparison, i.e., compare the production of two different sounds along time to check differences/similarities. For example, vowel [o] and its nasal congener [ø].
- Comparison, over time, with a reference, i.e., analyse production of a sound using a particular (static) vocal tract configuration, as reference, to check how similar (and when) is the vocal tract configuration along the production to other relevant configurations (e.g., [e] compared to representative (static) configuration of [r]).

2.2.3. Dynamic Data Representation

The representation of the data gathered for the dynamic analysis has four different parts (refer to figure 3, for an example),
3. Application Examples

After a brief description of the image data considered, a few application examples covering EP nasal vowels are shown to highlight the main features provided by the proposed dynamic analysis methods.

3.1. Image Acquisition and Contour Extraction

Image sequences were acquired using RT-MRI, including the five EP nasal vowels and the eight oral vowels for three speakers (Teixeira et al. 2012). Synchronized audio was recorded during image acquisition and annotated manually using Praat (Boersma and Weenink 2013). Vocal tract segmentation has been performed semi-automatically for all image sequences. Figure 4 shows some examples of RT-MRI image frames and corresponding vocal tract segmentations.

3.2. Intra-sound Dynamics

All utterances of a particular sound – in our case, nasal vowels – are matched and cross compared. The resulting comparison data for each considered frame is averaged and represented in the line plot to depict evolution over time. Figure 3 shows, as representative examples, variation data for vowels [˜ 5] and [˜ u]. Since all utterances are matched and the vocal tract comparison method provides normalized data, the presented dynamic analysis for vowel [˜ u] was performed taking into account the data for all three speakers in the RT-MRI dataset. Visualizations made possible by the proposed method show no major discrepancy of vocal tract configuration across utterances, for both vowels, is observed, i.e., all utterances exhibit a similar vocal tract variation pattern: a slight difference is observed for lip aperture (LA), in vowel [˜ 5], at the beginning and end of the production, but the remaining parameters keep above 0.75 with low standard deviations.

3.3. Inter-vowel production differences

Figure 5 shows two examples regarding the assessment of dynamic differences between sounds. In the first example (top row), the production of [˜e] was compared with the production of [˜e]: both start very similar with only a slight difference at tongue tip and gradually differ at lip aperture and velum. This gradual variation at the velum is in agreement with previous studies (Oliveira et al. 2012; Martins et al. 2012).
The second example (figure 5, bottom row) shows the comparison between vowels [ı] and [u]. As expected there are strong differences at the tongue tip (TT), tongue back (TB) and tongue dorsum (TD). Nevertheless, note how no substantial difference exists at the velum (VEL) between the two sounds. Since both are nasal vowels, the velum is expected to gradually open across the production (Martins et al. 2012). The dynamic analysis shows that both sounds exhibit a similar velum aperture pattern.

4. Conclusions

The results provided by the proposed method seem promising and provide grounds to quantitatively characterize intra- and inter-sound variability along the production of different sounds. In the particular application shown, as an example, jointly using data for three speakers, the studied EP vowels did not present notable inter-utterance variability and the proposed methods allowed a depiction of how and when different sounds differ, over time. The quantitative nature of the data is also important for further exploration and analysis using computational tools.

As future work the application of the proposed methods to other sounds (e.g., laterals) and the study (and inclusion in the proposed visualizations) of landmark trajectories and their variability, considering all available speakers, should provide further insight into the usefulness of the proposed methods.

These methods are computationally demanding and generate large amounts of data. Therefore, their deployment in a cloud computing scenario would provide a more suitable scenario for further developments and an important first step towards their validation and use by third parties such as Phoneticians.

Acknowledgements: Research partially funded by FEDER through the Program COMPETE and by National Funds (FCT) in the context of HERON II (PTDC/EEA-PLP/098296/2008), IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-FEDER/IEET/010127/2011), Project Marie Curie IAPP “IRIS” (FP7-PEOPLE-2013-IAPP) and project Cloud Thinking (funded by the QREN Mais Centro program, ref. CENTRO-07-ST24-FEDER-002031).

5. References


Systematic and Quantitative Analysis of Vocal Tract Data: Intra- and Inter-Speaker Analysis

Samuel Silva¹, António Teixeira¹²

¹JEETA, University of Aveiro
²Dep. of Electronics, Telecommunications and Informatics, University of Aveiro

Abstract

Articulatory analysis has traditionally been performed visually, by comparing, in general, a reduced number of vocal tract profiles. Nowadays, technological advances in speech production analysis methods, for which imaging techniques such as real-time Magnetic Resonance Imaging (RT-MRI) is a good example, allow the collection of large amounts of data providing resources to advance in numerous application areas such as articulatory studies and articulatory synthesis. In this context visual analysis is insufficient and mostly infeasible. Furthermore, the subjective nature of such assessment makes comparisons between speakers and languages almost impossible, prone to non-uniform criteria used by different researchers. It is, therefore, of paramount importance to move to a quantitative articulatory analysis framework.

The authors have recently proposed a method to perform quantitative comparison between vocal tract profiles and an abstract representation of the gathered data for easier interpretation (Silva et al. 2013b). That first proposal was only applied to compare intra-speaker data and therefore required further developments to serve comparisons among speakers and analysis gathering data for all speakers. Extending this previous work, the work carried out so far provides the following additional contributions: a) quantitative assessment of per vowel intra-speaker variability, b) inter-speaker differences and c) overall inter-vowel differences, using data from all speakers in a single representation.

To illustrate, on a real scenario, the potential and different applications made available by the proposed methods, the comparison among EP (European Portuguese) vowels is presented.

To the best of our knowledge this is the first time (even considering other languages) a quantitative articulatory analysis is presented encompassing the assessment of per vowel intra-speaker variability, inter-speaker differences and overall inter-vowel differences using data from all speakers in a single representation.

This article is structured as follows: Section 2 describes the main features of the proposed framework; Section 3 provides some examples depicting the possible applications of the proposed framework; finally, Section 4 presents some conclusions and ideas for further developments of the work carried out.

2. Vocal Tract Comparison Framework

In what follows we present the main aspects of the proposed framework regarding vocal tract comparison and visualization of the compute data in a difference diagram.

2.1. Vocal Tract Comparison

For comparison between vocal tract profiles, the middle frame for each oral vowel occurrence and last frame for nasal vowels (Oliveira et al. 2012) (based on the audio annotation) is selected. Vocal tract contours are divided in segments considering different anatomical regions and articulators: lower lip, tongue tip, tongue dorsum, tongue back, larynx/pharynx, velum, hard palate and upper lip. Vocal tract comparison is performed comparing contour segments (e.g., tongue dorsum and velum) or comparing extracted distances between articulators (e.g., lip aperture and protrusion).

The pharynx/larynx (Ph), tongue back (TB), tongue dorsum (TD) and velum (VEL) are compared by computing the Pratt index Pratt 2007 for each pair of corresponding contours and given by:

\[ P = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + \alpha d_i^2}, \]

where \( N \) is the number of corresponding points between the compared contours (e.g., tongue back), \( d_i \) is the euclidean distance between two corresponding points, and \( \alpha \) is a constant set to 1/9, based on Pratt’s work and similar works in the literature. At this stage the same constant value has been used for all regions but it might be tuned for each region if different sensitivities to differences were desired. To obtain corresponding points between contours, the contour with the smallest number of points is selected and for each point in this contour the closest point in the other contour is considered the corresponding point.

The tongue tip (TT) position is compared by computing the distance between the tongue tips in both contours normalized by the longest distance from each tongue tip to the alveolar ridge.
Lip protrusion (LP), i.e. horizontal lip movement, is obtained by computing the horizontal displacement of the mid-point between the upper and lower lips (LMP). The mid-point is computed considering the line that connects the lowest point of the upper lip with the highest point of the lower lip. To perform normalization the horizontal distance between the mid-points and the alveolar ridge (AR) is obtained \((d_A(TT-AR), dB(TT-AR))\) and used as follows:

\[
LP = 1.0 - \frac{\max(d_A(LMP-AR), dB(LMP-AR))}{d_A(LMP-AR) + dB(LMP-AR)}
\]

Lip aperture (LA) is computed based on the lip aperture values for both vocal tract profiles \((L_A, L_B)\) and is normalized by considering the longest of the two:

\[
LA = 1.0 - \frac{|L_B - L_A|}{d_A(LMP-AR), dB(LMP-AR)}
\]

The resulting seven values are a quantitative representation of the differences found between vocal tract profiles yielding several important features, such as an objective, repetitive measure of difference and suitability for computational analysis, paving the way for methods such as cluster analysis.

### 2.2. Comparison Data Visualization

The quantitative comparison method presented above, resulting in a set of numerical values, demands a visualization method that can help users to interpret/compute and disseminate in an easy way as opposed, for example, to value tables. One possible representation would be to present the values in a line plot, with the parameters shown in a fixed order, in the X-axis. Nevertheless, we considered that a different representation might bring easier interpretation if, somehow, one could build some analogy with the vocal tract and the position of the different articulators.

Figure 1 illustrates the proposed representation (Silva et al. 2013b). The computed comparison values are represented over the unitary circle in a radar-like representation. Its design follows the principles described below.

**Parameter Location** — The location of each parameter in the representation is chosen as if the vocal tract was inscribed in the unitary circle and selecting, at each time, the parameter concerning the closest articulator. Furthermore, the orientation associated with each articulator/parameter is, where applicable, related with its movement direction. Starting at zero degrees, for the tongue back (TB), it follows the velum, tongue dorsum, tongue tip, lip protrusion and lip aperture.

**Difference Level** — To help on the interpretation of the presented representations, three circular coronas are proposed. The first, between 0.75 and 1.0, in green, is suggested to mean no relevant difference. The interval lower bound is chosen based on the value typically interpreted as meaning a good match between contours compared using the Pratt index. The second circular corona, between 0.50 and 0.75 is proposed to be interpreted as a mild difference and the last, from 0.0 to 0.50, as a strong difference.

These are proposed interpretations and different criteria can be adopted, but their intent is to help users to have common interpretation criteria and to move away from direct numerical comparison which would lead, yet again, to a rather subjective assessment. Furthermore, the intent is also not to have too much difference levels.

**Displacement Direction** — One important aspect regarding the differences found when comparing vocal tract profiles is to characterize how that difference occurred from one configuration to the other, e.g., by a rise in the tongue dorsum or lowering of the velum. To provide such information, a vector is used, with origin in the parameter value and the direction of displacement. For the sake of simplicity displacement direction is only represented for values corresponding to mild or strong differences.

**Variability** — Since differences are often computed for multiple occurrences of two sounds, and the average difference represented, it is also important to depict the variability associated with each value. The standard deviation for each parameter is represented over its direction and all parameter points used to draw an orange polygon.

**Overall Interpretation** — As a first approach, trying to provide easier analysis of the proposed representation, namely a faster perception of each point location, and a simple way of “grouping” the points belonging to the same comparison, when multiple comparison data is shown, the points are connected using line segments, forming a polygon.

### 3. Application Examples

In this section we show some examples of what can be accomplished using the current state of the proposed framework regarding different tasks relevant for articulatory analysis performed over an existing RT-MRI corpus.

The main purpose is not to perform a detailed characterization of the articulatory characteristics of EP, but to highlight those innovative aspects provided by the proposed quantitative analysis framework. Examples focus on the EP vowels [a][i][u] and their nasal congers.
3.1. Image Acquisition and Contour Extraction

Using RT-MRI (Teixeira et al. 2012), the corpus, acquired for three speakers (CM, CO and SV), includes:

- the five EP nasal vowels ([i], [e], [i], [6], [i]) in word initial, medial and final positions (e.g. [6p]), [p6p], [p6]);
- the eight oral vowels ([a], [r], [r], [e], [i], [s], [o] and [u]) inserted in CV, CV sequences (e.g. [papa] [pap], [papa] [pup]), where C is a voiceless bilabial plosive. Synchronized audio was recorded during image acquisition and annotated manually using Praat Boersma and Weenink accessed Mar. 2014.

Vocal tract segmentation has been performed semi-automatically for all image sequences as described in Silva et al. 2013a.

3.2. Intra-Speaker Analysis

All occurrences of each vowel can be cross compared. This comparison allows assessment of the variability associated with the production of each vowel, for each speaker. Figure 2 presents some example difference diagrams. In general, for the studied vowels, no major differences were found with some exceptions for SV at the tongue tip or lip aperture for vowel [a].

Quantitative analysis can also be performed to compare among different vocal tract configurations for the same speaker. As an example, figure 3 shows comparisons among different vowel pairs. In EP, such comparisons are useful, for example, to study oral vs nasal vowels. Given the normalized nature of the computed differences, they can be presented using a common representation. Some contour superpositions are also presented to show how the representation relates with the traditional visual assessment.

3.3. Inter-Speaker Comparison

One important feature provided by the proposed framework is the comparison between speakers. Since the provided comparison values are normalized, it is possible to visualize the differences for various speakers on a common representation. The difference polygons for each speaker and for each vowel pair were superimposed to assess inter-speaker differences (see figure 4). In general, the difference polygons for all speakers, for the same vowel pair, were similar. Noticeable inter-speaker differences (i.e., were sometimes observed for lip aperture.

3.4. Overall Characterization of Articulatory Differences

The data gathered for all speakers can now be used to assess the overall articulatory differences between vowels. Observing Fig-
Figure 4: Inter-speaker comparison of differences between vowels. From left to right: [a] vs [i], [a] vs [u] and [i] vs [u].

Figure 5: Differences among vowel pairs computed using data from all speakers. From left to right: [a][i][u], [ɪ][ʊ][a] and [ʊ][u].

A notable aspect, evidenced by the standard deviation, is that, overall, the strongest variability among the different productions is mostly observed at the tongue tip and lip aperture (protruding vertices of the orange polygon at the centre).

4. Discussion and Conclusions

This article proposes a quantitative framework for vocal tract profile comparison. To the best of our knowledge, this is the first time a quantitative analysis framework is presented encompassing the assessment of per vowel intra-speaker variability, inter-speaker differences and overall inter-vowel differences using data from all speakers in a single representation. This shows the potential of the proposed framework to take advantage of the increasing amounts of articulatory data available and move towards quantitative inter-speaker and inter-language comparison.

This is an ongoing work that is open to improvements and we have been gathering feedback from different researchers concerning the proposed representations. A first informal evaluation was presented in Silva and A. Teixeira 2014 and we intend to further explore their inputs. For example, some researchers considered that connecting the points corresponding to lip aperture and the pharynx was misleading, since it might be understood as an anatomical connection. We argue that the main intention is just to improve readability of the representation, but have to gather more feedback regarding this aspect.

As future work, application of the proposed framework to a complete study including all EP vowels is envisaged, exploring several factors such as word position and context.

The proposed method, together with suitable vocal tract segmentation methods, allows to deal with large amounts of image data, generating data that can be used for a computational approach to articulatory studies. This processing pipeline is computationally demanding and generates large amounts of data and would, therefore, profit from its implementation in a cloud scenario. Its implementation in such a scenario is also important to allow other research groups to use the framework over their data, paving the way for comparisons between dialects and languages. We are currently starting this migration process in the scope of projects Cloud Thinking and IRIS.

Acknowledgements:
Research partially funded by FEDER through the Program COMPETE and by National Funds (FCT) in the context of HERON II (PTDC/EEA-PLP/098298/2008), IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEst-C/EEI/UI0127/2011), Project Marie Curie IAPP “IRIS” (FP7-PEOPLE-2013-IAPP) and project Cloud Thinking (QREN Mais Centro, ref. CENTRO-07-ST24-FEDER-002031).

5. References


The Role of Auditory Information on Gestural Intrusions and Reductions
Anneke Slis1, Pascal Van Lieshout2

1, 2Department of Speech Language Pathology, Oral Dynamics Lab, 160-500 University Avenue, University of Toronto, Toronto, Ontario, M5G 1V7, Canada
Anneke.slis@mail.utoronto.ca, p.vanlieshout@utoronto.ca

Abstract
The present study reports findings on kinematic data investigating the role of auditory information in intrusions and reductions. Intrusions and reductions arise as a consequence of destabilized coupling of gestures. The study sought to answer the question whether the presence of auditory information stabilized articulatory coordination. The hypothesis was that the presence of auditory information stabilized speech and thus that fewer intrusions and reductions would occur with auditory information present than without. Articulatory movements of the tongue tip, tongue dorsum and lower lip were recorded with Electromagnetic Articulography. Auditory feedback was manipulated by a masking condition. Speech rate and part of the trial (start, middle and end) formed two other independent variables. The findings revealed that more intrusions were made with auditory information available than without. Fast speech resulted in more intrusions than normal speech and more intrusions and reductions were found at the end of a trial. The data suggest that factors such as attention and a larger role for proprioceptive information during masked than unmasked speech stabilized speech coordination.

Keywords: coordination dynamics, entrainment, auditory information, gestural intrusions and reductions

1. Introduction
Listeners employ auditory information when correcting their speech in case of errors (Postma & Noordanus, 1996). Little is known about how auditory information influences the occurrence of speech errors. The present study elaborates on recent kinematic studies that defined errors as gestural intrusions and reductions arising from entraining movement patterns (Goldstein, et al., 2007; Slis & Van Lieshout, submitted). Entrainment is the process in which two independently moving oscillators, i.e. limbs or articulators, tend to synchronize over time. This entrainment of movement is considered an autonomous mechanism in which patterns form spontaneously and adjust in a flexible manner to changing conditions as specified in Coordination Dynamics theory (Beek et al., 1995; Van Lieshout, 2004). Studies showed that the presence of auditory information has a stabilizing role on coordination dynamics for limb control (Lagarde & Kelso, 2006; Namavivayam et al., 2009). It is so far untested whether this information is relevant for preventing or changing the nature of gradual gestural intrusions and reductions in speech.
Traditionally, speech errors have been explained as involving abstract static speech units, such as a phoneme, that have been transposed to a different location (Fromkin, 1971). Recent studies have shown that these so-called phonological errors are more gradual and present themselves as intruding or reducing articulatory movements (Goldstein, et al., 2007; Slis & Van Lieshout, submitted). In the case of intrusions, non-target articulatory movements are intruding during a target constriction. For example, in the word pair “top cop” an extra tongue dorsum intrudes during the /t/ in the word “top”, resulting in two sequential tongue dorsum activations. Reductions are characterized by reduced articulatory movements of a target constriction. In the studies that typically employ repetitive speech, intrusions and reductions were found to be building up over the time course within a single trial, especially at a higher speaking rate. These intrusions have been explained as originating from a stabilizing mechanism in which a 1:1 mode is more stable than a 1:2 ratio (Goldstein, et al., 2007). Higher speaking rates reduce the coupling strength between gestures and consequently tend to destabilize movement coordination (Van Lieshout & Neufeld, 2014) explaining the higher rate of intrusions at a fast speaking rate. Because the presence of auditory information is assumed to strengthen the coupling in speech movement coordination, it is hypothesized that the 1:2 mode of coordination is easier maintained with auditory information present than without. In line with this, lack of auditory information would result in more intrusions and reductions.

2. Methods
2.1. Participants and procedure
Fifteen monolingual speakers of Canadian English between 19 and 35 years of age were asked to produce CVC-CVC word pairs with alternating onset consonants and identical rhymes, such as in the word pair “cop top”, 15-17 times repetitively. The stimuli consisted of the word pairs “cop top”, “kip tip”, “pick tick”, “pock tock”, “pot cot”, and “pit kit”. Two different speaking rates were employed, normal and fast. Speaking rate was guided by a visual metronome. To investigate the role of auditory information on the occurrence of intrusions and reductions during repetitive speech, two conditions were employed: a first condition in which the participant could perceive his own speech and a second condition in which noise prevented the speaker from perceiving his own speech.
For the purpose of this study, the maximum vertical displacements of the target movements of the tongue tip, tongue dorsum, and lower lip during the respective onset consonants /t/, /d/ and /p/ in the first and second word were retrieved from data collected with the EMA AG500 system. The position of a non-target articulator, i.e. the intruding articulator, was measured at the time when the target articulator was maximally constricted.
An intrusion was defined as an outlier from a distribution of normalized movement ranges for non-target articulators. Likewise, a reduction was defined as an outlier from a distribution of normalized movement ranges of target articulators. To determine these outliers, two median values were found.
were calculated: one based on the movement ranges of successive target maxima of a specific articulator within a trial and one based on movement ranges of the non-target amplitude values. Movement range values, two Median Absolute Deviations (MADs) or more below the median target value were considered reductions, and those two or more MADs above the median non-target value were considered intrusions. To evaluate the prediction that the intrusions and reductions build up faster without auditory information available, the trials were divided in three parts: start (repetition 3 to 6), middle (7 to 10) and end (11 to last).

2.2. Statistical analysis

A repeated measures ANOVA was used to determine if (lack of) auditory information affected the ratio of intrusions and reductions. The dependent variables were ratio of intrusions and reductions. The independent variables assessed the influence of speaking rate (normal and fast), part of trial (start, middle and end), and the presence of auditory information (masked or not masked). All tests were performed with \( \alpha < 0.05 \).

3. Results

The results showed that masking speech affected the rate of intrusions (F(1, 13) = 6.67, p = 0.02). More intrusions were measured in unmasked than masked speech. Fast speaking rate resulted in more intrusions than normal speaking rate (F(1, 13) = 34.56, p < 0.0001). As expected, the intrusions were building up over the course of the trial, showing a significant Part effect (F(2, 26) = 11.05, p < 0.001). With respect to reductions, a Part effect was revealed as well (F(2, 26) = 7.81, p < 0.01). Neither noise nor speaking rate affected the ratio of reductions. The interactions were not significant.

4. Discussion and conclusion

Based on these analyses, it can be concluded that speakers made more intrusions when auditory information was available. It is speculated that, when no auditory information is present, speakers attend more closely to their articulatory movements and employ different sensory modalities, such as proprioception. These factors strengthen the coupling to prevent the intrusions to build up towards the end of a trial.

5. Acknowledgements

The study was supported by a Social Sciences and Humanities Research Council (SSHRC) grant and partly, by funding from the Canada Research Chairs program both awarded to Pascal van Lieshout. The authors would like to thank James Le for his technical support during the EMA sessions and Radu Craioveanu for his help analyzing the data.

6. References


Complex Tongue Shaping in Lateral Liquid Production Without Constriction-Based Goals

Caitlin Smith

1Department of Linguistics, University of Southern California, Los Angeles, California, USA
smithcm@usc.edu

Abstract

Liquids are complex consonants that employ multiple gestures of the tongue in order to achieve specific tongue shapes. In the case of the lateral liquid /l/, that specific shape involves lowering and curling of the tongue along the blade. Many models of speech treat /l/ simply as a combination of tongue tip closure and tongue body retraction gestures, with tongue shaping receiving little attention. Here we examine the sequencing of achievement of both the constriction gestures and complex tongue shaping of American English /l/. This will allow us to discern the interaction between constriction formation and tongue shaping, shedding light on the true goals of production for lateral liquids. The results suggest that the so-called primary tongue tip constriction gesture of /l/ is actually a form of tongue bracing that aids in the achievement of tongue curling in tandem with tongue body retraction, the true goals of production of this complex consonant.

Keywords: liquids, laterals, tongue shaping, speech tasks, syllable structure

1. Introduction

Liquids have long brought up questions regarding the representations of their goals of production. This is especially true within the framework of Articulatory Phonology (Browman & Goldstein 1986, 1992), in which the complex tongue shaping associated with liquids is modeled via the coordination of multiple constriction-based gestures. The lateral liquid /l/ is captured by two gestures, one for tongue tip constriction at the alveolar ridge, and one for tongue body constriction at the pharynx. The complex tongue shaping of /l/, characterized in particular by tongue blade lowering or curling in order to create lateral air channels, is considered only indirectly.

The sequencing of the gestures that make up /l/ and lend it its complex shape has been shown to vary by syllable position. In general, the consonantal tongue tip gesture precedes the vocalic tongue body gesture in onset position, but the tongue body gesture precedes the tongue tip gesture in coda position (Sproat & Fujimura 1993; Browman & Goldstein 1995). It remains unclear what role of tongue shaping is in terms of syllable position, and how that tongue shaping transpires over time when the sequencing of the tongue tip and tongue body gestures of /l/ varies.

It is especially unclear how the sequencing of the constriction gestures and tongue shaping of /l/ interact when one of those gestures is not fully realized. It is common for /l/ in coda position to be realized without any discernible contact between the tongue tip and alveolar ridge, a phenomenon known as /l/ vocalization (Hardcastle & Barry 1989). Previous work has shown that even during the production of vocalized /l/, which is seen in all syllable positions but most prevalent in nucleus and coda positions, substantial curling of the tongue blade is achieved (Smith & Lammert 2013). Understanding how all of these events are structured and sequenced within the syllable may shed light on why /l/ vocalization is prevalent in some syllable positions and not others, as well as further informing its effect on tongue blade curling.

This work aims to discern the goals of production of the /l/ of American English by examining the production of its composite gestures across syllable positions in terms of their individual achievement, their relative coordination, and their relationship to tongue blade lowering/curling. This is done through the use of real-time magnetic resonance imaging (MRI), which provides dynamic information about the entire vocal tract, allowing us to examine both the relative timing of multiple gestures and changes in the shape of the tongue throughout the achievement of those gestures.

2. Methods

Real-time MRI is a methodology that is well suited to the examination of lateral liquid production. Electropalatography can only provide information about whether the tongue tip has made contact with the alveolar ridge and tells us nothing about the complex shaping of the tongue body. Ultrasound is able to image the tongue body, but is usually unable to capture the tongue tip. MRI overcomes this problem, providing a full view of the vocal tract from glottis to lips. However, static MRI requires speakers to maintain production of target phones for several seconds, essentially allowing us to see only the equivalent of a syllabic /l/. The use of real-time MRI allows us to embed /l/ in different contexts, particularly different syllable positions, as well as to examine how the production of /l/ unfolds over time in each of these positions.

2.1. Data

This work examines the speech of two subjects in the USC-TIMIT real-time MRI database, compiled by the University of Southern California’s SPAN (Speech Production and Articulation kNowledge) group and publicly available for research purposes (Narayanan et al. 2011). Each speaker in the USC-TIMIT database produced a phonetically balanced set of 460 sentences. Between both subjects examined here, this yielded tokens of /l/ in various syllable positions, as listed in Table 1.

Table 1: Number of /l/ tokens at each syllable position.

<table>
<thead>
<tr>
<th>Syllable position of target /l/</th>
<th>Number of Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onset</td>
<td>54</td>
</tr>
<tr>
<td>Intervocalic/Ambisyllabic</td>
<td>44</td>
</tr>
<tr>
<td>Coda</td>
<td>42</td>
</tr>
<tr>
<td>Nucleus</td>
<td>34</td>
</tr>
</tbody>
</table>

In order to minimize the influence of tongue shaping from nearby liquids on the liquid targeted for analysis, any instances of /l/ with an /ɹ/ in an adjacent syllable were not considered.
2.2. Image acquisition

A team of SPAN members performed the acquisition of real-time MRI images at Los Angeles County Hospital. Each subject lay supine in the scanner with his or her head held in place by a neck brace. A mirror was placed in the scanner allowing subjects to see out of the bottom of the scanner and read stimuli that were projected onto a screen. Stimuli were presented in five-sentence blocks, with a pause of approximately thirty seconds between each block in order to allow the scanner to cool. Subjects produced each sentence once.

Scans used a real-time MRI protocol specifically developed for the examination of speech production (Narayanan, Nayak, Lee, Sethy, & Byrd 2004). Images were acquired along the midsagittal plane with a complete view of the vocal tract: glottis, pharynx, and oral and nasal cavities. They were reconstructed at a rate of 23.18 frames per second with a field of view of 68 x 68 pixels, corresponding to 200 x 200 millimeters.

2.3. Image analysis

Edge tracking for the contours of the tongue and palate was performed using a custom-designed GUI (graphical user interface). The GUI introduces an image scaling factor of five, bringing the spatial resolution of each frame from 68 x 68 pixels to 340 x 340 pixels. A semi-polar grid of approximately thirty lines orthogonal to the vocal tract was laid over the video images. This grid was fit specifically to each speaker’s vocal tract by specifying points at the speaker’s glottis, palate, alveolar ridge, and lips.

![Figure 1: One frame of midsagittal MRI with grid overlaid and upper and lower surfaces of vocal tract outlined](image1)

Portions of the upper vocal tract surface’s contour were fixed at the palate and pharynx; to account for head movement, the grid was shifted as necessary to line up these fixed portions of the contour with the relevant anatomical landmarks. For each frame of video, the air-tissue boundary for the upper and lower surfaces of the vocal tract were automatically detected at each gridline and corrected manually if necessary. Further details on this method of image analysis can be found in Proctor, Bone, Katsamanis, & Narayanan (2010).

2.4. Measurements

This study utilizes two types of measures: those that capture the formation of constriction or tongue shaping that may be seen as goals of production, as well as temporal measures of the achievement of those articulations. First, in order to see whether some syllable positions are more conducive to /l/ vocalization than others, rates of vocalization were compared across syllable positions. Rate of vocalization is calculated as the number of vocalized tokens of /l/ in a given syllable position divided by the total number of tokens of /l/ in that syllable position. A token of /l/ is considered vocalized if there is no contact between the tongue tip and the upper surface of the vocal tract during its production. Only non-coronal adjacent tokens of /l/ were considered for the calculation of rate of vocalization.

Additionally, a quantification of curvature of the tongue contour was performed. First, fifteen evenly spaced points were laid along the tongue contour. The radius of the ellipse passing through each set of three contiguous points was calculated, and its inverse taken and multiplied by 100 to yield a curvature score. If the circle between three points lay outside of the tongue area, the curvature score was recorded as negative. Curling of the tongue during the production of /l/ was signaled by a high-magnitude negative curvature score at a point along the tongue blade.

![Figure 2: Evenly spaced points along tongue contour during production of /l/ in ‘excluded’ (above) and /d/ in ‘adult’ (below). Blue indicates positive curvature; red indicates negative.](image2)

In order to examine the sequencing of events during the production of /l/, several temporal measures were also made. In cases in which /l/ was not vocalized and tongue tip contact with the upper surface of the vocal tract was achieved, the frame of tongue tip contact was recorded. For all tokens, frame of greatest tongue body retraction and frame of greatest degree of tongue curling were recorded.

3. Results

3.1. /l/ vocalization

Vocalized tokens of /l/ were found in all syllable positions, but rate of /l/ vocalization was affected by syllable position. In
onset and intervocalic positions, vocalized tokens were quite outnumbered by those in which some kind of tongue tip contact was made. However, in coda and nucleus positions the great majority of tokens were vocalized. These results are reported in Figure 3.

### 3.2. Tongue curling & tongue tip contact

Lag times between the consonantal and vocalic tasks of /l/ were recorded to the nearest frame. Because many tokens of /l/ were produced with no discernable tongue tip constriction, achievement of the consonantal task was also measured by the achievement of maximal degree of tongue curling. The lag between the achievement of the vocalic gesture, measured by maximum tongue body retraction, and the achievement of the consonantal gesture, measured by both tongue tip closure and maximum degree of tongue curling, is reported in Figure 4.

In onset and intervocalic positions, achievement of the consonantal gesture, whether measured by tongue tip contact or maximal tongue curling, preceded retraction of the tongue body. However, the reverse sequence was seen when /l/ is in coda and nucleus positions. Here, maximum tongue body retraction preceded the achievement of the consonantal gesture. In all syllable positions, the lag between achievement of the consonantal and vocalic gestures was substantially less when measured from the frame of maximal degree of tongue curling rather than the frame of tongue tip closure. This indicates that whatever the sequencing of tongue tip closure and tongue body retraction, achievement of maximal tongue blade curling occurred between these two events.

### 4. Discussion

There is a fairly clear-cut distinction between American English /l/ in onset and intervocalic positions versus coda and nucleus positions in terms of both rate of vocalization and temporal organization. The fact that tongue tip closure precedes tongue body retraction syllable-initially but follows this retraction syllable-finally is consistent with previous studies such as Sproat & Fujimura (1993) and Browman & Goldstein (1995). This is generally seen as a tendency for the vocalic gesture to occur closer to the syllable nucleus than the consonantal gesture. This study has also determined that maximal tongue curling occurs at some point between the achievements of these two constriction gestures, however they are arranged. Thus, the sequence of events in the production of /l/ appears to be tongue tip closure \( \rightarrow \) maximal curvature \( \rightarrow \) tongue body retraction in onset and intervocalic positions, but tongue body retraction \( \rightarrow \) maximal curvature \( \rightarrow \) tongue tip closure in coda and nucleus positions.

However, one should not think of these two possible sequences as simple mirror images of one another, or think of the tongue curling event as somehow intermediate in its vocalic/consonantal nature between tongue tip closure and tongue body retraction. One should instead consider the fact that in the case of onset and intervocalic /l/, tongue tip contact is followed by greater curling of the tongue, while in nucleus and coda /l/ the degree of tongue curling is diminished once tongue tip contact is made. Put another way, it appears that in onset and intervocalic positions the tongue tip contact has a facilitative effect on tongue curling. It could be, then, that the main production goal for /l/ is not the tongue tip closure, but the curling along the tongue blade, for which tongue tip contact is merely an assistive act of bracing. Figures 5 and 6 illustrate the sequencing of tongue tip closure and maximal tongue curvature in onset and coda positions.
The idea of the tongue tip as a brace during the production of /l/ (and other consonants) in order to facilitate precise tongue shaping has been proposed previously by Stone (1990) and Stone, Faber, Raphael, & Shawker (1992). These studies suggest that tongue tip contact is not just facilitative but necessary to achieve shaping of the tongue beyond the repertoire of basic shapes available to unbraced vocalic postures. However, these studies deal only with tokens of /l/ that involved tongue tip contact. In fact, tongue curling occurs even when /l/ is vocalized, indicating that bracing is not strictly necessary for the production of /l/. Rather, there is a tendency to prefer tongue tip contact to assist in the achievement of tongue curling.

However, this tendency only seems to hold in onset and intervocalic positions, for which the rate of /l/ vocalization is low. The fact that /l/ vocalization varies according to syllable position indicates that bracing plays a different role depending on whether /l/ occurs in onset/intervocalic position or nucleus/coda position. This may be explained by the fact that the relative timing of tongue body retraction is different between these two sets of syllable positions. In onset and intervocalic positions, the tongue body is not yet retracted when tongue tip contact and maximal tongue curling are achieved. The more anterior position of the tongue body may make curling more difficult and in need of assistance in the form of a bracing maneuver of the tongue tip against the alveolar ridge. On the other hand, in coda and nucleus positions, for which the rate of /l/ vocalization is quite high, the degree of tongue curling is actually diminished once tongue tip contact is made. The tongue body is retracted earlier for /l/ in these syllable positions, bringing the tongue into a position that makes tongue curling easier. Thus, tongue tip contact is at best unnecessary for the achievement of tongue curling, and may even be inhibitory. This could explain, at least partially, why there is greater tendency toward vocalization of /l/ in coda and nucleus positions, consistent with the findings of Hardcastle & Barry (1989).

This theory is in stark contrast with previous accounts of the goals of production for liquids within Articulatory Phonology. In this framework, lateralization is captured indirectly through tongue tip and tongue body gestures that stretch the tongue, narrowing it enough to open lateral air channels, rather than the tongue blade curling that is evident in the data presented here. According to this view, the consonantal tongue tip constriction is the primary gesture of the liquid rather than a form of bracing that aids in the achievement of the true goals of production, tongue curling and tongue body retraction. In fact, a shaping-based goal of production is not something that can be currently implemented within the theory, which relies solely on construction goals. However, this study provides evidence that in the case of lateral liquids this stance should be reevaluated.

5. Conclusion

The relative timing of the constriction gestures of American English /l/ and their relationship to the achievement of tongue blade curling indicate that /l/ has a production goal that is based on tongue shaping. Tongue tip contact, rather than being the primary gesture or goal of this consonant, appears to be a form of tongue bracing to achieve complex shaping of the tongue blade and body. Whether this bracing is facilitative to tongue curling depends upon whether the tongue body is retracted, and thus depends upon the syllable position of /l/. At the very least, these findings bring up interesting questions about models of speech that are based entirely on constriction tasks and largely set aside issues of tongue shaping.

6. Acknowledgements

This work was funded through NIH Grant DC007124 to the University of Southern California.

7. References


Investigating the effects of posture and noise on speech production

Ingmar Steiner¹, Peter Knopp², Sebastian Musche², Astrid Schmiedel², Angelika Braun², Slim Ouni³

¹Multimodal Speech Processing, Cluster of Excellence MMCI
²Language Technology Lab, DFKI GmbH
³Computational Linguistics & Phonetics, Saarland University, Germany

Abstract

In recent years, speech production research has benefited greatly from magnetic resonance imaging (MRI). Two problem areas can be identified in conjunction with MRI, however: (a) subjects are typically recorded in supine posture and (b) they have to produce speech in noise. This paper investigates both of these issues by comparing articulatory behavior in upright and supine posture, with and without noise. The production data are recorded using electromagnetic articulography (EMA) and ultrasound tongue imaging (UTI) simultaneously. Preliminary analysis of the EMA data suggests that speakers are affected by posture, noise, and the combination of both in different ways, and use different strategies in compensating for these effects.

Keywords: speech production, posture effect, speech in noise, EMA, ultrasound

1. Introduction

Speech production research, like many other fields, is benefiting in ever-increasing amounts from advances in the area of magnetic resonance imaging (MRI). This non-hazardous, non-invasive imaging modality offers unprecedented insight into the human vocal tract. By acquiring a rapid succession of 2D slices, it is possible to scan the static vocal tract in 3D, or a 2D slice of articulatory movements at video frame rate, with synchronized acoustic recordings (25 to 30 Hz, cf. Narayanan et al. 2013, Niebergall et al. 2013). Within certain constraints, both approaches can even be combined for 4D vocal tract imaging (Zhu et al. 2013).

Two aspects of MRI scanning are however potentially problematic for speech production: (a) the speaker is required to lie in the scanner, and (b) the MRI scanner emits a very loud noise during acquisition. The effects of posture and gravity have been explored in several previous studies using cineradiography (Whalen, 1990), electromagnetic articulography (EMA) (Tiede, Masaki, Wakimoto, et al., 1997), optical tracking (Shiller, Ostry, and Gribble, 1999), X-ray microbeam (Tiede, Masaki, and Vatikiotis-Bateson, 2000), ultrasound tongue imaging (UTI) (Stone, Crouse, and Sutton, 2002; Stone, Stock, et al., 2007; Wrench, Cleland, and Scobbie, 2011), and MRI (Kitamura et al., 2005; Engwall, 2006; Traser et al., 2013). The number of subjects studied is generally very small, and results vary. While Wrench, Cleland, and Scobbie (2011) observe a “slight superior and posterior displacement of the tongue root” in all of their four subjects, Tiede, Masaki, and Vatikiotis-Bateson (2000) note consistent, but differing behavior patterns in the two speakers studied. Stone, Stock, et al. (2007) as well as Kitamura et al. (2005) point to significant between-subject differences. The reason for this seems to be that different subjects choose different strategies to cope with the unusual posture. Considerable caution should therefore be exercised when interpreting the data, and it seems best to start by considering each speaker individually before making generalizations.

The effects of noise on speech production have also been widely studied (Van Summers et al., 1988; Junqua, 1993; Liénard and Di Benedetto, 1999; Lu and Cooke, 2008). Analyses were carried out in the acoustic domain, with inferences made with regard to the underlying articulatory gestures. The previous studies are in good agreement as far as word and vowel duration, F0, overall intensity, and spectral tilt are concerned: most speakers will speak with increased duration, average F0 and intensity, as well as a reduced spectral tilt in a noisy environment (cf. e.g., Van Summers et al., 1988; Junqua, 1993). This is precisely what is generally described as the Lombard reflex (Lombard, 1911). The effect of ambient noise on formants is less clear. Results are fairly consistent with respect to F1: generally, an increase of F1 is observed. Changes in F2, on the other hand, seem to vary with individual speakers and possibly also gender. Junqua (1993) found an increase in female speakers only. Furthermore, effects on formant bandwidth and formant separation have been reported (e.g., Junqua, 1993).

In order to systematically investigate the interactions between posture, noise, and the production of sustained, reiterant, and running speech, we studied these conditions in a factorial design. Preliminary results for the edge vowels of German (/i/, /a/, and /u/) are presented in this paper.

2. Data and methods

Speakers were recorded in supine and upright posture, with and without masking noise, using 3D EMA and synchronized UTI.

A total of 7 speakers were recorded; 3 female, 4 male. All are native speakers of German and thus recorded stimuli in German; one (bilingual) speaker also recorded stimuli in English.

The decision about the materials was not an easy one taking into account that everything had to be recorded in four different conditions: Upright and supine, each in combination with and
without masking noise. Since the EMA coils tend to become detached after some time and may need to be reattached (which makes it more difficult to interpret the data), we decided to limit the material rather than risk having to reattach several coils with the possibility of missing their original placement.

In designing the material, we attempted to emulate established MRI speech production experiments by including a set of sustained speech sounds as well as simple nonsense utterances, and a small number of benchmark utterances. We thus recorded the following production tasks:

1. A set of sustained vowels and diphthongs. The vowels correspond to the “long” vowel phonemes of German and therefore represent extreme vowel qualities as well as the roundedness dimension: /i, e, æ, a, o, u, y, ø, oʊ, ø/.
2. The consonant phonemes of German in an [aCa] context, where C is each of /p, t, k, b, d, g, m, n, η, l, f, v, s, z, ç, j, ŋ, x, r, h/.
3. CV repetitions of the consonants /f, s, ŋ, ç, x, r, m, n, η, l/ in vocastic context /i, a, u/, e.g., [fifififi], do study coarticulatory effects. Since it would have been too time-consuming to include all German consonant phonemes, the plosives were dropped altogether, and only voiceless fricatives were recorded.
4. A repetition of the sustained vowels and diphthongs (see above), in order to study potential compensatory effects.
6. 10 sentences taken from a project corpus designed to study German vowels (kindly provided by Phil Hoole).

2.1. Acquisition setup
Each speaker was recorded in upright (sitting) posture, and in supine posture, lying on a non-ferromagnetic gurney constructed for an earlier study (Steiner and Ouni, 2011). The sequence of conditions was the following:
1. upright without noise;
2. upright with noise;
3. supine with noise;
4. upright without noise.

This allowed the speakers to become accustomed to the EMA coils first, before being presented with the noise condition; conversely, presenting the noise first in the supine condition allowed us to isolate (more or less) the posture effect during noise, before allowing the speakers auditory feedback in the final, supine condition. This rationale reflects the situation of speakers in an MRI speech production experiment, where they have little, if any, opportunity to compensate for posture before the noisy scanning procedure.

The recordings were made simultaneously using a Carstens AG501 articulograph with 16 channels and an Ultrasonix Mindray DP-6600 ultrasound imaging system. The audio was recorded with a Sennheiser MKH816 P48 directional microphone mounted approximately 2 m from the subject. In addition, the entire procedure, which lasted 90 to 120 min per speaker, was documented using a digital video camera.

In order to later synchronize the modalities, a “clicker” was used to record an audible pulse before and after each production task.

2.1.1. Ultrasound tongue imaging
The tongue contour was tracked using an electronic convex array transducer (Mindray 35C20EA).

The video signal from the UTI system was recorded twice: one output was fed to the Articulate Assistant Advanced (AAA) software package, which recorded individual production tasks; the second output was recorded directly in a digital video recorder, which multiplexed the UTI video with the microphone signal and the audio prompts (see section 2.1.3) into a continuous, uninterrupted MPEG-2 stream.

The probe stabilization headset normally used to maintain a constant probe position could not be utilized in our experiment, as its metal parts would have interfered with the magnetic field of the EMA device. Moreover, it would have physically hindered the speakers from resting their heads comfortably in the supine condition. The speakers therefore held the probe by hand during the recordings; its position was monitored and adjusted whenever it deviated from its optimum, while two EMA coils on the probe tracked its position relative to the speaker’s head (see section 2.1.2).

2.1.2. Electromagnetic articulography and acoustic recordings

The positions and orientations of the EMA coils were recorded at 250 Hz. Of the 16 available coils, 13 were attached as follows:
- three reference coils, behind each ear and on the bridge of the nose, in order to correct for head movements;
- two coils on the upper and lower lip, respectively;
- five coils on the tongue, three in the mid-sagittal plane (on the tongue tip, blade and dorsum), as well as one on either side of the tongue blade;
- one coil near the lower incisors to track jaw motion;
- two coils mounted on the UTI probe, one near the top, the second roughly 5 cm further down on the handle. These coils enable tracking the position of the probe throughout the experiment and registration of the UTI and EMA modalities. This technique is similar to the one described by Aron et al. (2006).

The three remaining coils were held back as spares. They were also used to capture 3D palate traces in both postures, and the speaker’s bite plane at the end of each recording session.

The acoustic signal from the microphone was recorded in synchronization with the EMA sweeps at 48 kHz, 16 bit.

2.1.3. Prompt presentation and noise

The stimuli for all production tasks were presented to the speakers via in-ear headphones. All stimuli except the last set had been recorded by a male speaker of Standard German without any regional markers; the 10 sentences were synthesized using a text-to-speech system. Each prompt was followed by two beeps spaced 2 to 5 s apart (depending on the task); the speakers were instructed to speak between these beeps. In the upright posture condition, the stimuli were additionally presented via a computer screen facing the speakers; this allowed them to familiarize themselves with the prompt list during the two upright repetitions with visual, as well as aural, input, before having to rely only on the audio prompts in the supine condition.

The simulated MRI noise was likewise presented via earphones, between the beeps. We selected a recording of gradient echo noise,1 chosen for its roughly uniform structure. The noise level was set according to subjects’ individual tolerance. Across the speakers, the sound pressure level varied from 75 to 90 dB.

1recorded in 2010 at the University of Iowa Hospital Radiology Lab, http://www.cornwarning.com/xfer/MRI-Sounds/
3. Results

3.1. Acoustic analysis

Even though acoustic analysis is not the focus of this paper, analyses were attempted regarding fundamental frequency, as well as the first two formants, in order to establish agreement with previous studies and thus the validity of the data. However, the quality of the audio recordings was far from perfect: while the directional microphone picked up a good-quality signal in upright position, the recording quality in supine position is seriously degraded. Furthermore, it is difficult to find a dedicated software package which will work with noisy recordings, let alone automatically. Therefore, due to the significant noise in the acoustic data, we cannot report reliable formant measurements in this paper.

3.2. EMA analysis

For a preliminary analysis of our data, we selected the sustained vowel prompts for four of the speakers (two male, two female). The vowels were manually labeled based on the recorded audio, and these annotations were used to automatically extract synchronized time segments of the recorded EMA data. For each of the eight measurement coils, we applied a principal component analysis to the position data for the vowel segments to compare the differences between the four experimental conditions for each vowel and coil separately. Figure 1 displays the first principal component (PC1) for female speakers VP05 and VP06 and male speakers VP07 and VP08. Each plotted box represents quartiles 1 to 3, with whiskers extending to ±1.5 interquartile range.

4. Discussion and conclusion

While more thorough analysis is of course planned, and only portions of the data have been annotated so far, the preliminary analysis shown here allow us to make several observations:

• Jaw movement is strongly affected by supine posture and noise for VP07 and VP08.
• All speakers show a clear effect of posture (and to a lesser extent, for noise, except for VP05) regarding lip motion, and characteristic effects of rounding. For the female speakers this is restricted to the lower lip, although this may also be influenced by the individual attachment of the coils.
• Tongue mid, left, and right coils are strongly correlated, with little lateral motion (as expected with the vowel).
• Noise seems to affect the tongue tip motion of VP08 in both positions, while for VP06, it is mainly posture.
• A few coils seem to have become detached or faulty during the recordings, notably those on the lower incisors and tongue back in the supine conditions of VP06. The nature of these erratic measurements is yet to be investigated.

We have yet to analyze the ultrasound recordings; the large amount of manual effort involved makes this a formidable task, but we expect to benefit from the registered EMA tongue coil data to improve the ultrasound processing.

Overall, we can confirm the influence of posture and noise on articulation, and that speakers are affected, or compensate, in different ways. Pending more detailed analysis, we should indeed be wary of these effects when interpreting speech production data acquired in a noisy environment, and in supine posture, such as during speech production MRI studies.

5. References


Figure 1: EMA results for sustained vowels from selected speakers
Intergestural coordination between tonal and oral gestures in Catalan, Italian, and Spanish
Antonio Stella1, Maria del Mar Vanrell2, Massimiliano Iraci1, Pilar Prieto3, & Barbara Gili Fivela1
1CRIL & Università del Salento, Lecce (Italy)
2Freie Universität Berlin (Germany)
3ICREA & Universitat Pompeu Fabra, Barcelona (Spain)
antonio.stella@unisalento.it, mariadelmar.vanrell@fu-berlin.de,
massimiliano.iraci@unisalento.it, pilar.prieto@upf.edu, barbara.gili@unisalento.it

Abstract
In this study we investigate the alignment of L+H* pitch accents used with different pragmatic functions in Catalan, Italian, and Spanish. We collected articulatory and acoustic data from three varieties of these languages with a 3D electromagnetic articulograph (Carstens AG500) and analyzed the stability of the tonal anchoring measuring the lag between tones and articulatory landmarks. Furthermore, we evaluated the differences in the phasing relations between tonal and oral gestures comparing our results with those obtained in different studies on other varieties of the same languages. Results shows that in the three varieties the low tone at the onset of the rising gesture is tightly anchored to the same articulatory landmark of the gestural chair, for both vocalic and consonantal articulators. The same holds true for the high tone at the offset of the tonal gesture, but in this case tonal anchoring shows a higher variability between the speakers and a linguo-specific difference in the lag between the tone and the gestural landmark.

Keywords: intergestural coordination, Articulatory Phonology, tonal alignment, Catalan, Italian, Spanish.

1. Introduction
The issue on tonal alignment is currently widely debated in the analysis of intonation because of its importance as correlate for the encoding of phonological categories. In the last decades, the stability of the synchrony between tones and segments has been widely investigated on acoustic data from the point of view of the Autosegmental-Metrical approach, with the aim to demonstrate the independency from linguistic variables such as syllable structure or time pressure. Unfortunately, the constant presence of small differences between language varieties forecloses a general theory of tonal alignment useful for phonological encoding.
During the last decade, it has been suggested that F0 contours could be seen as sequences of gestures in the sense of the Articulatory Phonology. Under this view, alignment could be regarded as a matter of intergestural coordination between laryngeal and supra-laryngeal gestures (Ladd, 2006; Prieto and Torreira, 2007) in such a way that a tonal gesture is coupled to oral gestures. Articulatory investigations on tonal alignment have already been conducted on various languages (D’Imperio et al., 2007; Prieto et al., 2007; Niemann et al., 2011; Mücke et al., 2012; inter alia) and showed that tonal targets tend to be more closely aligned with articulatory landmarks than with segmental boundaries. For example, Mücke et al. (2012) described the coordination between tones and supra-laryngeal features in Catalan and German LH pitch accents using the framework of the Coupled Oscillator Model (Goldstein et al., 2009) and showed that in Catalan the onset of the rising gesture is synchronous with the onset of the oral constriction gestures in both broad and contrastive focus conditions (the gestures are coupled in-phase), while in German contrastive accents the onset of the rising gesture is systematically delayed.
This work is part of a wider project that aims at investigating the tonal features involved in the production and the perception of different tonal features exploited in broad and narrow-contrastive focus productions of Majorcan Catalan, Lecce Italian and Madrid Spanish (see Vanrell et al., 2013). In these three varieties, a L+H* pitch accent is exploited to convey different pragmatic functions: in Italian it is used as non-focal accent (NF), i.e. in statements produced with a broad focalization, while in Catalan and Spanish it is used as narrow-contrastive focus accent (CF). Vanrell et al. (2013) shows that there is no difference in terms of L and H alignment in the realization of the L+H* accent as produced in Italian NF and Spanish and Catalan CF: indeed, in the three languages L is aligned at the beginning of the tonic syllable and H is aligned at the end.
In the present work we will test the articulatory alignment of the low and high tones at the boundaries of the rising gesture in the production of L+H* in the same varieties of Catalan, Italian, and Spanish. Our aim is to analyze the phasing of the rising gesture with the articulatory gestures involved in the production of the segments and to verify if the tonal anchoring stability is detected also in the articulatory domain of speech.
The determination of the phase relations between tonal and oral gestures, supported by our previous acoustic and perceptual investigation, could lead to a better understanding of the intonational analysis in the framework of the gestural model and of the relation between acoustic and articulatory dimensions of speech.

2. Methods
In order to compare the articulatory differences in alignment of the L+H* pitch accent, a corpus per language was created. The corpora are designed to permit a direct comparison of acoustic and articulatory data among the three languages. Each corpus consists of dialogues between two speakers (identified with the letters A and B). An example is given in Table 1. They are formed by two question-answer exchanges, built in such a way that B productions represent declarative sentences with NF or CF pitch accent in prenuclear position. Both accents occur on target pseudo-words, which are interpreted as a proper name due to the pragmatic context of the dialogue. Target words are composed by reiterated syllables with various segmental compositions ([ma], [na], [la]), as well as different syllable structures (open and closed) and word stress (paroxytones and proparoxytones). Furthermore, the target
words are embedded in carrier sentences controlled for the number of syllables and the stress pattern. In this study only the results related to the productions containing [ma] open syllables are discussed, in paroxytones ([mi.'ma.mi]) and proparoxytones ([mi.'ma.mi.ml]). Each dialogue was presented to the subject in randomized order through a computer screen. The subject listened to the question A (previously produced by a native speaker and recorded) and answered using the B sentence of the mini-dialogue. In this way, we wanted to simulate and reconstruct (as far as possible in a laboratory experimental session) a dialogic exchange for eliciting a natural prosodic style.

Table 1: Examples of mini-dialogues for each language.

<table>
<thead>
<tr>
<th>CATALAN</th>
<th>ITALIAN</th>
<th>SPANISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Què han dit?</td>
<td>A: Cosa ti han detto?</td>
<td>A: Que te han dicho?</td>
</tr>
<tr>
<td>B: Que na Mimami maneja sa mopa</td>
<td>B: La Mimami marina la scuola</td>
<td>B: La Mimami maneja la mopa</td>
</tr>
<tr>
<td>A: Na Mimomi?</td>
<td>A: La Mimami marina la scuola?</td>
<td>A: La Mimami maneja la mopa</td>
</tr>
<tr>
<td>B: no na Mimami maneja sa mopa</td>
<td>B: No, la Mimami maneja la mopa</td>
<td>B: No, na Mimami maneja la mopa</td>
</tr>
</tbody>
</table>

Acoustic and articulatory data were acquired by a 3D electromagnetic articulograph AG500 (Carstens Medizinelektronik, GmbH). The experimental sessions took place at the CRIL Research Centre of the University of Salento (Lecce, Italy) in a quiet room. Speech gestures were tracked by 4 sensors glued on the tongue, 2 on upper and lower lips, 2 on upper and lower incisors and 2 behind the ears, used for the normalization of the head movements. In this study we will analyze the trajectories of the lower lip (LL) to track the production of [m], and of the tongue dorsum (TD) involved in the production of vowels.

We acquired articulatory and acoustic data of 3 speakers for Majorcan Catalan (2 F, 1 M, age: 30–36), 3 speakers for Lecce Italian (3 F, age: 24–27), and 2 speakers for Madrid Spanish (2 F, age: 20–22). Each speaker performed 10 readings of the corpus. In this study a total of 160 tokens have been analyzed (8 speakers x 2 word compositions x 10 repetitions).

Tonal targets and segmental boundaries were automatically labeled and then manually corrected through a Praat script. The onset of the rising gesture (L1) has been labeled at the inflection point of the F0 valley before the tonal rise. In many cases the detection of a single high F0 point was impossible, since the accents were realized with a high attainment phase after the rise. Thus, in all the productions we isolated the attainment of the tonal gesture, identifying the highest F0 value in the accent (H) and calculating the attainment onset and offset (H1 and H2, respectively; see Figure 1). These two points were identified as the first left and right F0 value below the 4% threshold of the H F0 value. This threshold is reported as the perceptually equality range of a tone (House et al. 1999). In our analysis, we considered the point H2, i.e. the end of the attainment phase, as the offset of the rising gesture. In addition, such interpretation of the tonal gesture composition is more in line with the interpretation of the articulatory gesture given by Articulatory Phonology, which considers the attainment phase at the target as part of the gesture itself.

Articulatory data labeling was performed using MAYDAY (Sigona et al., in prep.), a MatLab graphical user interface-based software for multimodal articulatory data inspection and analysis. After importing the compiled Praat TextGrids containing segmental and tonal labels, we labeled the onset, offset and peak velocity of each opening and closing gesture involved in the production of consonants and vowels in the target words. We performed the labeling through a semi-automatic procedure implemented in MAYDAY, identifying the articulatory landmarks on the vertical velocity plot (z-axis) of lower lip (LL), to isolate the consonantal gestures, and tongue body (TD), to isolate the vocalic gestures. The labeling was then inspected and, if necessary, manually corrected.

The computation of the latencies between tonal targets and articulatory landmarks, as well as the graphical plot of the alignment pattern based on the mean values of the measures, was performed through a procedure implemented in a MatLab script. The alignment plot is formed by 4 separate tiers (Segments, Tones, LL, TD), which reports the mean temporal values for the 10 repetitions of each speaker separated for syllable and word type. In each plot, the temporal values of articulatory, acoustic, and tonal landmarks were normalized at the onset of the first segment of the target word. Mean temporal values of segmental boundaries, tones, and articulatory landmarks were plotted on the alignment pattern, which permits the visual inspection of the synchrony between tonal and oral gestures in the productions of each speaker.

3. Results

Figure 2 shows the alignment patterns exhibited in Catalan CF (top) by speaker 2ca, in Italian NF (central) by speaker 3it, and in Spanish CF (bottom) by speaker 2sp. These are the most recurrent patterns in the different conditions. Each pattern is built on the basis of the mean of the target word [mi.ma.ml] produced 10 times by the same speaker. In order to evaluate the anchoring stability and the phasing of oral and tonal gestures, we will statistically analyze the lags of L1 and H2 tones from the nearest LL and TD anchors, identified by the inspections of the alignment pattern. In the next sections the lag measures of each tone from the nearest articulatory landmark will be mentioned as follows:

- **L1 LL**: L1 lag from LL anchor, i.e. maximum constriction of C gesture in the tonic syllable;
- **L1 TD**: L1 lag from TD anchor, i.e. peak velocity of [a]-i gesture in the tonic syllable;
- **H2 LL**: H2 lag from LL anchor, i.e. onset of the C gesture in the final syllable;
- **H2 TD**: H2 lag from TD anchor, i.e. maximum constriction of [a]-i gesture in the post-tonic syllable.

In the following sections we will evaluate the statistical difference in lag measures running a One-Way among the 8 speakers separately (factor SPEAKER). Furthermore, we will evaluate the results of a Two-way ANOVA (3x2) on the same measures to evaluate the statistical difference among Catalan, Italian and Spanish (factor LANGUAGE) in paroxytones and proparoxytones (factor WORD STRESS).
3.1. Onset of the rising gesture (L₁)

Figure 3 shows the boxplots of L₁_LL (top panel) and L₁_TD (bottom panel) lag measure for the 8 speakers in paroxytones (white boxes) and proparoxytones (grey boxes). The dotted line at 0 value represents the position of the articulatory landmark. The measure L₁_LL shows that the low tone at the onset of the rising gesture is tightly anchored to the maximum constriction for the formation of [m] at an average distance of 17 ms from the tone. The measure L₁_TD shows very similar lags of the tone from the peak velocity of the [i] gesture, with an average distance of 25 ms. A One-way ANOVA was conducted on L₁_LL with SPEAKER as factor. It shows a significant difference between speakers [F(7,150)=12.520; p=.000]. A Tukey post-hoc test reveals that the difference is due to the means of 1ca and 2sp compared with the other speakers. A Two-way ANOVA was conducted that evaluated the effect of LANGUAGE and WORD STRESS on L₁_LL. No statistically significant difference is found for WORD STRESS [F(1,152)=.378; p=.540], nor for the interaction between the two factors [F(2,152)=.073; p=.930]. A significant difference is found only for LANGUAGE [F(2,152)=4.300; p=.015] and is due to a difference between Italian and Catalan (p=.023), as a Tukey post-hoc test reveals. A One-way ANOVA test conducted on L₁_TD with SPEAKER as factor shows a significant difference [F(7,150)=12.951; p=.000], with a high variability between the speakers as reported by a Tukey post-hoc test. A Two-way ANOVA conducted on L₁_TD reports no significant difference for LANGUAGE [F(2,152)=.603; p=.549] and WORD STRESS [F(1,152)=.427; p=.514] and no interaction between the two factors [F(2,152)=.509; p=.602].

In sum, both measures L₁_LL and L₁_TD show very tight lags between the tone and the articulatory anchors, and very slight differences between the speakers. The test conducted on the speakers separately shows that not all the speakers of the same language behave in the same way. However, a tight anchoring of the onset of the tonal gesture is found with the offset of the consonantal gesture. This alignment pattern is common to the majority of the subjects, and, above all, with no substantial differences among the three languages in both word stress patterns.

3.2. Offset of the rising gesture (H₂)

Figure 4 shows the boxplots of H₂_LL (top panel) and H₂_TD (bottom panel) lag measure for the 8 speakers. The two measures shows a higher variability in anchoring, if compared to the stability of the L₁ tone. The first variation to notice is linguo-specific. Indeed, Italian speakers tend to align the H₂ tone in a more advanced position with respect to Catalan and Spanish speakers. However, both measures show a high interspeaker variability, also within the same language group. Indeed, as Figure 5 and Table 3 show, H₂_LL is 37 ms on average for Italian, and 39 ms on average for Italian and 43 ms on average for Catalan and Spanish together. The same holds true for H₂_TD, which is 39 ms on average for Italian and 44 ms for H₂_TD for Spanish speakers. The test conducted on the speakers separately shows that not all the speakers of the same language behave in the same way. However, a tight anchoring of the offset of the tonal gesture is found with the offset of the consonantal gesture. This alignment pattern is common to the majority of the subjects, and, above all, with no substantial differences among the three languages in both word stress patterns.
LANGUAGE \[F(2,152)=37.011; p=.000\] and WORD STRESS \[F(1,152)=4.581; p=.034\] and a significant interaction between the two factors \[F(2,152)=8.576; p=.000\]. A Tukey post-hoc test shows no significant difference between Catalan and Spanish (p=.623), while Italian is again statistically different compared to the other two languages.

To sum up, H2 shows higher lags and higher variability in LL and TD anchoring among the speakers of the same language. However, there is a significant difference between Italian vs. Catalan and Spanish, which show no differences in anchoring with both LL and TD. On the other hand, even if a difference is found statistically, there is no change in the articulatory anchors of H2 in the three languages.

![Boxplot of H2 lag from LL (top) and TD (bottom) anchors for the 8 speakers.](image)

### 4. Discussion and conclusion

The lag measures obtained in our study show a high stability in the articulatory alignment of the rising tonal gesture as produced in Catalan, Italian, and Spanish. The onset of the rising gesture, i.e. L1 tone, shows the greatest stability, and is anchored with the maximum constriction of the consonantal gesture for [m] and the peak velocity of the vocalic gesture [i]-[a] in the tonic syllable. Furthermore, such an alignment seems to be stable in the three languages under investigation and across the majority of the speakers. A stable pattern is also shown by the offset of the rising gesture, i.e. H2 tone, but the lag increases in Italian with respect to Catalan and Spanish. However, such an increase does not influence the anchoring of the tone. Indeed, in the three languages it is alwaysaligned with the onset of the consonantal gesture after the post-tonic syllable and the maximum constriction of the [a]-[i] gesture.

In the framework of the Coupled Oscillator Model of speech production (Goldstein et al., 2009), the intergestural coordination shown in our data would put the rising tonal gesture in anti-phase relation with both consonantal and vocalic gesture. These results are not in line with previous investigations on Central Catalan (Mücke et al., 2012), in which tonal, consonantal, and vocalic gestures are claimed to be synchronous in CF productions, showing an in-phase coordination. The same happens comparing our results to the ones shown for Bari and Bologna Italian (Niemann et al., 2011), in which the same in-phase pattern is found for BF production. It is likely that such discrepancies are due to differences in the experimental design, as in the case of Mücke et al. (2012), or, more probably, to differences in the intonational system of the varieties under study, as in the case of Niemann et al. (2011). Indeed, as shown in Gili Fivela et al. (to appear), where the data of 13 varieties of Italian are discussed for the AH (Atlas of Italian Intonation) project, Italian intonation changes consistently depending on the variety under consideration, with many varieties having highly different patterns for the same pragmatic function.

The analysis of the other part of our corpus produced by the same speakers could shed some light on this issue. Indeed, our data involve different phonological categories, i.e. H*+L produced in Italian CF, and L+<H* produced in Catalan and Spanish BF. Furthermore, our data contain also target words formed by reiterated [na] and [la] syllables, e.g. [na.na.na] and [na.na.na.na]. These data should show more controlled alignment patterns since there should be no competition between consonantal and vocalic gestures, and could be crucial to obtain a wider view of the phase relation between tonal and oral gestures in these three languages.

### 5. References


Adaptation of trajectory shapes during conversation

Mark Tiede1, Christine Mooshammer2,1, Dolly Goldenberg3,1, Douglas N. Honorof4

1Haskins Laboratories, New Haven, USA
2Humboldt-Universität, Berlin, Germany
3Yale University, New Haven, USA
tiede@haskins.yale.edu, mooshamc@hu-berlin.de, dolly.goldenberg@yale.edu, honorof@haskins.yale.edu

Abstract

Speech audio and articulatory movements of age- and gender-matched speaker pairs have been recorded in face-to-face spontaneous conversation using two electromagnetic articulometer (EMA) systems simultaneously. Elicited tasks included synchronized, imitated and spontaneous speech interspersed with repeated baseline utterances for evaluation of mutual accommodation. Convergence measures have been obtained using between-speaker (for magnitude) and within-speaker (for symmetry) distances computed on coda velar gestures and the preceding vowel from baseline tasks produced before and after conversational interaction. The linearly normalized velar trajectories have also been compared for evidence of increased similarity in shape. Results are equivocal, indicating that in some cases convergence can persist as an aftereffect, but more generally supporting the view that it is active interaction that drives mutual accommodation, and without it the effect collapses, leaving production to drift.

Keywords: phonetic convergence, speech kinematics, EMA

1. Introduction

When two speakers interact in a conversation their patterns of speech may gradually become more similar. This adaptation effect, known as convergence, has been observed on many linguistic levels (Pickering & Garrod, 2004). To date most evidence for convergence on the phonetic level has been acquired from acoustic studies and perceptual similarity ratings (e.g. Pardo 2006). Recently we have begun a project to collect kinematic data from interacting speaker pairs to investigate whether convergence can also be observed in the underlying articulation.

Although interaction between two speakers may fail to produce convergence, operate asymmetrically, or even result in diverging speech patterns, in general it has been observed that convergence is facilitated when two individuals are aligned along such parameters as age, gender, and social background (Babel 2012). It is also known that convergence is enhanced by direct access to visual cues provided by the face (e.g. Hazan et al. 2005). Our view is that if convergence does occur it is driven by production-perception links that organize a loose coupling between two interacting systems (Beek et al. 1992). This entrainment serves as a forcing function on internal dynamics, leading to adjustments in existing patterns of behavior within their intrinsic range and resulting in relative coordination between the two speakers (as contrasted with absolute coordination between physically coupled systems). Convergence observed in the acoustic signal is thus a surface manifestation of underlying alignment in articulation.

In the current study we are interested in whether adaptation can be found on purely motoric levels, as in the tongue dorsum movements associated with velar stops. It has been previously observed that following back vowels the tongue dorsum slides forward along the palate during closure (Mooshammer et al. 1995). These so-called forward loops have been found in many languages and show speaker-specific patterns in the extent of the forward movement. Conversely, following front vowels the tongue dorsum does not typically slide during velar closure. Listeners are aware of the naturally occurring shape variation of the trajectory and rate voiced velar stops in back vowel context as more natural if they are produced with a forward loop compared to a straight movement or a backward loop (cf. Nam et al. 2013). In this work we compare the gestural characteristics and shape of velar loops in various contexts obtained from interacting speaker pairs pre- and post-adaptation to a series of conversational tasks. We expect the effects to be most pronounced in those cases where changes towards increased similarity are also observed within such acoustic parameters as vowel quality.

2. Method

2.1. Participants

Ten native speakers of American English were matched for age, gender, social background, personality (assessed from responses to a questionnaire), and dialect. Each passed a 20 dB pure-tone hearing test and had no self-reported speech deficits. One male and one female pair are presented here.

2.2. Experimental tasks

In each experiment participants performed a range of tasks that included synchronized (choral production), imitated and spontaneous speech, interspersed with repeated baseline utterances used to evaluate mutual accommodation. These baseline tasks were produced separately by each participant and consisted of focus words (Table 1) produced within a consistent context (“Say ___ it again”).

| Table 1: Focus words produced in baseline tasks. |
| bag  | pad  | pack |
| beg  | peg  | peck | deck |
| big  | pig  | pick | Dick |
| bog  | pod  | pock | dock |

Within the baseline tasks each word was elicited three times in randomized order.

2.3. Experimental procedures

To record speech articulator movements for two interacting speakers simultaneously we rely on separate electromagnetic
articulometer (EMA) devices, one for each talker. Two types of commercially-manufactured EMA systems are used together. The AG500 (Carstens MedizinElektronik, GmbH) has six narrowly tuned transmitters that operate continuously at different frequencies (7.5kHz – 13.75kHz). The WAVE (Northern Digital, Inc.) uses eight strobed transmitters, all operating at 3kHz. Both systems resolve three spatial and two angular orientation measurements per sensor at sampling rates of at least 100 Hz. Crucially both systems permit unrestricted head movement and provide an unimpeded view of the face.

In pilot work to validate the assumption that the different operating frequencies of the AG500 and WAVE devices would support simultaneous operation, a series of benchmark tests were performed. With the measurement centers of each system positioned 2 meters apart, the stability of fixed distances between sensors attached to a rotating rigid body within the field of each device was assessed, with and without the other device in active operation. Results based on rotational symmetry and the standard deviations between pairs of fixed sensors showed no significant effect of dual operation on either system (Tiede et al. 2012).

For each participant sensors were glued to two points on the tongue (blade and dorsum), the lower incisors, and the upper and lower lips. Additional sensors placed on the upper incisors and left and right mastoid processes were used to correct for head movement. Independent audio tracks were recorded at 44.1 kHz using separate directional microphones located about 50 cm from the mouth. Participants were seated such that their anterior vocal tracts were centered within the respective device fields, for a face-to-face distance of slightly less than 2 meters, and with a clear view of their partner’s face. All tasks were presented on separate monitors specific to each speaker. Stimulus presentation and data acquisition were coordinated by custom software (Marta, Haskins Laboratories). Figure 1 shows the experimental arrangement.

Data for each speaker were remapped to a movement-corrected standard orientation aligned with the occlusal plane established with a biteplane trial. To reduce noise, reference sensor trajectories were low-pass filtered at 5 Hz and movement sensor trajectories at 20 Hz. Alignment between the two EMA data streams was effected through cross-correlation of their respective audio.

2.4. Measurements

F0 and formant values were computed at the midpoint of each focus word vowel. Tongue dorsum movements for the velar gestures of the focus words were labeled using Mview (Haskins Laboratories). The /g/ of “again” in the context sentence was also labeled. A 20% thresholding criterion applied to local peak sensor velocity was used to identify the closing gesture onset (GONS) and opening offset (GOFFS), together with closing and opening peak velocities and the point of maximum constriction. Gestural ‘stiffness’ (STIFF\textsubscript{close}, STIFF\textsubscript{open}) was computed as the peak velocity divided by the sensor distance traveled during the closing and opening movements. Normalized trajectories for morphological comparisons were obtained using linear interpolation to a standard number of samples over the GONS : GOFFS range.

Phonetic convergence has been quantified in three ways, using in each case corresponding focus words from the initial (PRE) and final (POST) series of baseline tasks. The first measure, called BETWEEN, is the average between-speaker Euclidean distance between each pair’s matched values contrasting the initial and final distances; it shows the extent of any convergence (speakers who converge will show smaller BETWEEN values for the POST task). Distances are computed on [F1xF2] pairings in mel space for formant measures, and as absolute differences for monodimensional values.

The second measure, called WITHIN, is the within-speaker Euclidean distance between initial and final matched values for each speaker. It is used to assess the asymmetry of convergence: speakers who accommodate to one another symmetrically will show approximately equal WITHIN values.

The third measure, called DIRection, contrasts the averaged initial and final values of the measured parameter separately by speaker. It is used to determine the direction of any change, in particular whether such change was in the (converging) direction of the partner’s values.

Differences between velar loops are computed on the normalized trajectories by summing the Euclidean distances between corresponding samples, with comparisons made BETWEEN and WITHIN speakers as described above.

3. Results

Despite the instance of vowel-specific convergence shown in Figure 2, results obtained for male pair C01 in general support an asymmetric pattern for the acoustic measures (i.e., larger...
changes in WITHIN values for the second member of the pair) and noisy results for the kinematics:

\[
\begin{align*}
F0 & \quad \text{PRE} & \quad \text{POST} \\
\text{BETWEEN:} & \quad 7.5 & \quad 6.9 & \quad t(10170) = -4.8 * \\
\text{WITHIN:} & \quad 6.9 & \quad 8.3 & \quad t(10130) = -12.2 * \\
\text{FMTS} & \quad 114.1 & \quad 108.5 & \quad t(52) = 0.2 \\
\text{WITHIN:} & \quad 37.0 & \quad 108.8 & \quad t(52) = -2.6 * \\
\text{STIFF\_close} & \quad 0.07 & \quad 0.08 & \quad t(38) = -0.9 \\
\text{WITHIN:} & \quad 0.02 & \quad 0.01 & \quad t(38) = 1.1 \\
\text{STIFF\_open} & \quad 0.07 & \quad 0.06 & \quad t(43) = 0.8 \\
\text{WITHIN:} & \quad 0.04 & \quad 0.03 & \quad t(43) = 0.9 \\
\end{align*}
\]

The normalized trajectories for the male pair show a pattern of divergence, both in the summed distances (Table 2) and in the shapes of the loops (Figure 3).

### Table 2: Summed Euclidean distances between corresponding samples of normalized trajectories (male pair C01)

<table>
<thead>
<tr>
<th></th>
<th>BETWEEN</th>
<th>WITHIN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pre</td>
<td>post</td>
</tr>
<tr>
<td>/a/</td>
<td>199.4</td>
<td>410.0</td>
</tr>
<tr>
<td>/æ/</td>
<td>232.8</td>
<td>473.1</td>
</tr>
<tr>
<td>/h/</td>
<td>330.7</td>
<td>273.1</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>303.7</td>
<td>346.4</td>
</tr>
<tr>
<td>‘again’</td>
<td>342.1</td>
<td>514.3</td>
</tr>
</tbody>
</table>

Figure 3: Mean normalized trajectories (male pair C01). “o” marks initial sample.

The female pair (C09) shows some evidence for acoustic convergence, but it is asymmetric with the second member of the pair adapting more. The gestural stiffness measures show divergence:

\[
\begin{align*}
F0 & \quad \text{PRE} & \quad \text{POST} \\
\text{BETWEEN:} & \quad 21.8 & \quad 16.9 & \quad t(11412) = 7.8 * \\
\text{WITHIN:} & \quad 16.3 & \quad 18.2 & \quad t(11376) = -12.1 * \\
\text{FMTS} & \quad 397.6 & \quad 256.1 & \quad t(52) = 1.4 \\
\text{WITHIN:} & \quad 248.9 & \quad 345.1 & \quad t(52) = -0.9 \\
\text{STIFF\_close} & \quad 0.02 & \quad 0.06 & \quad t(42) = -5.9 * \\
\text{WITHIN:} & \quad 0.04 & \quad 0.02 & \quad t(42) = 2.3 * \\
\text{STIFF\_open} & \quad 0.02 & \quad 0.07 & \quad t(42) = -4.1 * \\
\text{WITHIN:} & \quad 0.05 & \quad 0.03 & \quad t(42) = 1.1 \\
\end{align*}
\]

However, normalized trajectories show convergence not only on the summed distance measures (Table 3), but in increased similarity of loop shape as well (Figure 4).

### 4. Discussion and conclusion

The preliminary results presented here are useful in illustrating methods for quantifying kinematic convergence while at the same time underscoring their limitations. In particular, the lack of a consistent response across acoustic and kinematic measures calls into question the sensitivity of the PRE vs. POST tasks for detecting possibly subtle convergence effects. The underlying assumption is that once shifted within their respective production ranges speakers would stay shifted as a form of production aftereffect, and indeed the female pair does show some evidence for a persistent shift towards greater shape similarity in the velar trajectories (Figure 4).

However, because participants perform these baseline tasks reading from a screen and without interaction with their partner it now seems more likely to us that in the POST phase the coupling driving the effect collapses, leaving production to drift. To test this idea we are currently investigating whether more systematic effects can be observed on the same focus words embedded in the speech tasks during which partners were actively engaged with one another.

### Table 3: Summed Euclidean distances between corresponding samples of normalized trajectories (female pair C09)

<table>
<thead>
<tr>
<th></th>
<th>BETWEEN</th>
<th>WITHIN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pre</td>
<td>post</td>
</tr>
<tr>
<td>/a/</td>
<td>797.7</td>
<td>640.1</td>
</tr>
<tr>
<td>/æ/</td>
<td>675.7</td>
<td>444.6</td>
</tr>
<tr>
<td>/h/</td>
<td>591.5</td>
<td>341.6</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>838.2</td>
<td>502.2</td>
</tr>
<tr>
<td>‘again’</td>
<td>365.9</td>
<td>132.1</td>
</tr>
</tbody>
</table>

Figure 4: Mean normalized trajectories (female pair C09). “o” marks initial sample.

### 5. Acknowledgements

This work is supported by a grant from the U.S. National Institutes of Health (NIDCD R21-DC012350).
6. References


Vowel articulation affected by word frequency

Fabian Tomaschek1, Benjamin V. Tucker2, Martijn Wieling3, R. Harald Baayen1

1Department of Linguistics, Quantitative Linguistics, University of Tuebingen, Germany
2Department of Linguistics, University of Alberta, Canada
3Department of Information Science, University of Groningen, The Netherlands
fabian.tomaschek@uni-tuebingen.de, btucker@ualberta.com, wieling@gmail.com, harald.baayen@uni-tuebingen.com

Abstract
A frequently replicated finding is that the frequency of words affects their phonetic shape. In English, high frequency words have been shown to contain more centralized vowels than low frequency words. By contrast, a recent study on vowel articulation in German has shown a contrary finding. At the gestural level, tongue movements in HF words showed more extensive vowel targets and less coarticulation with consonants. This paper further evaluates the later finding by taking into account a large set of verbs covering the continuum between high and low frequency. In addition to frequency the effects of two factors were analyzed: inflection (sagt vs. sagen) and speech rate (normal vs. fast). Our results imply that language experience increases the proficiency with which words are articulated: speakers are able to plan and target tongue movements earlier.

Keywords: articulography, vowels, word frequency, learning

1. Introduction
It is well known that how often a word is pronounced affects its phonetic form. In contrast to low frequency words (LF), high frequency (HF) words tend to have a lower number of segments (Zipf 1935) and shorter acoustic durations (Gahl 2008) as well as a higher probability of deleting a segment (Aylett and Turk 2004; Munson and Solomon 2004). At the segmental level, acoustic data indicates that vowels in English HF words form a more contracted and centralized vowel space than vowels in LF words (Munson and Solomon 2004; Munson 2007). Zipf (1935) explained such reduction processes by the principle of least effort. The more often an articulatory movement is executed, the more efficient it is performed insofar as ‘unnecessary’ movements are omitted. Aylett and Turk (2004) in light of their Smooth Signal Redundancy Hypothesis -SSRH- further show that higher contextual predictability of a word increases the reduction process.

In Tomaschek et al. (2013), we investigated frequency effects at the articulatory level in German. We found that HF produced stonger peripheral articulation in [i] vowels and less coarticulation in [a] vowels. We reasoned that these effects indicate another important process in language experience: learning. Higher frequency not only enables the speaker to articulate more efficiently but also more precisely. We can consider this effect as another mode of articulation, which we want to call the articulatory proficiency theory -APT-. The present study is a follow up of Tomaschek et al. (2013) further investigating APT in words containing [a]. Unlike Tomaschek et al. (2013), who used words from the extremes of the continuum in a categorical way, we increased the number of words along the entire frequency continuum.

In addition to frequency, we introduced the following two factors: speech rate and morphology by means of inflection. It has been shown that faster speech rate leads to a stronger temporal contraction (Hoole, Mooshammer, and Tillmann 1994) and centralization (Moon and Lindblom 1994) of vowels. The SSRH predicts that vowels in HF words produced at a fast speech rate should show the most reduction. By contrast, the APT predicts that vowels in HF words will show less reduction due to a fast speech rate than those in LF words as learning should enable the speaker to counteract reduction due to fast speech.

The manipulation of the inflection addresses the question of how inflected forms, especially regulars, are stored in the mental lexicon. Are they generated by rules that transform a stem from the lexicon? Or are they stored independently as suggested by studies with lexical decision tasks (Milin, Filipovic Durdevic, and Moscoso del Prado Martin 2009). For example, Stemberger and MacWhinney (1986) investigated the occurrence of speech errors by having subjects pronounce past tense forms of regular verbs. They found that regular HF verbs were less prone to errors than LF verbs. On the one hand, their results indicate that inflected forms, even regulars, are stored in the lexicon. On the other hand, they show that the experience with specific words results in improved mastery of those words. We hypothesize that if the inflected form is stored in the lexicon, higher frequency should facilitate the planning/articulation of these forms, insofar that the articulation of the inflection is anticipated and produced earlier.

2. Stimuli and Methods
2.1. Stimuli
27 German verbs were used as stimuli. Their first syllable was stressed and contained the phonologically long [a] vowel (see Table 1). In the disyllabic condition, words were produced in a "sie ...." they .... context: sie zahlen [zi: tsalan], sie mahnen [zi: ma:nan]. Nine of these were monosyllabic inflected forms produced in a "ihr ...." you pl. .... context: ihr zahlt [i: htsalt], ihr mahnt [i: hmant]. We used the logarithmic relative frequency (henceforth frequency or log(P)) of a word in the SDEWAC corpus (Shaoul and Tomaschek 2013). In addition, the consonants before and after the vowel were controlled for place of articulation: coronal-V-coronal, coronal-V-labial or labial-V-coronal.
Table 1: Stimulus material, ordered by frequency: C-C = place of articulation of consonants next to the vowel: coronal (c) or labial (l). log(P) = logarithm of relative frequency. Monosyllabic words are written in italics.

<table>
<thead>
<tr>
<th>Word</th>
<th>C-C</th>
<th>log(P)</th>
<th>Word</th>
<th>CVC</th>
<th>log(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>zahlen</td>
<td>c-c</td>
<td>-14.20</td>
<td>waten</td>
<td>l-c</td>
<td>-21.35</td>
</tr>
<tr>
<td>schlafen</td>
<td>c-l</td>
<td>-15.53</td>
<td>lahnnt</td>
<td>c-l</td>
<td>-21.49</td>
</tr>
<tr>
<td>zahnt</td>
<td>c-c</td>
<td>-15.79</td>
<td>labern</td>
<td>c-l</td>
<td>-21.51</td>
</tr>
<tr>
<td>schaden</td>
<td>c-c</td>
<td>-16.59</td>
<td>faseln</td>
<td>l-c</td>
<td>-22.06</td>
</tr>
<tr>
<td>baden</td>
<td>l-c</td>
<td>-17.99</td>
<td>schaben</td>
<td>c-l</td>
<td>-22.24</td>
</tr>
<tr>
<td>mahnt</td>
<td>l-c</td>
<td>-18.45</td>
<td>latschen</td>
<td>c-c</td>
<td>-22.36</td>
</tr>
<tr>
<td>blasen</td>
<td>c-c</td>
<td>-19.07</td>
<td>schlaft</td>
<td>c-l</td>
<td>-22.45</td>
</tr>
<tr>
<td>bahnt</td>
<td>l-c</td>
<td>-19.14</td>
<td>schabt</td>
<td>c-l</td>
<td>-23.18</td>
</tr>
<tr>
<td>bahn</td>
<td>l-c</td>
<td>-19.25</td>
<td>tafeln</td>
<td>c-l</td>
<td>-23.35</td>
</tr>
<tr>
<td>mahnen</td>
<td>l-c</td>
<td>-19.31</td>
<td>latscht</td>
<td>c-c</td>
<td>-23.72</td>
</tr>
<tr>
<td>stapeln</td>
<td>c-l</td>
<td>-19.66</td>
<td>blast</td>
<td>c-c</td>
<td>-24.91</td>
</tr>
<tr>
<td>fahnden</td>
<td>l-c</td>
<td>-20.42</td>
<td>zahnnt</td>
<td>c-c</td>
<td>-25.03</td>
</tr>
<tr>
<td>tadeln</td>
<td>c-c</td>
<td>-20.60</td>
<td>zahlen</td>
<td>c-c</td>
<td>-25.24</td>
</tr>
<tr>
<td>lahnen</td>
<td>c-l</td>
<td>-20.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

2.2. Recording method

All recordings were conducted in a sound proof booth at the Department of Linguistics of the University of Tubingen. A total of 17 native German subjects (mean age: 26, sd = 3) were instructed to read the stimuli aloud after being presented on a computer screen. Each word in each context was presented once. The list was pseudo-randomized for each participant and divided into three parts. Each part was presented once in a slow (inter-stimulus-time: 600 ms; presentation-time: 800 ms) and once in a fast speaking condition (inter-stimulus-time: 300 ms; presentation-time: 450 ms).

Articulatory movements of the tongue were recorded with the NDI wave articulograph at a sampling frequency of 100 Hz. Simultaneously, the audio signal was recorded (Sampling rate: 22.05 kHz, 16bit) and synchronized with the articulatory recordings. To correct for head movements and to define a local coordinate system, a reference sensor was attached to the subjects’ forehead. Before the tongue sensors were attached, a recording was made to determine the rotation from the local reference to a standardized coordinate system. The standardized coordinate system was defined by a bite plate to which three sensors in a triangular configuration were attached. Tongue movements were captured by three sensors: one slightly behind the tongue tip (TT), one at the tongue middle (TM) and one at the tongue body (TB; distance between each sensor: around 2cm). The present analysis focuses on the sensors TT and TB.

2.3. Preprocessing

The recorded positions of the tongue sensors were centered at the mid-point of the bite plate and rotated in such a way that the front-back direction of the tongue was aligned to the x-axis with more positive values towards the front of the mouth, and more positive z-values towards the top of the oral cavity. No filtering was applied as this would artificially increase the autocorrelation of the data. To determine segment boundaries, the audio signal was automatically aligned with phonetic transcriptions by means of a Hidden-Markov-Model-based forced aligner for German (Rapp 1995). Alignments were manually verified and corrected where necessary. The beginning (henceforth CV transition) and offset time points (henceforth VC transition) of each vowel in every word were used to identify the movement trajectories of the three tongue sensors.

3. Analysis and results

3.1. Analysis

Since the duration of each vowel differs from utterance to utterance per person and word, vowel duration was normalized between 0 and 1 (henceforth called time). Separate analysis of the vertical movement as a function of time in each of the sensors (TT, TB) was performed by means of generalized additive models (GAMs) (R version 3.0.2, package mgcv, Version 1.7-28, Wood 2006). GAMs model the nonlinear relationships between the numeric predictor and the response variable with thin plate regression splines. Interactions between two gradual predictors are modelled by means of tensor product smooths with cubic spline basis functions and result in wiggly surfaces. The estimated degrees of freedom (edf) reflect the number of parameters required for modeling a wiggly curve, surface or hypersurface and measure how wiggly it is. More wiggly curves, surfaces or hypersurfaces require more edfs.

As the exact tongue movements might differ across subjects due to different morphologies of the oral cavity, by-subject random factor smooths as a function of time were included. In order to account for random item effects, random factor smooths for time per word were included. These random factor smooths have the same function as the combination of random intercepts and random slopes in a standard linear mixed-effects regression analysis. Random factor smooths were also included for CV and VC consonant place of articulation. Random wiggly curves were significant for the variation by participants, words and place of articulation in CV and VC consonants (Tables 2 and 3).

Articulatory data constitute time series with strong autocorrelation (i.e. one can predict the value in X+1 given the value in X). Residual autocorrelation results in anti-conservative p-values. We therefore included a parameter rho to remove autocorrelation noise from residuals. Remaining errors were Gaussian and uncorrelated. Autocorrelation was estimated on the basis of the first model. This first estimate was used during model optimization. After the optimal model was found, autocorrelation was estimated anew and optimized.

Model selection was based on model comparison with the maximum likelihood (ML) test. The optimal models are presented in Table 2 and 3. A detailed description of this approach can be found in Kryuchkova et al. (2012).

3.2. Results

3.2.1. Contour plots

We use contour plots to show fitted regression surfaces. For example, the contour plot in Fig. 1 displays the tongue tip sensor’s vertical position (in mm) as a function of time (x-axis) and as a function of frequency (y-axis). Lighter shades indicate high positions, darker shades indicate low positions. Contour lines connect points with equal elevation. The movement for a certain probability is represented in the vertical axis. For example, the tongue tip’s movement at log(P) = -20 starts at a height of 4 mm above the mean, falls to -2 mm (i.e. 2 mm below mean) and then rises to 4 mm again.
3.2.2. The tongue tip sensor (TT)

The model for the movement in the TT sensor yields an R-squared of 0.791 and explains 79.5% of the deviance in the data (ML: 24526, edfs: 388). No significant effects of speech rate and morphological alternation were found. The tensor model indicates that the time-by-frequency interaction is significant (Table 2, first row). In Fig. 1, one can see that the time point of the vowel target changes minimally as a function of frequency. Furthermore, the CV transition onset is higher at log(P) = -24 than at higher frequencies. Also, the VC transition raises earlier at this frequency than at higher frequencies. In all, the effect of frequency was tiny. Neither the falling and rising pattern of the tongue tip, nor the depth of the vowel target was drastically affected by frequency.

![Figure 1: Vertical position of the tongue tip sensor (TT) as a function of time (x-axis) and logarithmic relative frequency (y-axis): Lighter shades indicate high positions, darker shades indicate low positions. Contour lines connect points with equal elevation.](image)

Table 2: GAM table for TT

<table>
<thead>
<tr>
<th>Effect</th>
<th>edf</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>tensor(time, frequency)</td>
<td>12.16</td>
<td>80.34</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>smooth(time, participant)</td>
<td>141.50</td>
<td>54.47</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>smooth(time, onset place)</td>
<td>8.67</td>
<td>26.43</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>smooth(time, word)</td>
<td>215.83</td>
<td>10.61</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

3.2.3. The tongue body sensor (TB)

The model for TB sensor yields an R-squared of 0.40 and explains 41.6% of the deviance in the data (REML-score: 20358, edfs: 314). No significant effect of speech rate was present. The model indicates a significant morphology-by-frequency interaction (Table 2 first and second row; Fig. 2). In the disyllabic condition (stem + /ə/), the vowel is produced lower with increasing frequency. Simultaneously, the CV and VC transitions become steeper. In the monosyllabic condition (stem + /h/), the movement pattern is reversed. With higher frequency, the vowel target becomes higher and the onset of the VC transition starts earlier.

![Table 3: GAM table for TB](image)

<table>
<thead>
<tr>
<th>Effect</th>
<th>edf</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>tensor(time, frequency: 2syll.)</td>
<td>8.31</td>
<td>22.87</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>tensor(time, frequency: 1syll.)</td>
<td>15.64</td>
<td>26.12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>smooth(time, participant)</td>
<td>136.35</td>
<td>23.91</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>smooth(time, onset place)</td>
<td>7.04</td>
<td>6.71</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>smooth(time, offset place)</td>
<td>6.86</td>
<td>5.22</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>smooth(time, word)</td>
<td>137.30</td>
<td>1.49</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

The present study investigated the effects of morphological inflection, speech rate and frequency of occurrence on the articulation of the German vowel [a]. Two sensor positions, tongue tip and tongue body, were analyzed. The present findings indicate that none of the factors under investigation affected the movement pattern of the tongue tip. One possible reason for this finding might be that [a] is primarily articulated by the body of the tongue whereas the tongue tip is used primarily for the production of coronal consonants. It is possible that word frequency effects on the CV and VC transitions were confounded by the articulation of the surrounding consonants. As Munson (2011) has shown that more frequent two-consonant clusters are articulated shorter than less frequent ones, phonotactic frequencies would be probably a better measure to investigate usage effects in the tongue tip.

Hoole, Mooshammer, and Tillmann (1994) report that faster speech rate results in a contraction of the vowel. However, we found no effect of speech rate at any of the sensors under investigation. This might be a result of the present analysis technique which required normalization of vowel duration between 0 and 1. Possible contractions might have been normalized out.

Frequency affected the movement patterns of the tongue body. In the monosyllabic condition, the [a] vowel was produced with a less extensive, i.e. a higher and more centralized target. Simultaneously, the outgoing VC transition became smoother. This finding seems in line with the SSRH (Aylett and Turk 2004) and the principle of least effort (Zipf 1935), which state that the more frequent a certain word is, the less effort is invested into producing it. This is why the vowel would be centralized and stronger coarticulated with the consonantal context (Munson and Solomon 2004; Munson 2007).

In the disyllabic condition, the opposite pattern is visible. With increasing frequency, the vowel target is articulated more extensively and the CV and VC transitions become steeper. Steeper transitions indicate less coarticulation between vowel and consonant, as has been show by Katz and Bharadway (2001). This finding replicates Tomaschek et al. (2013).

How is it possible that the same vowel is affected in two different ways by frequency, depending on whether it is produced in a monosyllabic or in a disyllabic word? One could argue that in the disyllabic condition, there is no need to reduce the vowel in the first syllable. Rather, reduction occurs in the second unstressed syllable where the [a] is often non-existent in modern German and the sonorant becomes syllabic (Becker 1998). Since in the monosyllabic condition there is no unstressed syllable where reduction might be focussed, it is realized on the vowel itself.

However, we would like propose another explanation for this
finding. In the monosyllabic condition, a [t] had to be attached to the stem. In order to produce the [t], the tongue body raises so that the occlusion which occurs between the VC consonant and the [t] can be produced. We observed that the tongue body raises earlier, the higher the frequency. This early raising implies earlier preparation of the articulation of the [t]. Given word frequency as a measure of experience, our results indicate a learning effect in both conditions: In the disyllabic condition, the vowel shows less coarticulation with increasing frequency; in the monosyllabic condition, the final segment is anticipated earlier.

In summary, our data show that word frequency does not affect the movement patterns of the tongue tip. Frequency manifests itself in the tongue body, where it interacts with the word’s inflectional form. Our articulatory proficiency theory APT captures this phenomenon insofar as acoustically measured reductions at the spectral and temporal level turn out to be an earlier preparation of articulatory movements. Nevertheless, on the basis of the present data, no conclusion can be drawn and open questions have to be solved in future studies.

5. Acknowledgements

We would like to thank our student assistants Franziska Bröker, Dankmar Enke, Lea Hofmaier and Samuel Thiele for their in-exhaustible willpower while correcting the annotations. This paper was funded by the Alexander von Humboldt Professorship.

6. References


Munson, B. (2007). “Lexical access, lexical representation, and vowel production”. In: Laboratory Phonology 9, 201–228.


Shaoul, C. and F. Tomaschek (2013). “A phonological database based on CELEX and N-gram frequencies from the SDEWAC corpus”. In: Personal communication.


'Hearing tongue and seeing voices': neural correlates of audio-visuo-lingual speech perception.

Avril Treille¹, Coriandre Vilain¹, Thomas Hueber¹, Jean-Luc Schwartz¹, Laurent Lamalle², Marc Sato¹

¹GIPSA-lab, Département Parole & Cognition, CNRS & Grenoble Université, Grenoble, France
²Inserm US 17 / UMS IRMaGE, CHU de Grenoble / CNRS UMS 3552, Unité IRM 3T Recherche, Grenoble, France

avril.treille@gipsa-lab.grenoble-inp.fr, coriandre.vilain@gipsa-lab.grenoble-inp.fr, thomas.hueber@gipsa-lab.grenoble-inp.fr, jean-Luc.Schwartz@gipsa-lab.grenoble-inp.fr, laurent.lamalle@ujf-grenoble.fr, marc.sato@gipsa-lab.inpg.fr

Abstract
The present fMRI study examined the neural substrates of auditory, visual and audio-visual speech perception in relation to either labial or lingual movements (acquired with a camera and an ultrasound system). Common overlapping activities between modalities were mainly observed in the posterior part of the left superior temporal gyrus/sulcus as well as in the premotor cortex and inferior frontal gyrus. Stronger activity of the premotor and somatosensory cortices was observed during the observation of lingual compared to labial speech movements. Conversely, greater activation of the visual and auditory cortices was observed for labial movements. Altogether these results suggest that audio-visual-labial and audio-visuo-lingual speech perception recruit a common sensory-motor neural network and are partly driven by the listener’s knowledge of speech production.

Keywords: audio-visual speech perception, ultrasound, fMRI.

1. Introduction
Audio-visual speech perception is a special case of multisensory processing that interfaces with the linguistic system. In face-to-face interaction, visual cues from the speaker’s face can benefit the listener, notably by improving speech perception in noise or the understanding of a semantically complex statement or a foreign language (Sumby and Pollack, 1954; Reisberg et al., 1987; Navarra et al., 2005). Conversely, seeing incongruent articulatory gestures may also modify auditory speech perception (McGurk & McDonald, 1976).

At the brain level, audio-visual speech perception is known to rely on both primary and associative auditory and visual regions (Calvert et al., 1997; 2000). Because an enhancement of neural responses to audio-visual compared to unimodal speech inputs has been observed in the posterior part of the left superior temporal gyrus/sulcus, it has been proposed that the acoustic and visual speech signals are integrated in this multisensory region, and that modulation of activity within sensory-specific brain areas might partly be caused by backward projections and would represent the physiological correlates of the perceptual changes experienced after audio-visual speech integration (Calvert et al., 2000). In addition, audio-visual speech integration might partly be mediated not only by sensory-specific and multisensory brain regions but also by the speech motor system (including the posterior part of the inferior frontal gyrus and the adjacent ventral premotor cortex), with increased motor activity observed during audio-visual compared to unimodal auditory and visual speech perception (Skipper et al., 2005; 2007), as well as during audio-visual speech perception under adverse listening or viewing conditions (Callan et al., 2003; 2004).

From these studies, one unanswered issue is whether cross-modal speech interactions only depend on well-known auditory and visuo-facial modalities (in relation to labial movements of a speaker) or, rather, might also be triggered by less familiar visual modalities. In the present fMRI study, we examined the neural substrates of auditory, visual and audio-visual speech perception in relation to either labial or lingual movements (acquired with a camera and an ultrasound system, respectively). Since labial and lingual biological speech movements naturally exhibit temporal proximity with auditory speech inputs, evidence for cross-modal speech interactions in relation to both lip and tongue movements would strengthen the hypothesis that multisensory speech perception is partly driven by the listener’s knowledge of speech production (Skipper et al., 2007; Schwartz et al., 2012).

2. Methods
2.1. Participants
Twelve healthy adults, native French speakers, participated in the study. All participants were right-handed, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders.

2.2. Stimuli
Multiple utterances of /pa/, /ta/ and /ka/ syllables were individually recorded by one male and one female speakers in a sound-proof room. Synchronous recordings of auditory, visual and ultrasound signals were acquired by a Teracons T3000 ultrasound system (Hueber et al., 2008; see Figure 1) including a 140° microconvex transducer with 128 elements (tongue movements acquired with a sampling rate of 60 fps with a 640x480 pixel resolution), an industrial USB color camera (facial movements acquired with a sampling rate of 60 fps with a 640x480 pixel resolution) and an external microphone connected to the built-in soundcard of the T3000 ultrasound system (audio digitizing at 44.1 kHz).

Two clearly articulated /pa/, /ta/ and /ka/ tokens were selected per speaker (with the speaker initiating each utterance from a neutral mid-open mouth position). Sixty stimuli were created
consisting of twelve /pa/, /ta/ and /ka/ syllables related to five conditions: an auditory condition (A), two visual and two audio-visual conditions related to either lip or tongue movements of a speaker (Vt, Vf, AVt, AVf).

Figure 1: An example of lip (left) and tongue (right) visual stimuli.

2.3. Procedure
Before the fMRI session, participants were first presented with a subset of the recorded speech stimuli, with short explanations on the ultrasound system and on the tongue movements required for the production of /pa/, /ta/ and /ka/ syllables. Then participants underwent a three-alternative forced-choice identification task, being instructed to categorize as quickly as possible each perceived syllable with their right hand. The experiment consisted on 60 trials presented in a randomized sequence, with 12 trials related to each modality of presentation (A, Vt, Vf, AVt, AVf). The intertrial was of 3s and the response key designation was fully counterbalanced across participants.

The fMRI session consisted of one anatomical scan and one functional run. During the functional run, participants were instructed to passively listen to and/or watch speech stimuli presented in five different modalities (A, Vt, Vf, AVt, AVf). There were 144 trials, with an 8s intertrial, consisting of 24 trials for each modality of presentation and to a resting condition without any sensory stimulation.

2.4. Data acquisition
Magnetic resonance images were acquired with a 3T whole-body MR scanner (Philips Achieva TX). Participants were laid in the scanner with head movements minimized with a standard birdcage 32 channel head coil and foam cushions. Visual stimuli were presented using Presentation software (Neurobehavoiral Systems, Albany, USA) and displayed on a screen situated behind the scanner via a mirror placed above the subject’s eyes. Auditory stimuli were presented through the MR-confon audio system (www.mr-confon.de).

A high-resolution T1-weighted whole-brain structural image was acquired for each participant before the functional run (MP-RAGE, sagittal volume of 256x224x176mm³ with a 1mm isotropic resolution, inversion time = 900ms, two segments, segment repetition time = 2500ms, segment duration = 1795ms, TR/TE = 16/5 in ms with 35% partial echo, flip angle = 30°).

Functional images were obtained in a subsequent functional run using a T2*-weighted, echo-planar imaging (EPI) sequence with whole-brain coverage (TR = 8s, acquisition time = 3000ms, TE = 30ms, flip angle = 90°). Each functional scan comprised fifty-three axial slices parallel to the antero-posterior commissural plane acquired in a non-interleaved order (72x72 matrix; field of view: 216mm; 3x3mm² in plane resolution with a slice thickness of 3mm without gap). In order to reduce acoustic noise, a sparse sampling acquisition was used (Gracco et al., 2005). This acquisition technique is based on neurophysiological properties of the slowly rising hemodynamic response, which is estimated to occur with a 4–6s delay in case of speech perception and production (Grabinski et al., 2013). In the present study, functional scanning therefore occurred only during a fraction of the TR, alternating with silent interscanning periods, where stimuli were presented. The time interval between each stimulus onset and the midpoint of the following functional scan acquisition was set at 5s. All conditions were presented in a pseudorandom sequence. Altogether, 144 functional scans were therefore acquired ((5 perceptual conditions + 1 baseline) x 24 trials). In addition, three ‘dummy’ scans at the beginning of the functional run were added to allow for equilibration of the MRI signal and were removed from the analyses.

2.5. Data analyses

2.5.1. Behavioral analysis
For each participant and modality, the percentage of correct responses and mean reaction-times (RTs), from the onset of the acoustic syllables, were computed. For each dependent variable, a repeated-measures ANOVA was performed with the modality (A, Vt, Vf, AVt, AVf) as the within-subjects variable. For both analyses, the significance level was set at p = .05 and Greenhouse–Geisser corrected (for violation of the sphericity assumption) when appropriate. When required, posthoc analyses were conducted with Newman–Keuls tests.

2.5.2. fMRI analysis
fMRI data were analyzed using the SPM8 software package (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK) running on Matlab (Mathworks, Natick, MA, USA). Brain activated regions were labeled using the SPM Anatomy toolbox (Eickhoff et al., 2005) and, if a brain region was not assigned or not specified in the SPM Anatomy toolbox, using the Talairach Daemon software (Lancaster et al., 2000).

For each participant, the functional series were first realigned by estimating the six movement parameters of a rigid-body transformation in order to control for head movements between scans. After segmentation of the T1 structural image and coregistration to the mean functional image, all functional images were spatially normalized into standard stereotactic space of the Montreal Neurological Institute (MNI) using segmentation parameters of the T1 structural image. All functional images were then smoothed using a 6mm full-width at half maximum Gaussian kernel, in order to improve the signal-to-noise ratio and to compensate for the anatomical variability among individual brains.

For each participant, neural activations related to the perceptual conditions were analyzed using a General Linear Model, including five regressors of interest (A, Vt, Vf, AVt, AVf) and the six realignment parameters as covariates of no-interest, with the silent trials forming an implicit baseline. The BOLD response for each event was modeled using a single-bin finite impulse response (FIR) basis function spanning the time of acquisition (3s). Before estimation, a high-pass filtering with a cutoff period of 128s was applied. Beta weights associated with the modeled FIR responses were then computed to fit the observed BOLD signal time course in each voxel for each condition. Individual statistical maps were calculated for each perceptual condition with the related baseline and subsequently used for group statistics.
In order to draw population-based inferences, a second-level random effect group analysis was carried out with the modality (A, V_L, V_T, AV_L, AV_T) as the within-subjects variable and the subjects treated as a random factor. In order to determine common neural activity related to auditory, visual and audio-visual speech perception, in relation to lip and tongue movements, two conjunction analyses were separately performed (i.e., A ∩ V_L ∩ AV_L and A ∩ V_T ∩ AV_T). Then, an analysis by modality was conducted in order to determine which regions were more activated during lip compared to tongue movements, and vice-versa (i.e., V_L ≠ V_T and AV_L ≠ AV_T).

All contrasts were calculated with a significance level set at $p = .05$, family-wise-error (FWE) corrected at the voxel level with a cluster extent of at least 30 voxels.

### 3. Results

#### 3.1. Behavioral results

Overall, the mean proportion of correct responses was of 81%. The main effect of modality was significant ($F(4,44) = 38.1$, $p < .001$), with more correct responses in the A, AV_L, AV_T conditions than in the V_L condition, and in the V_L condition than in the V_T condition (on average, A: 99%, AV_L: 98%, AV_T: 94%, V_L: 69%, V_T: 47%).

For RTs, a significant effect of the modality was also observed ($F(4,44) = 18.2$, $p < .001$), with faster RTs in the AV_L condition than in the AV_T and V_L conditions, and in the AV_T and V_L conditions than in the V_T condition (on average, A: 837ms, AV_L: 732ms, AV_T: 926ms, V_L: 984ms, V_T: 1187ms).

### 3.2. fMRI results

#### 3.2.1. Conjunction analyses - see Figure 2

The conjunction analysis on A, V_L, and AV_L conditions demonstrates common activity in the posterior part of the superior temporal gyrus/sulcus (pSTG/STS), extending rostrally to the Heschl’s gyrus and insular cortex, ventrally to the posterior middle temporal gyrus (MTG) and dorsally to the parietal operculum and the ventral part of the supramarginal (SMG) and angular gyri (AG). Common neural responses were also observed in the prefrontal cortex, the inferior frontal gyrus (pars opercularis and right pars triangularis), the middle frontal gyrus and the left primary sensorimotor cortex. Additional activity was found in the cerebellum, the supplementary motor area (SMA) and adjacent anterior cingulate cortex, and the precuneus.

Similarly, the conjunction analysis on A, V_T, and AV_T conditions demonstrates common activity in pSTG/STS, extending ventrally to the left posterior MTG and dorsally to SMG, AG and the left parietal operculum. Common neural responses were also observed in the prefrontal cortex, the inferior frontal gyrus (pars opercularis and right pars triangularis), the middle frontal gyrus, the insular cortex and the left primary sensorimotor cortex. Additional activity was found in the cerebellum, the SMA and adjacent anterior cingulate cortex, the precuneus, and the associative extrastriate visual cortex.

#### 3.2.2. Analyses by modality - see Figure 3

V_L ≠ V_T: Compared to tongue movements, audio-visuo-labial speech perception induced a greater activation of the auditory
(including the Heschl’s gyrus, the temporopolar area, pSTG/STS, MTG) and visual cortices (from the primary visual cortex extending to the extrastriate visual cortex and to the dorsal part of the cerebellum). Greater activity of frontal regions was also evident (middle frontal and dorsolateral prefrontal cortices), especially in the right hemisphere. Conversely, the audio-visuo-lingual speech perception entailed greater activity in motor and premotor cortices as well as in parietal regions (including parts of the sensorimotor cortex, intraparietal sulcus, inferior and superior parietal cortices).

$AV_T \neq AV_f$: Compared to tongue movements, audio-visuo-labial speech perception induced a greater activation of the visual cortex (from the primary visual cortex extending to the right extrastriate visual cortex and to the dorsal part of the cerebellum). Conversely, the audio-visuo-lingual speech perception entailed greater activity in left frontal (prenotor and prefrontal cortices), parietal (including parts of the intraparietal sulcus and the surrounding inferior and superior parietal cortices) and auditory areas (pSTG/STS) as well as in the bilateral ventral cerebellum.

Figure 3: Activity differences between lip (L) and tongue (T) visual and audio-visual conditions.

4. Discussion

The present fMRI study examined the neural substrates of cross-modal binding during audio-visual speech perception in relation to both facial and tongue movements. Our results first demonstrate for both labial and lingual stimuli, common overlapping activity between modalities in the posterior part of the superior temporal gyrus/sulcus. These results appear line with previous studies indicating a key role of this region in biological motion perception (including face perception), speech processing and audio-visual integration (Calvert et al., 1997, 2000; Beauchamps et al., 2004). In addition, while more activity in visual and auditory cortices was observed during the perception of lip movements, more activity was observed in motor and somatosensory areas during the perception of tongue movements. This latter result likely indicates that participants simulated, covertly or overtly, the motor consequence of the perceived actions, particularly for less familiar visual information as in the case of lingual movements. Altogether, these results suggest that audio-visuo-labial and audio-visuo-lingual speech perception recruit a common sensory-motor neural network and are partly driven by the listener’s knowledge of speech production (Skipper et al., 2007; Schwartz et al., 2012).

5. References


Non-local duration differences caused by consonantal length contrasts
Giuseppina Turco, Bettina Braun
University of Konstanz, Germany
giuseppina.turco@uni-konstanz.de, bettina.braun@uni-konstanz.de

Abstract
Our study deals with the durational structure of Italian words with a medial geminate or singleton consonant (e.g., palla “ball” vs. pala “shovel”). Specifically, we investigated the duration of the word-initial consonant (e.g., [p]) and found that the onset consonant was longer in geminate than in singleton words. A comparison to German word pairs with a vowel length contrast in the first syllable (e.g., bitten “to ask” vs. bieten “to offer”) showed no such duration increase for the word-initial consonant [b]. On the basis of previous studies on long-distance anticipatory effects, we argue that the strength of the geminate articulation in Italian is already foreshadowed in the word-initial consonant, whereas the German vowel length contrast does not lead to such anticipatory duration differences. Our results hence provide evidence for a distinction between consonantal and vocalic subsystems for a different set of data.

Keywords: articulatory strengthening, non-local anticipation, length contrast, consonant, vowel, rhythm

1. Introduction
Word-initial consonantal strengthening involves a spatio-temporal increase of articulatory strength, resulting in a more forceful articulatory gesture (Cho, 2004; Fujimura, 1990; Keating, Cho, Fougeron, & Hsu, 2003). The amount of articulatory strengthening is affected by prosodic factors: if a given word starts a higher prosodic domain, the articulation of the initial consonant is stronger than when this word is domain-medial and starts a lower-level prosodic domain (e.g., Fougeron & Keating, 1997). The present study investigates another aspect of word-initial strengthening, one that is related to the rhythmic structure of the word itself. Specifically, we contrast words that contain either a geminate or singleton consonant in word-medial position (e.g., palla [palːa] “ball” vs. pala [palːa] “shovel”). Previous work has shown that the geminate-singleton contrast is signalled by the duration difference of the medial consonant as well as by the duration difference of the preceding vowel (i.e., [a] of palla is shorter than the [a] of pala, Pickett, Blumstein, & Burton, 1999, Esposito & Di Benedetto, 1999; see also Payne, 2005). In particular, Pickett et al. (1999) found that the duration ratio between medial consonant and preceding vowel discriminated minimal pairs with either a geminate or singleton consonant across different speaking rates. Similar findings were reported in Esposito & Di Benedetto (1999). Both studies suggest the presence of anticipatory ‘compensation’ (see also Lindblom & Rapp, 1973) in the realisation of the minimal pair difference in Italian: speakers appear to aim at maintaining the duration between adjacent syllables by balancing out the duration of the medial consonant in relation to the preceding vowel. In this paper we investigate whether the length of the word-medial consonant also results in more distant temporal adjustments in Italian, i.e., whether the rhythmic structure of the word also affects the word-initial consonant.

Recent research seems to suggest that certain sounds or sound contrast exert adjustments to the realization of sounds that are not immediately adjacent. For instance, anticipatory strengthening of word-initial consonants is also shown for English. Hawkins and Nguyen (2004), for instance, tested the influence of syllable-coda voicing on the spectral and durational properties of the onset [l] in (British) English CVC monosyllables (e.g., led vs. let). Speakers produced longer onset [l] consonants when they occurred in words with voiced codas (i.e., led) than in words with voiceless codas (i.e., let), mimicking the allophonic duration adjustment of the vowel (longer before voiced codas). These results show that a consonant can not only affect adjacent segments (see Farnetani & Recasens, 1993 for Italian) but has more far-reaching effects extending to syllable-onsets. Evidence on long-distance effects also comes from speech perception (e.g., Speeter Beddor, Hamsberger, & Lindemann, 2002). However, to date, our understanding of the rhythmic organization of syllables and words is not fully resolved.

On the basis of previous observations on compensatory rhythmic strategies in Italian and long-distance effects in other languages, we investigated whether the strength of the geminate articulation is already foreshadowed in the duration of the word-initial consonant, i.e., the [p] in palla should be longer than the [p] in pala (Experiment 1). Note that this articulatory strengthening is different from the phenomenon of lexical initial gemination (e.g., [p]unta vs. [p]unta, Romano, 2003) as well as from the widely discussed post-lexical initial gemination known as raddoppiamento sintattico (Nespore & Vogel, 1986; Payne, 2005), a process that lengthens initial consonants after words that end with a stressed vowel (e.g., virtù [dɨːrˈvutː] “different virtue”). To exclude the possibility that this duration adjustment is caused by durational differences in the adjacent vowel, we conducted a control experiment with German minimal pairs differing in vowel length in that position (Experiment 2). The German vowel length contrast is signalled by an increased duration of the vowel and, for most vowels, by a difference in vowel quality (such that the short vowel is more central than the long vowel, e.g., bitten [ˈbitn̩] “to ask” vs. bieten [ˈbiːtɛn] “to offer”, e.g., Wiese, 2000).

According to some articulatory theories, consonantal and vocalic gestures operate on distinct levels and constitute different subsystems (e.g., Fowler, 1983; Ohmann, 1966; Smith, 1993). Durational adjustments in the word-initial consonant are hence only expected in Italian (with the word-medial consonantal length contrast) and not in German (which only features a vocalic length contrast).

2. Experiment 1: Italian
In Experiment 1 we analysed the duration of the word-initial consonant in Italian disyllabic minimal pairs that contained a word-medial singleton or geminate.
2.1. Methods

2.1.1. Materials

We chose 24 trochaic, disyllabic minimal pairs with a geminate-singleton contrast in word-medial position. To optimize generalizability across consonant types, 8 started with plosives, 8 with fricatives, and 8 with nasals. The word pairs were matched for lexical frequency: Singleton words had an average frequency of 276 occurrences per million (SD=1229), geminate words 137 o.p.m (SD=494), p>.6, according to the LIP corpus (http://badip.uni-graz.at/, last accessed March 2014).

We also chose 96 fillers to hide the presence of the minimal pairs. All fillers were common Italian words with different lengths and stress patterns: 16 monosyllabic words and 80 polysyllabic words. Of these 80 polysyllabic filler items, half were trisyllabic (20 with geminates in different positions of the word) and half were four-syllabic (20 with geminates in different positions). To avoid the occurrence of silence before the target word (which would have made it difficult to measure the closure duration of plosives), the words were embedded in the carrier sentence la parola <target>, questo è quello che dico (“the word <target>, this is what I am saying”).

2.1.2. Participants

Ten Italian native speakers (6 female, average age=27.2 years) took part for a small fee. They were unaware of the purpose of the experiment. They originated from different parts of Italy: 6 from Northern Italy (Como, Genova, Pavia) and 4 from Central-Southern Italy (Pisa, Chieti, Calabria, Potenza). All of them had been living in Konstanz at the time of testing.

2.1.3. Procedure

The members of a minimal pair were interspersed with the 96 fillers (separated by at least 10 other words). The reading list started with two fillers to familiarize participants with the task. Geminate and singleton words were equally often presented in the first and the second half of the experiment.

Participants were recorded individually in the phonetic laboratory at the University of Konstanz (44.1kHz, 16 Bit). They were instructed to read each sentence aloud at normal speed. In case of hesitations or disfluencies they were asked to repeat the sentence at the end of the session. The whole recording lasted approximately 10 minutes.

The 480 target words were manually annotated at the segmental level (word-initial consonant: C1, following vowel: V1, word-medial consonant: C2, word-final vowel: V2) using broadband spectrograms (Boersma & Weenink, 2012). Closure duration for plosives was measured from the offset of the modal voicing of the previous vowel (i.e., [a] of parola, see Figure 1) to the onset of the burst or the onset of voiceless aspiration (in the 26 cases in which there was no burst). The duration of fricatives was determined by the friction noise. Spectral changes and auditory information guided the segmentation in more problematic cases of breathiness and aspiration before and after the friction noise. For nasals the segmentation was based on abrupt spectral changes.

2.2. Results

The average durations of the three segments are shown in Table 1. The right column indicates the p-value for the factor LENGTH CONDITION (singleton vs. geminate), which was calculated using linear mixed effects regression models (Baayen, 2008). As the length contrast is claimed to differ across Northern and Central-Southern dialects (Bertinetto & Loporcaro, 2005), the random-effects structure of the model included speaker ORIGIN (Northern Italy vs. Central-Southern Italy) and SOUND CLASS (plosives, fricatives, nasals), allowing for random intercepts and slopes (cf. Barr, Levy, Scheepers, & Tily, 2013; Cunnings, 2012). SPEAKER and ITEM were treated as nested factors under ORIGIN and SOUND CLASS respectively (Bates, 2010). P-values are derived by comparing a model with a certain factor to an identical model that lacks that particular factor, using the log-likelihood test as implemented in the anova()-function in R. Note that the results are comparable, or even slightly stronger, if we analyse normalized segment durations (i.e., dividing the duration of a segment by the duration of the preceding word).

Figure 1: Waveform and Spectrogram of the Italian words 'papa' - pope (top) and 'pappa' - baby food (bottom). C1 represents the initial consonant, V1 the following vowel, C2 the medial consonant, V2 the final vowel.

Table 1: Mean values, standard deviation, and p-values of the initial consonant (C1), the following vowel (V1), and the medial consonant (C2) in Italian singleton and geminate words.

<table>
<thead>
<tr>
<th>Mean values</th>
<th>Italian</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>singleton</td>
<td>geminate</td>
</tr>
<tr>
<td>C1 duration</td>
<td>102 ms</td>
<td>110 ms</td>
</tr>
<tr>
<td>V1 duration</td>
<td>192 ms</td>
<td>139 ms</td>
</tr>
<tr>
<td>C2 duration</td>
<td>72.9 ms</td>
<td>178 sec</td>
</tr>
</tbody>
</table>

2.3. Discussion

Results of Experiment 1 show longer durations of the word-initial consonant in geminate words compared to singleton words. Note that the duration increase is phonetic in nature and therefore not comparable to the initial gemination caused by raddoppiamento sinasticco (cf. McCrary, 2002; Payne, 2005), neither qualitatively nor quantitatively.
As expected, the vowel preceding the word-medial consonant also differed in duration (see Table 1). Therefore, the duration increase of the word-initial consonant may also be caused by the duration difference in the adjacent vowel and not by the more distant difference in the word-medial consonant. Unfortunately, it is impossible to tease these explanations apart in Italian (since the vowel before a geminate is always shorter). Therefore, we investigated a language with a similar durational difference in the vowel of the first syllable as in Italian, but without a following consonantal length contrast (German).

3. Experiment 2: German

In Experiment 2 we tested the duration of the word-initial consonant in German minimal pairs with a vowel length contrast.

3.1. Methods

3.1.1. Materials

The structure of the items was similar to the Italian items (disyllabic trochaic word pairs), but the minimal pairs differed in vowel length and not in consonantal length. We selected 15 word pairs that were matched as a group for lexical frequency: 75.6 o.p.m (SE=88.9) for words with a short vowel (according to the CELEX word form dictionary, cf. Baayen, Piepenbrock, & Gulikers, 1995) and 51.3 o.p.m (SE=66.4) for words with a long vowel (p<.04). As in Italian, the target words started with different sound classes (7 with a plosive, 5 with a fricative and three with a sonorant). We furthermore selected 96 filler items (48 monosyllabic, 48 trisyllabic), half of which contained short vowels, half long vowels.

3.1.2. Participants

Nine German speakers (6 female, average age=24 years) took part in the recording. All were from Baden-Württemberg and were unaware of the goal of the experiment.

3.1.3. Procedure

We constructed an experimental list with the same constraints as in Experiment 1. Participants were tested and recorded under the same conditions as in Experiment 1.

3.2. Results

The productions were annotated at the segmental level with the same criteria as for Italian. Average values and standard deviations, as well as the main effect of length condition are shown in Table 2. In contrast to the Italian data, in German there was no effect of LENGTH CONDITION on the duration of the word-initial consonant, neither in raw nor in normalized values (p=.5), i.e., the zero model did not improve when adding LENGTH CONDITION.

To corroborate the differential effect of LENGTH CONDITION on the raw productions of the first consonant in the two languages statistically, we calculated a combined model, adding LANGUAGE as a fixed factor. Results showed a significant interaction between LANGUAGE and LENGTH CONDITION (ß=-.011, SE=0.004, t=-2.52, p<.05), in addition to a main effect of LANGUAGE (ß=-0.038, SE=0.012, t=2.98, p<0.05, see Figure 2).

Furthermore, the vowel duration differences are significantly larger in German than in Italian: Results revealed a significant interaction between LANGUAGE and LENGTH CONDITION on vowel duration (ß=-0.020, SD=0.008, t=-2.59, p<.001).

3.3. Discussion

In Experiment 2, the duration of the initial consonant was unaffected by the length of the adjacent vowel, despite the fact that the durational difference in the adjacent vowel was even larger in German than in Italian. Therefore, we conclude that the duration increase in the first consonant in Italian is more likely due to the singleton-geminate contrast in the word-medial consonant and not to the duration difference in the adjacent vowel. This interpretation is in line with the assumption that vocalic gestures differ from consonantal ones (see Section 1).

4. Discussion and Conclusion

The Italian data show longer durations of the word-initial consonant in words that contain a word-medial geminate compared to singleton words. A control experiment with German words that contrasted in the length of the vowel in the initial syllable – and not in the length of the medial consonant – did not lead to any changes in the duration of the word-initial consonant. This difference across languages suggests that the initial strengthening found in Italian operates on the consonantal level only (e.g., Fowler, 1983; Öhmann, 1966; Smith, 1993). Since German does not have a length contrast in that position anymore, no such consonantal duration adjustment is found there. More generally speaking, our findings hence suggest that initial strengthening is not only caused by higher-order prosodic domains (Fougeron & Keating, 1997) but may also be due to the internal rhythmic structure of the words. An alternative explanation for the cross-linguistic difference between Italian and German may be rooted in language-specific patterns of anticipatory temporal adjustments (e.g., Speeter Beddor et al., 2002). To decide between these two explanations, it will be necessary to analyse more data from different languages.
Note that our experiments focused on the temporal domain of initial strengthening. However, articulatory strengthening is also manifested by other changes, such as increased lingualpalatal contact (Fougner & Keating, 1997) or the release RMS amplitude (cf. Ridouane, 2010). On the other hand, the increase in duration in the word-initial consonant that we reported here may be entirely rhythmic (and hence durational) in nature, with the aim of keeping the duration of adjacent syllables equal in duration. In that respect the observed word-initial strengthening may differ from the one induced by higher-level prosodic phrase breaks. We have to leave this issue for future research.

Irrespective of the source of the word-initial strengthening, we hypothesise that it serves to enhance the upcoming consonantal length contrast, similar to the perceptual relevance played by the ratio between the medial consonant and the preceding vowel (Pickett et al., 1999). To get a better insight into the effects of word-initial strengthening on word recognition, the next question will be whether Italian listeners can use these non-local fine phonetic differences to speed up lexical activation (Cho, McQueen, & Cox, 2007; Tagliapietra & McQueen, 2010). Our findings on non-local duration differences caused by word-medial consonantal length contrasts in Italian open interesting issues for models of speech production, in particular with respect to the nature of the representation of prosodic units (syllable, foot, word) during speech planning (Dell, 1986; Levelt, Roelofs, & Meyer, 1999; Wheeldon & Lahiri, 1997). The next question to address is whether it is only word-medial geminates or also heterosyllabic consonant clusters that lead to word-initial consonantal strengthening (e.g., panna vs. panda).

5. Acknowledgements

This research was supported by an AFF research grant of the University of Konstanz, awarded to the second author.

6. References


Soft ‘g’ in Turkish: Evidence for Sound Change in Progress
Özlem Ünal Logacev\textsuperscript{1}, Susanne Fuchs\textsuperscript{1}, Marzena Żygis\textsuperscript{1,2}

\textsuperscript{1}Zentrum für Allgemeine Sprachwissenschaft (ZAS), \textsuperscript{2}Humboldt University, Berlin
ozlemonal@anadolu.edu.tr, fuchs@zas.gwz-berlin.de, marzena@zas.gwz-berlin.de

Abstract

The aim of this study is to propose a phonologically and phonetically based explanation of soft ‘g’ sound in Turkish which shows different outputs in spoken language. Our lexicon based search reveals that soft ‘g’ has a consonantal status and is found most frequently in the intervocalic context. Our phonetic study, however, demonstrates that soft ‘g’ is reflected in a lengthening of the preceding vowel. This conclusion holds for syllable-final position occurring word medially.

Keywords: soft g, Turkish, phonetics and phonology, sound change, lexicon

1. Introduction

One of the biggest disagreements concerning the phonology and phonetics of Turkish is the status of soft ‘g’, which is represented as ‘ğ’ in the orthography of the Turkish alphabet. Table 1 provides an overview of the phonemic descriptions of this sound given by various scholars. Depending on the surrounding context, soft ‘g’ is claimed to be a stop, a fricative, an approximant, or a vowel. Such a description makes this sound the chameleon of Turkish phonetics and phonology, which can be almost everything except a nasal.

<table>
<thead>
<tr>
<th>Table 1. Summary table of the literature concerning soft g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref.</td>
</tr>
<tr>
<td>[1] IA</td>
</tr>
<tr>
<td>[2] UP</td>
</tr>
<tr>
<td>[3] L</td>
</tr>
<tr>
<td>[4] PZ</td>
</tr>
<tr>
<td>[5] UP</td>
</tr>
<tr>
<td>[6] L</td>
</tr>
<tr>
<td>[7] VF</td>
</tr>
<tr>
<td>[8] VF</td>
</tr>
<tr>
<td>[9] SV</td>
</tr>
<tr>
<td>[10] SV</td>
</tr>
<tr>
<td>[12] VA</td>
</tr>
<tr>
<td>[13] L</td>
</tr>
</tbody>
</table>

IA: Inaudible
PG: Palatal glide
L: Lengthened vowel
UP: Unpronounced
VF: Velar fricative
A: Approximant
SV: Semi-vowel
V\textsuperscript{h}: Back vowel
V\textsuperscript{g}: Rounded vowel
VA: Velar Approximant
NV: Narrowing of previous vowel

As Table 1 shows, some authors treat soft ‘g’ as a stable phoneme that does not vary according to the context (e.g. [2], [3], [7], [8]), while others describe some contextual variations in detail. Moreover, as noted in [1], [2], and [5] there is a dialectical influence (without specification of the dialect) where soft ‘g’ is pronounced as a velar fricative ‘ɣ’. All studies, except for very few [3, 11, 12, 13] are rather based on impressionistic observations, which may, at least in part, explain their huge variability. It is, however, also possible that soft ‘g’ has been undergoing a sound change and therefore has so many different faces.

The aim of this study is to investigate the phonological and phonetic status of soft ‘ğ’ in a systematic manner.

2. Grapheme-to-phoneme relation in Turkish

Turkish orthography is very handy for phoneticians and phonologists, because there is a clear grapheme to phoneme relation [14, p. 688-689]. This relation goes back to an orthography reform in 1928 in the course of which the Ottoman letters were changed to Latin letters such that single phonemes usually correspond to single letters [12]. Soft ‘g’ was assigned the letter ‘ğ’. In the Ottoman alphabet, soft ‘g’ had two different symbols for different vowel contexts. In the context of front vowels it was symbolized as ‘ğ’ and in back vowel context it was used. Redhouse [15, p. 39] compared the Arabic ‘غ’ with French ‘g’ (grassoe) and claimed that the Turkish version of this sound is more softened and practically disappearing from the pronunciation.

3. Phonology: The occurrence of soft ‘ğ’ in the lexicon

“Historical sound change is ubiquitous and often fossilized in spellings how words used to be pronounced” [16, p.1]. This may be particularly true of Turkish, due to its extraordinary grapheme to phoneme relation. The directness of this relation allows us to search for the contextual environment of soft ‘ğ’ using a Turkish lexicon [17]. This lexicon consists of 62593 words, 9336 of which contain soft ‘ğ’ (proper nouns, homophones, suffixed or compound words were not excluded from the lexicon.) The text-editor “Sublime” was used to extract all tokens containing soft ‘ğ’, and the distribution analysis was carried out with R (version 2.15) [19].

We expected soft ‘ğ’ to occur adjacent to vowels, since Turkish phonology does not allow consonant clusters, except in word-final position. Moreover, adjacent vowels (hiatus) are not allowed in Turkish except in loan words. Based on the lexicon search we found that:

1. Soft ‘ğ’ never occurs in the word-initial position.
2. Soft ‘ğ’ is always preceded by vowels (/i, u, a, õ, o, y, æ/) and followed either by vowels or by alveolar or bilabial sonorants (see Figure 1). When soft ‘ğ’ is followed by a sonorant consonant, both are separated by a syllable boundary (VC.CV).
3. Soft ‘ğ’ never occurs adjacent to voiceless obstruents with the exception of a few loan words or suffixes.
4. Soft ‘ğ’ is found most frequently in the context of high vowels.
5. There are very few minimal pairs with soft ‘g’ versus g in the intervocalic position and they are typically synonymous (e.g., ‘öge’ and ‘öge’ [item, element] can be used interchangeably).

6. In almost all minimal pairs soft ‘g’ is followed by a sonorant consonant, e.g. ‘ari’ vs. ‘ağrı’ (‘bec’ vs. ‘pain’). There are very few exceptions such as ‘adali’ and ‘ağdali’ (‘islander’ vs. ‘pompous’).

In summary, our grapheme-based search has revealed that soft ‘g’ is mainly found in the intervocalic position. More specifically, it is most frequently found in the high vowel context. With respect to this, Turkish follows cross-linguistic patterns according to which high vowels ‘soften’ or palatalise adjacent consonants (see e.g. Bhat [25]). However, since the soft ‘g’ occurs in other context as well, even though these are considerably less frequent, it seems to be premature to call soft ‘g’ an allophonic variant of /g/ in this particular context. In addition there are still relatively many words where /g/ is found in the intervocalic context in general and in the high vowel context in particular.

In the light of the fact that the results are based on a lexicon search and that previous research is inconclusive regarding the phonetic realisation of this sound, we conducted acoustic experiments with the aim of examining acoustic realisation of soft ‘g’ in different contexts.

4. Experimental evidence

The experimental part of the study addresses three questions that are grounded in claims proposed in the literature:

1. Is soft ‘g’ a consonant or a lengthened vowel?
2. Is soft ‘g’ realized differently when adjacent to different vowels?
3. Did soft ‘g’ undergo a diachronic change?

In order to answer these questions, we conducted an acoustic experiment in which we focused on minimal pairs with and without soft ‘g’. We addressed the question about diachronic change by comparing the phonetic realisations of soft ‘g’ by young and old people. Following this line of reasoning we expected that old people might show more traces of soft ‘g’ as a consonant than young people would do, if soft ‘g’ did indeed undergo a change.

4.1. Participants

The experiment was run with 16 young (age range: 19-29; 9 females and 7 males) and 8 elderly participants (age range: 72-91; 4 female, 4 male). All participants were native speakers of Standard Turkish without any known speech, language or hearing disorders.

4.2 Speech Stimuli and Procedure

The stimuli of this experiment consisted of 22 minimal pairs without soft ‘g’ (Condition A) and with soft ‘g’ (Condition B).

A. Elif _değmek sözcüğü onudu_ (‘Elif read the word say‘)
B. Elif _değmek sözcüğü onudu_ (‘Elif read the word touch‘)

Minimal pairs with and without soft ‘g’ are only available for word-internal, syllable-final position. It is impossible to have minimal pairs with and without soft ‘g’ intervocically because Turkish phonology does not allow for two vowels to be adjacent to each other. Because of all these phonotactic restrictions we focused only on the word-internal, syllable-final position.

Target words were embedded in sentences to avoid prosodic influences that can take place at the beginning or end of a sentence.

Participants were requested to read aloud five lists composed of stimuli with different randomizations. In each list, experimental sentences were mixed with 22 filler sentences in order to prevent participants from noticing the aim of the
experiment and thus possibly emphasizing soft ‘g’. Participants read 44 sentences per list. They were instructed to read the sentences at a normal speech rate. For elderly participants the number of fillers was decreased a half to avoid fatigue effect.

4.3 Instrumental Measurement and Data Analyses

All acoustic analyses were carried out using Praat (version 5.3.53). The on- and offsets of the whole target word were labeled. Visual inspection of the spectral properties showed that soft ‘g’ could not be separated from the preceding vowel so that the whole segment consisting of soft ‘g’ and the preceding vowel was labeled.

Based on these labels, the duration of the segment, as well as average F1, F2 and F3 values over the course of the vowels were computed and analyzed.

We compared the values of F1, F2, and F3 at 25% and 75% of the sound duration in condition B. To test differences in formant values at different time points we used linear mixed-effects models [21, 22, 23] using the lme4 package [24] in R [19]. We included condition, age, gender, and vowels as fixed effects as well as a maximal random effects structure without correlation parameters.

4.4 Results

Mean durations are presented in Figure 4. Segment durations were two times longer in condition B than in condition A ($\beta = 72$ms, SE=$6.5$ms, $t = 11.1$). Durations of the segments produced by the younger group are shorter in comparison to those pronounced by the elderly people ($\beta = 14$ms, SE=$4.6$ms, $t = 3.08$) for both conditions. Another result is that the /u/ is shorter than /e/ and /a/ in all groups and conditions ($\beta = 33$ms, SE=$12.4$ms, $t = -2.7$).

Elderly female participants had slightly lower formants and a narrower formant range in F1 than young female participants. Both formants F1 and F2 were more peripheral for condition B than condition A in both age groups.

If soft ‘g’ really exists phonetically there should see a trace of it, at least towards - its end in condition B. So we compared the difference between formants measured at the 75% and 25% time points, which is presented in Table 3. According to these results F1 rises towards the end for /u/ but it remains flat...
for /a/ and /e/. There is also a gender effect on F2 rise in /e/ and F3 which is rising for /e/, but falling for /u/.

Table 3. Linear mixed models coefficients, Standard errors (SE) and t-values of difference between 75% and 25%-time point’s formants in condition B.

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-33 (11)</td>
<td>130 (59)</td>
<td>83 (41)</td>
</tr>
<tr>
<td>Age: young-old</td>
<td>3 (8)</td>
<td>0.4</td>
<td>15 (31)</td>
</tr>
<tr>
<td>Gender: male-fem</td>
<td>8 (7)</td>
<td>-1.1</td>
<td>-58 (29)</td>
</tr>
<tr>
<td>Vowel: /a/-/e/</td>
<td>49 (19)</td>
<td>2.6</td>
<td>-147 (70)</td>
</tr>
<tr>
<td></td>
<td>-15 (13)</td>
<td>-1.2</td>
<td>-70 (42)</td>
</tr>
</tbody>
</table>

5. Conclusions

Soft ‘g’ in Turkish is a challenging research topic for both phonologists and phoneticians. Its chameleon-like phonological behavior is accompanied by several exceptions that allow us only to draw a very general conclusion that soft ‘g’ behaves phonologically like a consonant and to some extent as an allophonic variant of ‘g’. It is most frequently found in the context of (high) vowels. Phonetically speaking, soft ‘g’ is realized as a long vowel. Our experimental evidence reveals that the only trace of the soft ‘g’ in the context of the following glide is a lengthening of the preceding vowel. Such processes are also cross-linguistically found where deletion of the consonants causes lengthening of the vowel. For instance, in German orthographically represented <i> is not pronounced but the preceding vowel is lengthened: Kohl ‘cabbage’ is pronounced as [ko:l].

Finally, our results support a statement provided in [26] according to which soft ‘g’ still has a phonological status in Turkish. Phonological rules apply to word-final soft ‘g’ in the way they would do in the case of a consonant-final word, cf. /dاغ-ı/ [da:ı] ‘mountain’ vs. /dاغ-l/ [da:l] ‘branch’ [26, p.1381]. In addition, as reported in [27] soft ‘g’ causes lengthening of the vowel in the word-final position (e.g., /dاغ-ı/ [da:] ‘mountain’). Our data expand the context of vowel lengthening by including soft ‘g’ in word-medial syllable-final position if it is followed by a sonorant.

Acknowledgements

We would like to thank all our informants for participating in the experiment, and Pavel Logachev for programming assistance. We also thank the maintainer of http://markantoniniou.blogspot.de/search/label/label/scripts for the invaluable Praat scripts. This research has been supported by Bundesministerium für Bildung und Forschung (BMBF, Germany) Grant Nr. 01UG1411 to Susanne Fuchs and Marzena Żygis.

References


Cortical inhibition of the brainstem frequency following response in falling and rising tone production

Eryk Walczak, Jingwei Zhang, Tim Schoof, Jyrki Tuomainen and Peter Howell

Division of Psychology and Language Sciences, University College London, UK
eryk.walczak.11@ucl.ac.uk, jingwei.zhang.12@alumni.ucl.ac.uk, t.schoof@ucl.ac.uk, j.tuomainen@ucl.ac.uk, p.howell@ucl.ac.uk

Abstract

A shadowing task was used to determine influences of efferent control on tone falls and rises. Cortical activity was measured, and estimates of the fidelity of pitch in the brainstem frequency following response were obtained. N1 and P2 ERP amplitude during falls and rises were used as measures of cortical activity and these were correlated with brainstem measures for three main tasks (perception, production-shadowing and tapping) separately for falling and rising stimuli. Some tentative evidence was obtained that showed that falls were mainly under early cortical control (N1) and rises under late cortical control (P2). There was also some evidence to suggest that early cortical responses enhanced pitch representation at the brainstem directly. The later cortical response enhanced production indirectly (by attenuating use of a mechanism that reduced error in the perception condition).

Keywords: auditory cortex, cortical inhibition, cortical pitch response, frequency-following response, fundamental frequency, pitch processing, shadowing

1. Introduction

Cortical EEG responses in speech tasks show an early negative component (N1) and a later positive component (P2). The amplitudes of these components provide some information about activity in different tasks. Production of falling versus rising tones are tasks where the brain mechanisms that control the laryngeal muscles may differentially weight on N1 and P2 components. A falling tone needs early control of laryngeal muscle tension that can then be passively released whereas a rising tone needs laryngeal tension to be controlled over the whole length of the syllable carrying the tone.

As well as involvement of activity at different times, cortical activity on rising and falling tones may also differ between production and perception, and may reflect efferent activity directed at brainstem mechanisms. It is known that self-vocalization attenuates auditory cortical responses (Agniew et al. 2013; Chen et al. 2013; Curio et al. 2000; Numminen and Curio 1999; Paus et al. 1996). There is one study which interpreted its results as showing how cortical activity influenced subcortical auditory processing (Papanicolau et al. 1986). Papanicolau et al. reported that peak V in the auditory brainstem response (ABR, most likely generated in the lateral lemniscus or inferior colliculus) to tone clicks was suppressed when speakers vocalized concurrently (normal or whispered speech). They attributed this to efferent inhibition by cortical systems. Since that date, more attention has been paid in ABR work to the sustained brainstem response that represents the low frequency periodic components of acoustic stimuli, called the frequency following response (FFR). The FFR has been reported to be attenuated in some cross-modality studies (mainly on animals) in auditory-visual tasks (Hairston et al. 2013; Oatman et al. 1980). The FFR to tonal stimuli that are novel to participants has also been reported to change after a short period of training (Song et al. 2008). Related to this, Krishnan and Gandour (2009) reported that long-term experience using a tonal language changes the representation of pitch in the FFR. Krishnan et al. (2012) have recently started examining cortical activity in tone perception and found that it correlates with brainstem activity. The majority of studies on tone processing have focused on perception. Cortical influences on the FFR in speech production have not been examined to date.

Therefore, we examined rising and falling tones in a production task (shadowing) as well as in perception, and tapping control tasks. The same stimuli were heard in all tasks. For example, shadowing involved the speaker listening to this signal and to imitate it. Shadowing a stimulus has been used in speech training, so Krishnan and Gandour’s (2009) results would lead us to expect this task to influence the FFR. We hypothesized that the speaker could use the shadowed signal to improve performance by either enhancing perception (EnhPer, obtain an accurate idea of what is tracked) or enhancing production (EnhPro, obtain an accurate idea of own production).

The principal topics for investigation were: 1) whether cortical activity differed across perception, production and tapping tasks; and 2) whether any differences were associated with changes in the fidelity of FFR measures that reflect how accurately pitch is represented. Concerning 1, both EnhPer and EnhPro predict that there should be differences in cortical activity between the tasks. In addition, it was expected that the cortical component involved would depend on tone (fall early and N1, rise late and P2). Concerning 2, EnhPer and EnhPro predict that correlations between cortical and brainstem activity would occur either in perception or in production tasks respectively. These represent direct effects. In addition, the shadowing condition allows indirect effects to be examined. The shadowing condition has the same stimulus presented that occurs in the perception condition. If there is EnhPer in the perception condition, but not the production condition, this suggests that the perception effect has been suppressed in shadowing. In this case an effect is established indirectly by comparison across tasks (perception and shadowing) to determine how perception changed during speech production. In the study, the cortical-brainstem correlation patterns were examined to see whether EnhPer or EnhPro occur and whether there are direct or indirect effects. As with the cortical analysis, the correlation patterns that indicate which of these mechanisms operate were assessed separately for N1 and P2 responses since these components might control falls and rises respectively.
2. Method

2.1. Participants

Four male and four female right-handed Mandarin speakers aged 20-24 years participated in this study, all of whom had received training on a musical instrument (musical experience might improve pitch-processing ability). Standard brainstem responses to click stimuli were assessed prior to the experiment and were normal. Also, all participants had normal hearing for octave frequencies between 250 and 8000 Hz.

2.2. Procedure

The task conditions of perception, production-shadowing and tapping were performed in random order. Participants heard a 170 ms /da/ with falling or rising tone at 75dB SPL in all conditions. The stimuli were presented at a 1s rate. The /da/ was the standard stimulus used in FFR work (Russo et al., 2004) and its pitch was modified using PRAAT. This was listened to in the perception condition. The same syllable/tone was heard and produced concurrently by the participant in the shadowing condition. Participants heard the target syllable and tone in the tapping condition and tapped bi-manually in time on a hard surface. Each condition was performed twice, once when a /da/ with a falling pitch was heard and once when a /da/ with a rising pitch was heard. The perception condition provided a baseline against which inhibition (decrease) or facilitations (increase) in the production-shadowing condition was assessed. Similarly, tapping provided a non-speech motor response baseline against which inhibition or facilitation in the production-shadowing condition was assessed.

Three PCs were used. The first delivered the stimuli and sent triggers that indicated when the stimulus had been presented to a second PC. This second PC recorded the EEG responses using ActiView (Biosemi’s acquisition software) and the trigger was recorded on an unused channel. The third PC used CoolEdit to record audio responses in the production-shadowing and tapping conditions. Biosemi software was used to record from scalp electrodes and the signals were then used to obtain concurrent ERPs and ABRs.

3. Results

3.1. Brainstem FFR analysis and ERP pre-processing.

ABRs were obtained using two electrodes – active (Cz) and reference (EXG2 on the seventh cervical vertebral). The data from these electrodes were collected as with the electrodes used for ERP. As ABRs and the Frequency-Following Response (FFR) in particular require higher sampling rates, the data were only down-sampled to 16384 kHz (one of the standard setting on the AD board).

ABR responses were extracted using the Brainstem Toolbox MATLAB scripts (Skoe and Kraus 2010) and EEGLab. The measures chosen to represent fidelity of pitch representation were pitch error, which is the absolute distance in Hz that the response pitch deviates from the stimulus pitch on average across the duration of the stimulus and F0 correlation which is Pearson’s r computed between the stimulus-track and the response-track. The averaged FFR response for rise and fall tones in the perception and production tasks are shown for one participant in Figure 1.

There was some noise in the ERP data which had not been removed by standard cut-off procedures that were employed during data acquisition so ICA was used to clean up the recordings by removing artifactual components (Delorme, Sejnowski, and Makeig 2007). The ICA was based on Repovs forum post on EEGLabList in 2010. Thirteen channels were used for ICA, including three external channels – eye (mounted on the right side), left and right mastoid. The remaining ten cortical channels were: Fp1, F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4. The Fp1 electrode was used to control for eye movements on the left eye (it was the closest electrode to the left eye). The remaining nine channels were selected as they were at locations that were good for measuring N1 and P2 evoked potentials. Each file’s sampling frequency was down-sampled to 512 Hz for ERP analysis.
3.2. Cortical ERP and brainstem ABR analysis

N1 and P2 were analysed separately, firstly for cortical influences and then for cortical-brainstem correlations. The cortical data were analyzed by repeated-measures ANOVA with electrode (nine levels), tone (two levels) and tasks (three levels) as factors. For N1, the production task showed a slight peak for most electrodes for falling tones, and a more marked dip (suppression) for the rising tones compared to the perception and tapping conditions. The analysis of N1 showed that task was not significant, but produced near significant interactions with electrode (two way, F(16, 112)=1.510, p=.108) and these factors with tone (three way, F(16, 112)=1.577, p=.087). We examined N1-brainstem correlations next. There were near significant correlations for perception fall with F0 correlation for electrodes P3 and Fz (r=-.705, p=.051 and r=-.637, p=.089) and for production fall for pitch error (r=-.687, p=.060). The two reported effects for perception falls both involved negative correlations with F0 correlation meaning that the higher the N1 amplitude, the lower the correlation (showing perception is suppressed by N1 activity). Production fall involved a negative correlation with pitch error, this time meaning that the higher the N1 amplitude the lower the error (showing production is facilitated by N1 activity).

For P2, a similar ANOVA to that conducted for N1 cortical activity showed there was a near significant effect of task (F(1.077, 7.537)=3.147, p=.115) corrected for sphericity and F(2, 14)=3.147, p=.074 when not corrected. For the cortical P2-brainstem correlations, there were several significant and near significant correlations shown in Table 1. No other correlations were significant. The most notable features are that these all involved perception and only for rising tones, and this applied to nearly all the electrodes which were recorded from. The signs of the correlations show higher N1 activity reduced error and improved F0 tracking.

### Table 1: Cortical P2-brainstem correlations in perception (rising tone).

<table>
<thead>
<tr>
<th>Electrode</th>
<th>Pitch error</th>
<th>F0 correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>F3</td>
<td>r=-.675, p=.066</td>
<td>r=.697, p=.055</td>
</tr>
<tr>
<td>C3</td>
<td>r=-.863, p=.006</td>
<td>r=.854, p=.007</td>
</tr>
<tr>
<td>P3</td>
<td>r=-.806, p=.016</td>
<td>r=.658, p=.076</td>
</tr>
<tr>
<td>Fz</td>
<td>r=-.776, p=.024</td>
<td>r=.752, p=.031</td>
</tr>
<tr>
<td>Cz</td>
<td>r=-.704, p=.051</td>
<td>r=.717, p=.045</td>
</tr>
<tr>
<td>Pz</td>
<td>r=-.709, p=.026</td>
<td>r=.818, p=.013</td>
</tr>
<tr>
<td>F4</td>
<td>r=-.775, p=.024</td>
<td>r=.778, p=.023</td>
</tr>
<tr>
<td>C4</td>
<td>r=-.674, p=.067</td>
<td>r=.704, p=.051</td>
</tr>
<tr>
<td>P4</td>
<td>r=-.709, p=.026</td>
<td>r=.818, p=.013</td>
</tr>
</tbody>
</table>

4. Discussion and Conclusion

The study allows some tentative conclusions. Firstly, as hypothesized, N1 was involved with falls and P2 with rises. Secondly, the cortical-brainstem correlations were performed to determine whether the speaker could use the shadowed signal to improve performance by either enhancing perception (EnhPer, obtain an accurate idea of what is tracked) or enhancing production (EnhPro, obtain an accurate idea of what is produced).

Looking firstly at the N1 component, that appears to be involved with falls, in its interaction with brainstem measures, a mix of perception and production effects occurred. The perception effects showed deleterious influence of high N1 activity on fidelity of the brainstem FFR representation whereas production showed beneficial effects when N1 activity was high (i.e. lower error). It is not clear how higher N1 activity increased error of the FFR response in perception. In production, inhibition of the stimulus at the brainstem would possibly be a useful thing in shadowing (and maybe other tasks) for speech production as it would provide a mechanism that allows own voice to be heard (suppression of perception). This would show that the early facilitatory cortical effect (N1) helps to keep FFR error low in production falls (i.e. a direct effect). A caution is that this was specific to particular electrodes.

The late cortical response (P2) show that a perceptual mechanism that is useful for listening to rises (reduces error) is not employed or suppressed in production-shadowing; for the whole electrode montage there was some evidence that higher P2 activity reduced brainstem error and improved F0-correlations. This could indicate that in shadowing the speaker has suppressed the perception-enhancement mechanism to enhance production which would allow him or her to focus on own voice control, rather than maintain improved fidelity of the stimulus being tracked. This interpretation would favor EnhPro for rises in this later cortical component and is an indirect effect involving comparison across production and perception conditions.

Overall the results provide some evidence to suggest that early cortical activity is associated with fall control and later activity with rise control. Fall control may have a direct effect on the brainstem FFR signal’s fidelity. Rise control has an indirect effect on production control; the mechanism that is useful in perception is switched off so that the speaker can focus on his or her own voice control.

The current work employed a shadowing task to investigate production. The responses required in shadowing tasks are following responses. This contrasts with the response in other commonly-employed tasks used to examine speech production in which feedback of the speaker’s own voice is altered. With some of these alterations, speakers’ responses are compensatory. The question arises as to how these opposing responses in shadowing and in altered feedback emerge, whether they implicate different EnhPer and EnhPro mechanisms, and whether these mechanisms influence speech control directly or indirectly.

5. Correspondence

Correspondence concerning this paper should be sent to Peter Howell (p.howell@ucl.ac.uk).
6. References


On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French

Xiaou Wang*, Thomas Hueber, Pierre Badin

GIPSA-Lab / DPC, UMR 5216, CNRS – Grenoble Alpes University, France

rosingle871113@gmail.com, (thomas.hueber, pierre.badin)@gipsa-lab.grenoble-inp.fr

Abstract

This study investigates the usefulness of a French Articulatory Talking Head (ATH) to train Mandarin Chinese speakers to produce French vowels that do not exist in their phonetic repertoire. Two groups of Mandarin Chinese speakers were trained using either auditory or audiovisual stimuli displayed with the ATH, based on a speech shadowing task of immediate repetition of vowels and VCV sequences. We found that the F1 and F2 formants of vowels improved for both groups and that the audiovisual group (AVG) improved more than the auditory group (AG), which tends to validate the interest of the vision of articulators for the pronunciation training.

Keywords: articulatory talking head (ATH), computer aided pronunciation training (CAPT), second language acquisition, Chinese Mandarin, French

1. Introduction

Human begins to use articulatory information in speech perception at a very young age. Kuhl & Meltzoff (1982) have found that 18- to 20-week-old infants recognize already the correspondence between auditorily and visually presented sounds. This behavior will persist into adulthood, as shown by the McGurk Effect (McGurk & MacDonald, 1976). The use of such information in second language pronunciation training could thus be beneficial and has already started to spread. Traditional studies tended to show learners their own articulation with different instruments for pronunciation training, including ElectroPalatoGraphy (Schmidt & Beamer, 1998) or ultrasound (Bressmann et al., 2005), while other studies used what we call here an “articulatory talking head” (ATH) capable of displaying the movements of both external and internal articulators (such as the tongue) (Massaro et al., 2008; Engwall, 2012). The present study adopted the second approach.

Although the overall contribution of ATH in pronunciation training has already been investigated (cf. e.g. Massaro & Light, 2003), the benefit of articulatory information compared with traditional trainings in which only auditory instructions are available has not so far received such attention. For this purpose we recruited two groups of Chinese students learning French to participate respectively in an auditory training (Auditory Group, AG) and an audiovisual training (AudioVisual Group, AVG). Due to the influence of the first language, Chinese learners typically have difficulty pronouncing correctly /ε o ɔ ø œ/ and the consonants in the VCV sequences were /p t k f s/. Since the durations of the original sequences were too short, we used a time stretching algorithm (based on Harmonic plus Noise modeling (Stylianou, 1996)) to artificially increase the vowel length so that the learners had enough time to observe the whole movement. In order to preserve natural coarticulation patterns, time stretching was applied only to stable parts of the vowels. The training with single vowels was organized into five blocks of three drills. Each block focused on one of the five vowels that the subjects should learn. In each block, the three drills presented successively a reference vowel and the vowel to be learned. At the end of the drill, the images corresponding to the stable state of the vowels were retained on the screen for 2 seconds, allowing the learners to better understand their articulation by comparison. Figure 1 illustrates this method with /œ/ and /ø/.

* At present: Laboratoire de Phonétique et Phonologie, Paris, France.
The training with VCV sequences was organized into three blocks of five drills. In each block we provided an intensive training for two confusing vowels combined with consonants /p t k f s/ to form five contrastive drills. This procedure was also designed with an eye to training coarticulation. Table 2 lists the corresponding drills and blocks.

Table 2: List of drills for each block (VCV sequences)

<table>
<thead>
<tr>
<th>Block</th>
<th>Drill</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i/ ~ /ɛ/</td>
<td>/i/ vs. /ɛ/, and so on</td>
</tr>
<tr>
<td>/o/ ~ /ɔ/</td>
<td>/o/ vs. /ɔ/, and so on</td>
</tr>
<tr>
<td>/œ/ ~ /ø/</td>
<td>/œ/ vs. /ø/, and so on</td>
</tr>
</tbody>
</table>

2.4. Pronunciation test corpus

Each learner was required to read a speech corpus before and after the training so that the changes of her pronunciation could be assessed in later analysis. The corpus contained all French oral vowels except /j/. 2 isolated productions and 5 VCV sequences were recorded for /a i u e/, while 9 VCV sequences and 5 words were recorded for the other vowels. The detail is presented in Table 3.

Table 3: Speech corpus used for pronunciation tests

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Syllable</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a i u e/</td>
<td>Two isolated productions + VCV sequences with /p t k f s/ as consonants, no words</td>
<td></td>
</tr>
<tr>
<td>/e/</td>
<td>Two isolated productions + VCV sequences with /p t k f s/ as consonants</td>
<td></td>
</tr>
<tr>
<td>/ɛ/</td>
<td>Pépé, mêmé, café, cassser, télé</td>
<td></td>
</tr>
<tr>
<td>/o/</td>
<td>Peau, mot, faux, moreau, tôt</td>
<td></td>
</tr>
<tr>
<td>/ɔ/</td>
<td>Poré, molle, folle, sort, sort</td>
<td></td>
</tr>
<tr>
<td>/œ/</td>
<td>Peur, fumeur, fleur, seul, autre</td>
<td></td>
</tr>
<tr>
<td>/ø/</td>
<td>Deux</td>
<td></td>
</tr>
</tbody>
</table>

2.5. Speech perception test corpus

A perception test (forced choice identification) was also presented before and after the training. The corpus contained one isolated vowel production and three VCV sequences with /t s l/ as consonants for /ɛ ɔ ø œ/. Each stimulus was played once. The results of this test could inform us of the learners’ discriminative capacities.

2.6. Protocol

The whole experiment was performed using the Presentation® software (Version 16.5, www.neurobs.com). During the experiment the operator could monitor the learner’s performance via a screen located outside the soundproof room where the learner was seated and could communicate with the latter by an intercom system. Before the experiment started, each learner was required to listen to an isolated production of /l/ to allow her to adjust the earphone volume to a comfortable level.

The experiment was divided into 6 phases. The learners were at first given two minutes to get accustomed with the phonetic symbols of each vowel (phase 1). Since the IPA was absent for some learners, we used /E O ou ou u/ to represent /ɛ ɔ ø œ/ y. They heard successively a word, an isolated vowel production and two VCV sequences using /t l/ as consonants. They were then submitted to the pronunciation test (phase 2). Next, they performed the perception test during which they gave their responses by clicking one of the six buttons representing /ɛ ɔ ø œ / after hearing a stimulus (phase 3). The stimuli in the perception test were randomized using the default algorithm in Presentation®.

Phase 4 was intended for training. At first the learners watched a three-minute video. At the beginning, some animations were displayed in order to get the AVG learners familiar with the ATH whereas the AG learners simply listened. Then the video explained the speech shadowing task (cf. Shockley et al., 2004) demonstrated by a French speaker. This task required each subject to repeat the stimulus as soon as they perceived it. The underlying idea was that it would help them imitate better the stimuli, as suggested by Shockley et al. (2004). The learners were then given one minute to practice the speech shadowing, using /i~/y/ and /u~/y/ as drills.

The training per se started with single vowels training. The learners saw and heard each drill twice. The first time they were requested to observe the tongue movement and the second time the tongue and lips movement. Then they saw the animations on the screen and started to imitate each animation using speech shadowing. The animations were also presented two times. The training with VCV sequences followed the same procedure, except that the learners saw and heard each drill only once. These two training periods lasted 15 minutes. After the training, the learners passed again the pronunciation test (phase 5) and the perception test (phase 6).

The experiment for the AG learners followed exactly the same procedure, except that the animations of the ATH were not provided. The whole experiment never exceeded 40 minutes.

2.7. Data collection and acoustic measurements

For each group, the 672 tokens of collected /a i u e/ consisted of 12 tokens each produced by 7 speakers before and after training (4×12×7×2) and the 1750 tokens of collected /ɛ ɔ ø œ/ consisted of 25 tokens each produced by 7 speakers before and after training (5×25×7×2).
The audio data were recorded in a soundproof room at a 44,100 Hz sampling rate with 16-bit resolution and later down sampled to 22,050 Hz before acoustic analyses. Vowels were segmented by means of the Praat software (Boersma & Weenink (2005)). We used the second formant as principal cue for the segmentation of the vowel in VCV sequences. Formant frequency analysis was performed using the LPC (autocorrelation) algorithms available in Praat with default settings for females. A rapid overview of the data revealed great variations among learners for the pronunciation of /e o ø/. Some diphthongized, while the others pronounced differently, probably due to their knowledge of American English. The pronunciation of /ø/ and /œ/ revealed a similar pattern, though not completely homogenous, as we shall see later. This observation led us to the following decision. Formants of the vowels other than /ø/ and /œ/ were extracted automatically, while formants of /e/ and /æ/ were extracted manually in the stable part of each vowel production. Some accidental pronunciation mistakes (/y/ for /e/) were also discarded during this process so that 170 tokens of /ø/ and /œ/ were used for acoustic analysis. For the other vowels, all the collected tokens were used.

3. Results

3.1. Overall improvement

Since the pronunciation of /e o ø/ varied considerably in each group, we decided to limit the analysis for the moment to /ø/ and /æ/. To visualize the overall change of all the learners, formants measured on vowels before and after training were displayed in the F1/F2 space, as shown in Figure 2 and Figure 3.

Figure 2: Centroids of F1-F2 for French vowels produced by Chinese learners before training (large symbols) averaged over both groups. The reference values (Calliope, 1989) are marked by small symbols.

Progress in each group

The mean frequencies of the F1 and F2 vowel formants before and after training were computed for each group. A paired-samples t-test was conducted to quantify the formant change before and after training, while an independent-samples t-test was conducted to quantify the difference of the formant change between two groups. To better understand the changes in each group, the statistical distributions of the F1 and F2 measured on all learners were displayed as boxplots, as shown in Figure 4 and Figure 5.

Figure 3: Centroids of F1-F2 for French vowels produced by Chinese learners after training (large symbols) averaged over both groups. The reference values (Calliope, 1989) are marked by small symbols.

Figure 4: Distributions of F1 for /ø/ and /œ/ before and after training. For each vowel the first bar represents the state before training and the second one the state after training (* denotes differences between before after training that are significant at p<0.05 and *** at p<0.001). The green squares represent the values from Calliope (1989).

Figure 5: Distributions of F2 for /ø/ and /œ/ before and after training. Same conventions as in Figure 4.
It can be seen that the overall differentiation of these two vowels was mainly due to the F1 shift of the two vowels in both groups and the F2 increase for both vowels in AVG. The changes of the means of F1 and F2 before and after training for each vowel in each group led to the first conclusion: all formants moved significantly towards the typical values given in Calliope (1989). Two exceptions must be noted: (1) the F2 of both vowels did not change for AG, and (2) the F1 of /æ/ shifted excessively for AVG. The second important conclusion is that AVG changed significantly more than AV in all cases, except for the F1 of /æ/ (t(338)=0.7, n.s.). Note that the AV and AVG groups, though randomly selected, were not homogeneous, possibly due to the differences among Mandarin varieties. Indeed, the F1 means of /æ/ and /æ/ for the four north-eastern Chinese (AVG) were respectively 637 Hz and 662 Hz, whereas for the rest of the learners the values were only 548 Hz and 561 Hz. Note also that the four north-eastern Chinese’s F1 of /æ/ shifted from 662 Hz to 744 Hz, whereas the other three leaners of AVG shifted from 603 to 664 Hz (to compare with 647 Hz in Calliope (1989)). This could explain the aforementioned excessive shift for the mean of F1 measured over all the learners of AVG.

4. Discussion and conclusions

Some other aspects of these results are worth discussing. The F2 increased for both vowels for AVG but not for AV. This might be attributed to the fact that the vision of the tongue may have prompted the learners to centralize their articulation, as shown by Figure 6. Note also that the F1 difference between /æ/ and /æ/ after training was 62 Hz for AG and 117 Hz for AVG, to compare with 178 Hz in the Calliope (1989) data: an unpaired t-test showed that the difference of these two F1 differences was highly significant. (t(338)=−4.56, p<0.001). This indicates some better learning effect for the AVG group, though more progress remains to be done.

Besides, the perception test showed that the correct response percentage for /æ/ rose from 61% (42-76%) to only 68% (49-82%) for AG and from 50% (32-67%) to 86% (68-94%) for AVG (the 95% confidence interval of correct response percentage was defined as the Wilson score interval (Wilson (1927))). This result contributes to validate the advantage of the audiovisual approach.

Some of the learners reported the wish to have a frontal view of lips instead of the sagittal view. If we consider the decrease of F3 as a sign of lip rounding for front vowels (Harrington (1927)). This result contributes to validate the advantage of the audiovisual approach.

5. Acknowledgements

We would like to thank Cédric Gendrot for the Praat scripts used in this study (http://gendrot.ilpga.fr/scripts.htm) and Gérard Bailly for the idea of using speech shadowing as a supplementary learning strategy. This work was partially supported by the Centre for Chemistry, Life and Healthcare Sciences and Bioengineering of University Joseph Fourier, Grenoble, France (Vizart3D project).

6. References


Articulatory vowel spaces of male and female speakers
Melanie Weirich, Adrian P. Simpson
Friedrich-Schiller-University, Jena, Germany
melanie.weirich@uni-jena.de, adrian.simpson@uni-jena.de

Abstract
Studies have shown that although females on average have a larger acoustic vowel space than males, they exhibit a smaller articulatory vowel space. From this it is hypothesized that sex-specific differences in undershoot might exist. Articulatory vowel space sizes and Euclidean distances between vowel positions are analyzed in nine German speakers (5m, 4f) by means of electromagnetic articulography. Analyses include different sentence accent positions and two different sequences varying in their expected coarticulatory induced degree of undershoot. Results show a relationship between undershoot and speaker sex with males being more affected by accent and coarticulatory induced undershoot than females.

Keywords: vowel space, sex-specific differences, articulation

1. Introduction
For several languages it has been found that female speakers have a larger acoustic vowel space than male speakers (e.g. Diehl et al. (1996) for American English; Whiteside (2001) for British English; Weirich & Simpson (2014) for German). Different reasons have been held responsible for this including behavioral and physiological factors. Despite having a larger acoustic vowel space, females have been found to have a smaller articulatory space (Simpson 2001, 2002). Furthermore, Simpson (1998) found sex-specific differences in correlations between formant values and duration. While male speakers showed the expected significant relationship in read speech between duration and e.g. F1 of /æt/ or /u/, female speakers often did not exhibit this relationship. From this and the mismatch between articulatory and acoustic vowel spaces the question arises whether there might be sex-specific differences in terms of undershoot. In other words, do females reach their articulatory targets more often than males? The phenomenon of undershoot has been found cross-linguistically in many studies investigating the role of accent and lexical stress (Öhmann 1967, Fowler 1981, Rietveld & Koopmans-van Beinum 1987, de Jong 1995, Palethorpe et al. 1999, Harrington et al. 2000, Cho 2004). Vowel undershoot patterns have been explained both in terms of paradigmatic enhancement and by varying degrees of coarticulation. With respect to paradigmatic enhancement, more peripheral vowels are supposed to be found in stressed or accented positions, whereas in unstressed or unaccented positions vowel centralization is expected. Following the coarticulatory explanation, undershoot in unstressed or unaccented positions is a product of a higher degree of coarticulation related to contextual reduction. Vowel reduction due to increased coarticulation is explained by target undershoot reflecting the shorter durations in unstressed/unaccented positions. Mooshammer & Geng (2008) examined acoustic and articulatory vowel reduction patterns of tense and lax vowels in German in stressed and unstressed syllables. They found a higher degree of coarticulation in unstressed vowels than in stressed vowels. Thus, a relationship between accent/stress, degree of coarticulation and target undershoot in terms of vowel reduction seems to exist. However, if females reach their articulatory targets earlier and more often than men, as is hypothesized above, then they should be less influenced by accent induced undershoot. To test this, articulatory vowel spaces are compared between male and female speakers and the potential interaction of accent and sex-specific differences is analyzed. In addition, if females reach their articulatory targets earlier, they should also be less affected by coarticulatory induced undershoot. Therefore, articulatory vowel spaces of male and female speakers are compared a) in temporally privileged positions where the effect of coarticulatory induced undershoot is expected to be minimal and b) in the sequence /gV/.

2. Method
2.1. Speakers and speech material
Articulatory recordings of 5 male and 4 female German speakers were made at Potsdam University with the NDI-Wave system. The speakers were between 23 and 43 years old and revealed no known speech or hearing impairments. All of them came from the Eastern Central German dialect area but showed very little dialectal influence based on the auditory impression of the authors. Altogether, six coils were attached to the tongue, the lips, and the lower jaw. Four coils, one above the upper incisors, one at the bridge of the nose and two behind the participants’ ears served as reference coils which could be used to compensate for head movements. The bite plane was established using three sensors attached to a set square which a speaker held between his/her teeth.

The speech material presented here is twofold and part of a larger corpus comprising 20 different target words in varying accent conditions. The first set of data used in the current study includes the three point vowels /æt/, /i:/, /u:/ contained in the vowel sequences in the abbreviations IAA, AUA and BIU. The articulatory positions of the vowels were expected to be extreme and only minimally affected by coarticulatory influences due to their temporally privileged occurrence in the abbreviations. Each target word was repeated 10-12 times and embedded in a carrier sentence, e.g. Sie fuhren letzte Woche zur IAA ganz schnell (‘They went to the IAA very fast last week’). The second set of data comprises the sequence /gV/ with V being /i:/ or /e:/ or /æ:/ or /u:/ in the name GVbi embedded in the carrier sentence Ich sah GVbi an (‘I looked at GVbi’). For the /gV/-material, three different accent conditions were recorded. First, the participants were asked to read the sentences presented to them from a screen (neutral condition, n). Second, speakers produced the name in response to questions from the experimenter. For the accented condition (a) the experimenter asked 1) Sahst du /gVbi/ oder /gVbi/ an? (‘Did you look at /gVbi/ or /gVbi/?’), eliciting in the reply an accentuation of the name. For the unaccented condition (u) the experimenter asked 2) Siehst du oder sahst du /gVbi/ an? (‘Do you or did you look...”)...
at /gVbi?/), eliciting in the reply an accentuation of the verb. Each vowel was again repeated 10-12 times in each accent condition. The records were presented in a randomized order with additional filler sentences in between.

2.2. Articulatory labeling and analyses

All articulatory labeling was done with the help of mview (software written by M. Tiede). Here, we will concentrate on the horizontal and vertical position of the backmost lingual sensor (tongue dorsum). First, the articulatory position at the acoustic midpoint of the double vowel sequences were measured (cf. Figure 1, left graph). This provided a notionally extreme articulatory vowel space for each speaker (the IAU-polygon). To compare the articulatory positions across speakers, the data was translated with the midpoint set to the origin (0/0). The vowel space size of the IAU-polygon and the Euclidean Distances (EDs) between the vowels were calculated. Also, the EDs from the midpoint of the polygon to each vowel were measured for each speaker. Second, the start and end of the opening gesture /gV/ was labeled for each speaker; stimulus and accent condition oriented on the tangential velocity of the tongue dorsum sensor (cf. Figure 1, right graph). The end points of the gestures (which correspond to the vowels) were used to parameterize the polygon area spanned by the vowels contained in the /gV/-sequences and the EDs from the midpoint. Third, each speaker’s /gV/-polygon was normalized by expressing it as a percentage of the temporally privileged IAU-vowel space. The percentages were calculated for the polygon size and the EDs between vowels and midpoint.

For statistical analyses, Welch two sample t-tests and linear mixed models (LMMs) were conducted. LMMs as implemented in the lme4 package (Bates et al. 2011) were run in R (version 2.14.1, R Development Core Team 2008) with the articulatory measurements as dependent variable, the fixed factors speaker sex, vowel and accent condition (for the /gV/-data), and the random factors speaker and repetition. Likelihood ratio tests were used for model comparisons with different factors included to find the model with the best fit to the data.

3. Results

3.1. The impact of speaker sex on the articulatory space of the IAU-polygon

Figure 2 shows the mean articulatory vowel spaces measured in the double vowel sequences separated by sex. The data was translated and centered on the origin. As expected, the females on average reveal a smaller vowel space than the males (66 mm² vs. 93 mm²). Moreover, the dimensions of the space differ between the sexes: while the male speakers reveal a larger vertical than horizontal expansion (around 1.3 times larger), the relationship between vertical and horizontal expansion for the females is ca. 1 (or even less). A significant difference between males and females was found for the Euclidean distance (ED) between /iː/ and /aː/ with a larger distance for males (Welch two-sample t-test, t = -2.7, df = 5.9, p < .05).

This is mirrored in the statistical results. For both vertical and horizontal tongue position as dependent variable, LMMs revealed the best fit to the data with the interaction term Sex*Vowel included. For the horizontal tongue position, significant differences between males and females were found for /aː/, for the vertical tongue position males and females differ significantly in /aː/ and /iː/ (see Table 1). Thus, the female vowel space is smaller in terms of a higher and more fronted position for /aː/ and a lower position for /iː/.

Table 1: Summary statistics for fixed factors Sex and Vowel, dependent variable is vertical tongue position (default level is vowel /iː/ for females, observation: 263, rep.: 12, speakers: 9)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>t-value</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.1</td>
<td>0.2</td>
<td>19.1</td>
</tr>
<tr>
<td>Sex [f vs. m]</td>
<td>1.7</td>
<td>0.2</td>
<td>5.8</td>
</tr>
<tr>
<td>Vowel /iː/ vs. /aː/</td>
<td>-10.4</td>
<td>0.3</td>
<td>-34.5</td>
</tr>
<tr>
<td>Vowel /iː/ vs. /aː/</td>
<td>-2.0</td>
<td>0.3</td>
<td>-6.5</td>
</tr>
<tr>
<td>Sex(m)*Vowel[iː/aː]</td>
<td>-3.7</td>
<td>0.4</td>
<td>-9.1</td>
</tr>
<tr>
<td>Sex(m)*Vowel[iː/aː]</td>
<td>-1.1</td>
<td>0.4</td>
<td>-2.8</td>
</tr>
</tbody>
</table>

3.2. The impact of speaker sex and accent condition on the articulatory space of the /gV/-sequence

Figure 4 (left graph) shows the polygon area spanned by the tense vowels from the /gV/-sequence separated by sex and...
accent condition. First of all, the figure reveals a higher interspeaker variability for the males than for the females for all conditions. A possible reason for this might be that males are less restricted than females in terms of physiological boundaries and thus show this articulatory freedom in more variability. The figure also shows that while males reveal on average larger polygon sizes for the neutral and the accented condition, the picture is reversed for the unaccented condition. The same was found for the lax vowels, but here high inter-speaker variability was found for males and females. The right graph of Figure 4 shows the EDs measured from the midpoint of the polygons to the tense vowels /iː eː aː uː/. Again, the data is separated by speaker sex and accent condition. It is apparent that the difference between the accented and unaccented condition is much higher for the males than for the females.

![Polygon sizes spanned by the tense vowels of the /gV/-sequence](image)

**Figure 4:** Polygon sizes spanned by the tense vowels of the /gV/-sequence (left graph) and ED from the midpoint to the tense vowels of the /gV/ sequence (right graph) separated by sex and accent condition

In order to increase the number of data points that can be used, but also to enable a vowel specific analysis to be undertaken, the EDs from the midpoint of the vowel space to each vowel were used as dependent variable for the statistical analysis (and not the overall polygon size). Model comparisons showed that the LMM with the interaction terms Sex*Accent and Sex*Vowel as fixed factors revealed the best fit to the data (random factors included were speaker and repetition). Table 2 gives the summary statistics for the fixed effects. To summarize, there is a significant different reference between males and females in the EDs only for the vowel /aː/. Second, males show a significant difference between the accented and unaccented condition, while females do not. Since there was no three-way interaction of Accent, Vowel and Sex, this sex-specific difference in the effect of accent seems to hold for all vowels.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>t-value</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.1</td>
<td>0.6</td>
<td>9.1</td>
</tr>
<tr>
<td>Accent [a vs. n]</td>
<td>0.5</td>
<td>0.2</td>
<td>2.7</td>
</tr>
<tr>
<td>Accent [a vs. u]</td>
<td>-0.1</td>
<td>0.2</td>
<td>-0.3</td>
</tr>
<tr>
<td>Sex [f vs. m]</td>
<td>2.3</td>
<td>0.7</td>
<td>3.1</td>
</tr>
<tr>
<td>Vowel [/iː /aː /uː]</td>
<td>2.3</td>
<td>0.3</td>
<td>9.0</td>
</tr>
<tr>
<td>Vowel [/aː /iː /uː]</td>
<td>2.4</td>
<td>0.2</td>
<td>9.7</td>
</tr>
<tr>
<td>Vowel [/iː /aː /uː]</td>
<td>1.9</td>
<td>0.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Vowel [/aː /uː]</td>
<td>1.2</td>
<td>0.2</td>
<td>4.9</td>
</tr>
<tr>
<td>Accent [u]*Sex [m]</td>
<td>-0.8</td>
<td>0.3</td>
<td>-3.1</td>
</tr>
<tr>
<td>Accent [u]*Sex [m]</td>
<td>-1.4</td>
<td>0.3</td>
<td>-5.1</td>
</tr>
<tr>
<td>Sex[m]*Vowel[/eː ]</td>
<td>-1.4</td>
<td>0.3</td>
<td>-4.0</td>
</tr>
<tr>
<td>Sex[m]*Vowel[/iː ]</td>
<td>-1.6</td>
<td>0.4</td>
<td>-4.4</td>
</tr>
<tr>
<td>Sex[m]*Vowel[/aː ]</td>
<td>-2.8</td>
<td>0.3</td>
<td>-8.6</td>
</tr>
<tr>
<td>Sex[m]*Vowel[/uː ]</td>
<td>-1.9</td>
<td>0.3</td>
<td>-5.7</td>
</tr>
</tbody>
</table>

**Table 2:** Summary statistics for factors Sex, Vowel and Accent, dependent variable is absolute ED between midpoint and all tense vowels (in mm) (default level: vowel /aː/ in accented condition for females, observations: 1104, rep:s:12, speakers:9)

3.3. The impact of speaker sex and accent condition on the relationship between IAU-polygon and /gV/-sequence

To take the differences in speaker-specific physiology into account (in terms of its effect on the articulatory space that can be used in a temporally privileged condition), the /gV/-sequence was analyzed with respect to the speaker specific (extreme) IAU - vowel spaces. Figure 5 visualizes this relationship for four speakers. While the filled red points show the IAU - vowel space, the black and grey markers show the gestures /gV/ (start /g/ = black crosses, end /V/ = grey dots). The genders vary in terms of the use of their articulatory vowel space: both male speakers show a bigger vowel space measured in the temporally privileged IAU (red) – especially in terms of a lower tongue position for /aː/ – than in the /gV/ sequence (black and grey symbols). The females do not seem to differ between the two conditions.

![Tongue dorsi positions during the IAU (extreme) vowel space](image)

**Figure 5:** Tongue dorsum positions during the IAU (extreme) vowel space (red) and /gV/sequence of two male (M1, M2) and two female (F1, F2) speakers

Analogous to Figure 4 above, Figure 6 shows the polygon spanned by the tense vowels of the /gV/-sequence separated by sex and accent condition (left graph). However, this time the values are expressed in percent of the IAU-polygon (calculated for each speaker and accent condition separately). Interestingly, the males reveal smaller values than the females in every accent condition, but especially in the unaccented one. The right graph shows the EDs from the midpoints to the tense vowels /iː /aː /uː/ expressed as percent of the EDs measured in the IAU-polygon. Here again, the males differ between the accented and unaccented condition, while the females do not.

![Normalized polygon sizes (in % of the IAU-space)](image)

**Figure 6:** Normalized polygon sizes (in % of the IAU-space) spanned by tense vowels of /gV/-sequence (left) and EDs between midpoint and /iː /aː /uː/ of the /gV/-sequence (in % of the IAU-EDs) separated by sex and accent condition (right)
For the statistical analysis LMMs were run with the normalized EDs as dependent variable and likelihood ratio tests were used to find the model with the best fit to the data. This time the model with only the interaction term Accent*Sex as fixed factor turned out to be best (random factors were again speaker and repetition). Table 3 summarizes the results of the fixed factors. In contrast to the absolute values, there is no sex-specific difference anymore in the ED for /æ/ or for the other vowels. However, as in the model on the absolute data, the factor Accent reveals its significance in terms of sex-specific differences regarding its influence: while male speakers differ between the accented and unaccented condition, females do not. Again, the three way interaction of Sex. Vowel and Accent was not significant and including it did not reveal a better fit to the data which indicates that the interaction of sex and accent is valid for all vowels.

Table 3: Summary statistics for factors Sex and Accent condition, dependent variable is normalized ED between midpoint and tense vowels (in %) (default level is vowel /æ/ in accented condition for females, observations: 664, rep.: 12; speakers 9)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>t-value</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>83.1</td>
<td>5.4</td>
<td>15.4</td>
</tr>
<tr>
<td>Accent [a vs. n]</td>
<td>7.1</td>
<td>3.1</td>
<td>2.3</td>
</tr>
<tr>
<td>Accent [a vs. u]</td>
<td>0.5</td>
<td>3.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Sex [f vs. m]</td>
<td>5.1</td>
<td>7.3</td>
<td>0.7</td>
</tr>
<tr>
<td>Accent [i]*Sex [m]</td>
<td>-12.4</td>
<td>4.4</td>
<td>-2.9</td>
</tr>
<tr>
<td>Accent [u]*Sex [m]</td>
<td>-18.1</td>
<td>4.4</td>
<td>-4.1</td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

To summarize, our results replicate the findings of Mooshammer & Geng (2008) regarding more undershoot in unaccented than in accented condition in German vowels. At the same time, we have found evidence for sex-specific differences in the degree of undershoot already hinted at by Simpson (1998). Males reveal larger articulatory vowel spaces, as found in a temporarily privileged context in terms of a lower and more retracted tongue position for /æ/ and a higher tongue position for /ɪ/. The ED between these two vowels is significantly larger in males and thus, the vowels are articulatory further apart than in females. However, these sex-specific differences were found to be less in the coarticulatorily more vulnerable /gV/-sequence (in the neutral and accented condition) or even reversed (in the unaccented condition). Thus, sex-specific differences strongly depend on the accent condition. In detail, this means that while females did not differ in articulatory dimensions between accented and unaccented tense vowels, males revealed smaller excursions in the unaccented condition. Hence, the sexes differed in vowel reduction patterns that are due to target undershoot in an unaccented condition. In addition, in terms of normalized articulatory dimensions expressed in percent (with the help of a speaker-specific extreme articulatory vowel space measured in a temporarily privileged context) sex-specific differences decrease or disappear (as for the ED from the midpoint to /æ/1/) or even go in the other direction with higher values for females than males (as for the polygon sizes).

Although the neutral accent condition was elicited as well, our analysis here has focused mainly on comparing the accented and the unaccented condition. But as we can see from the figures and the tables for the EDs, there is a tendency for the female values to be highest in the neutral condition, for males in the accented condition. To conclude, we interpret our findings in terms of a higher amount of undershoot in unaccented conditions in male than in female speakers. We suggest that this might be caused by a stronger impact of coarticulation in males than in females. Since females have on average a smaller articulatory space due to physiological restrictions, they reach their targets earlier and thus are less affected by coarticulatory influences and also by accent induced undershoot.

5. Acknowledgements

This work was supported by a German Research Council Grant (SI 743/6-1/2) awarded to the second author. We would like to thank all participating subjects, our student assistant Miriam Oschkinat and Jana Brunner, Christian Geng and Adamantios Gafos at the Dept. of Linguistics, University of Potsdam where the articulatory recordings were made.

6. References


Mumbling is morphology?
Melanie Weirich\textsuperscript{1}, Adrian Simpson\textsuperscript{1}, Susanne Fuchs\textsuperscript{2}, Ralf Winkler\textsuperscript{2} & Pascal Perrier\textsuperscript{3}

\textsuperscript{1}Friedrich-Schiller-Universität Jena, \textsuperscript{2}ZAS, Berlin, \textsuperscript{3}GIPSA-lab, CNRS & Grenoble INP, France

melanie.weirich@uni-jena.de, adrian.simpson@uni-jena.de, fuchs@zas.gwz-berlin.de, winkler@zas.gwz-berlin.de, Pascal.Perrier@gipsa-lab.grenoble-inp.fr

Abstract
This study offers a possible physiological explanation for the observation that male speakers have repeatedly been found to exhibit less clear speech than females. We investigate one correlate of clear speech: jaw opening. In two articulatory data sets (American English, German) the jaw angle is examined as a function of sex and, additionally for the German data, accent condition. While no sex-specific differences are apparent in the American data, female speakers are found to have larger jaw opening angles than males in the accent condition in the German data. This empirical finding is borne out by the modeling study which compares the articulatory consequences of similar jaw opening settings in a typical male and a typical female articulatory model. In its most extreme setting the same angle of jaw opening in a male model gives rise to a complete radico-pharyngeal closure (but not in the female). The empirical and modeling findings suggest a possible physiological component in sex-specific differences in speech clarity.

Keywords: sex-specific differences, articulatory modeling, jaw angle, pharyngeal constriction

1. Introduction
Several studies have highlighted cross-linguistic sex-specific differences regarding clarity of speech, e.g. the larger female acoustic vowel space (Hillenbrand et al. 1995, Weirich & Simpson 2014). Heffnerman (2010) found a relationship between perceived masculinity and vowel space dispersion, consonant labialization and vowel length contrast, and concluded that “mumbling is macho”. In the majority of cases behavioral factors have been suggested as an explanation for clearer female speech. This in turn has been attributed to females being the primary care-givers and therefore at pains to produce clearer speech (Labov 1990).

A more pronounced articulation is often associated with a localized hyperarticulation (de Jong 1995) and a lower jaw position (Macchi 1985). Hyperarticulation is often associated with prosodically prominent positions. For prosodically prominent low vowels the jaw position has been found to be lower (Beckman & Edwards 1994). Interestingly, a lower jaw position in a prosodically prominent position has also been found for high vowels (Erickson et al., 1999, Harrington et al. 2000), which has been explained both in terms of increased loudness (Schulman 1989) and sonority (Harrington et al. 2000). Erickson (2002) found formants and tongue dorsum positions to be more peripheral in emphasized vowels for three speakers of American English. However, the jaw position was lower regardless of vowel height. The lowest vertical jaw position for /ai/ differed by about 2mm between the emphasized and unemphasized condition, but varied between speakers. Even though the study includes only 3 speakers and does not focus on sex-specific variability, differences between the sexes were present: the female speaker revealed a lower jaw position than the two male speakers in both stress conditions. In addition, the difference in jaw position between the conditions was largest for the female speaker.

Thus, the question arises whether the degree of jaw opening might also be affected by the sex of the speaker. If this is the case, then, besides behavioral factors, morphological parameters might also influence these sex-specific differences. Indeed, morphological factors have been found to influence speaker-specific characteristics in articulation (e.g. Winkler et al. 2006, Fuchs et al. 2008, Brunner et al. 2009, Lammert et al. 2013, Weirich & Fuchs 2013, Weirich et al. 2013). For example, Brunner et al. (2009) found that the height of the palate influences the amount of intra-speaker articulatory variability in vowel production (with less variability in speakers with flat palates). Weirich & Fuchs (2013) showed that the alveo-palatal steepness of a speaker affects the articulatory realization of the /ɪ/-/ɨ/ contrast. Lammert et al. (2013) found in their articulatory, acoustic and simulation study that while simulations revealed an impact of palate shape on formants, no such relationship was detected in real speech data. However, speaker-specific palatal morphology was correlated with lingual articulation. Thus, the authors assume that speakers vary in their articulatory strategies due to individual morphology adapting for potential differences in the acoustic output. Fuchs et al. (2008) and Winkler et al. (2006) found in their modelling study that the proportional length of the oral to the pharyngeal cavity is likely to affect speaker-specific articulation. This is especially interesting in the light of sex-specific differences since males on average have a proportionally longer pharynx than females (Vorperian et al. 2011). The longer pharynx generally correlates with larger facial height (Honda et al. 1996). This may also affect the distance between the temporomandibular joint and the lower incisors. The greater this distance is, the earlier it may cause a constriction of complete closure in the pharynx. In other words, imagine two pendulums of different lengths. Although both may move to the same degree (e.g. 10°) the longer pendulum will have a larger displacement at its extremity than the shorter pendulum. Transferring this example to speech production, one might expect that males avoid opening the jaw widely, since this can cause a constriction in the pharyngeal region if the tongue does not compensate for it.

The present study has two components. First, two experimental articulatory studies investigate the potential impact of speaker sex on jaw opening. The first of these comprises male and female English speakers of the Wisconsin X-ray microbeam speech database (Westbury, 1994) and focusses on possible sex-specific differences in jaw angle. The second is an EMA (electromagnetic articulography) study comprising female and male German speakers and takes into account the potential effect of different sentence accent conditions. In the second half, a modeling study analyzes the possible relationship between differences in vocal tract anatomy between males and females and sex-specific variability in the degree of jaw opening. The two experimental studies serve to elucidate potential sex-specific differences in
jaw opening and thus investigate whether “mumbling” - resulting from a smaller jaw opening - is indeed a male rather than female characteristic of speech. In other words the modeling study examines whether “mumbling” might follow from sex-specific morphological differences, and not just from behavioral and cultural factors.

2. Method

2.1 Wisconsin data base

By using the Wisconsin data base a large corpus of speakers could be investigated: in all, 18 male and 22 female speakers were included in the present analysis. A comprehensive description of the data collection is provided in Westbury (1994). We selected the target word coat in the sentence “The coat has a blend of both light and dark fibers”. The target word was chosen due to its expected large jaw opening movement, important to maximize the probability of finding sex-specific differences.

We calculated the difference in jaw opening between two sounds, the alveolar fricative in has and the vowel /oo/ in coat. We assumed a high and consistent jaw position in the alveolar fricative (Shadle 1990, Mooshammer et al. 2007). Thus, the jaw opening was normalized with reference to a speaker specific closed jaw position during speech. Figure 1 visualizes the calculation of the jaw angle using the intersection of the bite plane and the pharyngeal wall (which is included in the Wisconsin data base) as a fixed reference point (RP). Of the three repetitions of the sentence in the corpus one token per speaker was chosen which was clearly accented. The selection was based on the perceptual impression of one annotator.

We assumed a high and consistent jaw position in the alveolar fricative (Shadle 1990, Mooshammer et al. 2007). Thus, the jaw opening was normalized with reference to a speaker specific closed jaw position during speech. Figure 1 visualizes the calculation of the jaw angle using the intersection of the bite plane and the pharyngeal wall (which is included in the Wisconsin data base) as a fixed reference point (RP). Of the three repetitions of the sentence in the corpus one token per speaker was chosen which was clearly accented. The selection was based on the perceptual impression of one annotator.

We assumed a high and consistent jaw position in the alveolar fricative (Shadle 1990, Mooshammer et al. 2007). Thus, the jaw opening was normalized with reference to a speaker specific closed jaw position during speech. Figure 1 visualizes the calculation of the jaw angle using the intersection of the bite plane and the pharyngeal wall (which is included in the Wisconsin data base) as a fixed reference point (RP). Of the three repetitions of the sentence in the corpus one token per speaker was chosen which was clearly accented. The selection was based on the perceptual impression of one annotator.

2.2 EMA study

The EMA study comprises 4 female and 5 male German speakers. By analyzing both German and English speakers results can be interpreted cross-linguistically. In addition, the controlled realization of different accent conditions was included and the number of observations was increased by using repetitions.

Articulatory recordings (with the NDI-Wave system) were conducted at the Dept. of Linguistics at the University of Potsdam. Vertical and horizontal movements of the tongue (tongue tip, tongue mid and tongue dorsum), the lips (upper and lower lip), and the jaw were recorded. Reference sensors at the nose, the incisors, and behind the speakers’ ears were used for correction of head movements. The bite plane was established using three sensors attached to a set square held between a speaker’s teeth. Analogous to the Wisconsin data, the jaw movement was analyzed during the sequence /za/ (“sah”) that was part of the carrier sentence “Ich sah GVbi an” (I looked at GVbi). In addition to the control condition (asking the subjects to read the sentence from a list), two accent conditions were elicited by making the subject answer two questions: a) accented condition (with /za/ under focus), and b) pre-accented condition (with GVbi under focus). The number of repetitions varies between conditions and speakers due to untrackable sensor positions, synchronization problems and speech errors: on average 75.8 (SD: 22.6) repetitions for the control condition, 58.4 repetitions (SD: 22.6) for the accented condition and 71.7 (SD: 18.3) repetitions for the pre-accented condition were recorded for each speaker. The articulatory labeling was done with the help of mview (software written by M. Tiede). The start and end of the vocalic opening gesture /za/ was labeled for each speaker and stimulus based on the tangential velocity of the jaw sensor. Then the degree of jaw opening (α) was calculated (as in the Wisconsin study) between the two lines going from the determined jaw positions for /za/ and /a/ to a speaker specific reference point. Here, a fixed point behind the speakers’ left ear was used as the reference point (RP in Figure 1). In total, 1852 observations are included in the analysis.

2.3 Modeling study

Based on MRI data of a male speaker with a typically long pharynx and a female speaker with a typically short pharynx we built two models that allow us to test whether the degree of jaw opening has an impact on the degree of constriction of the tongue in the pharyngeal cavity. The position of the temporomandibular joint (TMJ) has been inferred for each subject from an MRI sagittal view. To do so we first located the most anterior point of the sphenoid bone and we applied a translation deduced from anatomical data to define the centre of the TMJ. The TMJ is modelled as a simple straight line with a constant slope, similar for male and females. Jaw movements are modelled as rotations around a reference point that represents the condyle. This condyle is allowed to move but it is constrained to stay on the line that simulates the TMJ. When the jaw moves down, the condyle slides forwards, when the jaw moves up the condyle slides backwards. In our model the relationship between jaw rotation and condyle translation is the same for males and females. The influence of jaw movements on the tongue position is modelled according to the method proposed by Zandipour (2006). The jaw is considered to be a simple independent carrier of the tongue (i.e. the dynamic interaction between the two articulators is not accounted for). We modelled the degree of jaw opening with respect to a reference position (0° jaw opening) corresponding to the rest position of the tongue. In the two models (male and female), different results may be attributed to a combination of anatomical differences. To further evaluate the different effects that may contribute to a pharyngeal constriction, we additionally measured a) the distance between the condyle and the top of the lower incisor (pendulum example in the introduction, see arrows in Figure 2) and b) the position of the condyle with respect to the pharyngeal wall.

Figure 1: Schematic visualization of jaw angle (α) measurement between /z/ and vowel (V=/oo/ in the Wisconsin data, V=/a/ in the EMA data)

Figure 2: Schematic representations of the jaw opening angle (α) as a function of varying condyle positions (C=Condyle)
The position of the reference point (here condyle) may crucially determine the calculation of the jaw angle (see Figure 2). While the position of the lower incisor stays constant, the more posterior the condyle position (with respect to the vocal tract) is, the longer the distance between the condyle and the lower incisors will be. This, in turn, will affect the jaw angle which is smaller for a more posterior condyle position (see red angle vs. black angle in Figure 2).

3. Results

3.1 Wisconsin data

Figure 3 shows the distribution of jaw angle in coat separated by gender. The closer the value is to 0, the smaller the jaw opening. From the figure it is apparent that female speakers exhibit a distribution with the highest density around 5 degrees. The males, however, show a distribution with a lower peak at around 3 degrees. Even though this tendency for a greater jaw opening in females is apparent from Figure 3, a Welch Two Sample t-test revealed that the influence of gender was not strong enough to result in a significant difference between the angle means. Of course, a possible reason for that can be the limited number of observations with only one token per speaker.

![Distribution of jaw angle (in degree) during coat for 22 female (red, left) and 18 male (blue, right) speakers](image)

Figure 3: Distribution of jaw angle (in degree) during coat for 22 female (red, left) and 18 male (blue, right) speakers

3.2 EMA study

The angle of jaw opening was calculated for all repetitions and each speaker and accent condition separately. Figure 4 shows the results separated by sex and accent condition. While there is an overall tendency for female speakers to have higher angles (i.e. greater jaw opening), the difference is only substantial for the accented condition. Here, male speakers have a mean jaw angle of 3.4 degrees (SD = 1.4), which is similar to the values in the Wisconsin study. Females exhibit on average a jaw angle of 6.6 degrees (SD = 1.6), which is higher than the values in the Wisconsin data.

![Angle of jaw opening (in degree) during /za:/ separated by sex and accent condition](image)

Figure 4: Angle of jaw opening (in degree) during /za:/ separated by sex and accent condition

Linear mixed models as implemented in the lme4 package (Bates et al. 2011) were run in R (version 2.14.1, R Development Core Team 2008) with jaw angle as dependent variable. We included the fixed factors Sex, Accent (condition) and Repetition, and an accent condition specific random intercept for speaker. Likelihood ratio tests revealed that the model with the interaction term Sex*Accent included shows the best fit to the data (p < .05). After calculating the model, p-values based on Markov-chain Monte Carlo sampling were computed and Table 1 gives a summary of the results of the fixed effects. The reference level is the accented condition of the female speakers. As Figure 4 already shows, a significant sex-specific difference in jaw angle exists only for the accented condition (such that males have a smaller jaw opening of about 3 degrees), while for the control and pre-accent condition the differences between the sexes are negligible and insignificant (less than 1 degree). Thus, only if the angle reaches a certain value, as it does for the accented condition, do sex-specific differences appear.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>t-value</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>6.62</td>
<td>0.53</td>
<td>12.48</td>
</tr>
<tr>
<td>Repetition</td>
<td>0.00</td>
<td>0.00</td>
<td>-1.84</td>
</tr>
<tr>
<td>Sex [f vs. m]</td>
<td>-2.98</td>
<td>0.71</td>
<td>-4.19</td>
</tr>
<tr>
<td>Accent [accent vs. control]</td>
<td>-3.28</td>
<td>0.75</td>
<td>-4.38</td>
</tr>
<tr>
<td>Accent[accent vs. pre-accent]</td>
<td>-3.89</td>
<td>0.75</td>
<td>-5.20</td>
</tr>
<tr>
<td>Sex[m]*Accent[control]</td>
<td>2.08</td>
<td>1.00</td>
<td>2.07</td>
</tr>
<tr>
<td>Sex[m]*Accent[pre-accent]</td>
<td>2.32</td>
<td>1.00</td>
<td>2.31</td>
</tr>
</tbody>
</table>

3.3 Modeling

Comparing both models, a narrower linguo-pharyngeal constriction was found in the male model than in the female model for the same angle of jaw rotation. This difference in the size of the constriction is particularly evident when the jaw is opened to 9 degrees. The male model even produces a complete radico-pharyngeal closure (Figure 5, right), whereas this is not the case in the female model (Figure 5, left).

![Jaw opening from the tongue rest position to 9 degrees of jaw opening, left: female, right: male](image)

Figure 5: Jaw opening from the tongue rest position to 9 degrees of jaw opening, left: female, right: male

The narrower constriction in the male model can be attributed to two anatomical differences: First, the distance between the condyle and the lower incisor is 103 mm for the male and 98 mm for the female. Comparable length differences between males and females have also been found in the EMA study. Second, the condyle position (red dots is Figure 5) is further back with respect to the rear wall of the pharynx for the male than for the female (when drawing a line from the condyle in parallel to the pharyngeal wall, the closest distance to the tongue is 20 mm for the male and 12 mm for the female), thus positioning the male tongue relatively closer to the rear pharyngeal wall.

4. Discussion and conclusion

This study brings together articulatory findings from two different articulatory data sets. In the first, the Wisconsin set,
we have access to a large number of speakers, but a comparatively small number of tokens. Unfortunately this data set did not provide ideal data, i.e. the same word in two different accent positions. In the second, German, data set a much smaller speaker sample produced a large number of repetitions of relevant items in three different accent conditions. And here, we were able to gain a first insight into possible sex-specific differences in accent related jaw opening, in line with previous studies which have found a larger jaw opening in females.

Modeling these data, we were able to show that such jaw opening differences may indeed be, in part at least, the result of sex-specific differences in vocal tract morphology. In particular, a longer distance between the condyle and the incisors leads more rapidly to a pharyngeal constriction than a shorter distance. Differences become especially apparent when a particular angle of jaw opening is reached, consistent with our empirical finding that angle differences are present in the accented but not in the unaccented condition. This in turn may provide some account for a lack of a significant finding in the Wisconsin data set. Moreover, defining the jaw reference points in the different data sets is an approximation that may have added noise to the results. The findings of our empirical and modeling study are not designed to champion physiological factors above all else but we were able to show that certain correlates of speech clarity might not be purely grounded in behavioral differences. Future work should look at potential male-female differences in the position of the condyle with respect to the tongue and relative to the vocal tract. From the model, the more posterior this position is, the greater the influence on the pharyngeal constriction. In other words, we suggest that mumbling may not only be macho or due to laziness, but physiologically inevitable. Males may avoid large degrees of jaw opening, as it occurs in accented syllables, in order to avoid narrow constrictions that might lead airflow to switch from laminar to turbulent.

5. Acknowledgements

This work was supported by two German Research Council Grants (SI 743/6-1/2 and FU 791/1-1) awarded to the second and third author respectively. EMA-recordings were made at Dept. of Linguistics, University of Potsdam: many thanks to Jana Brunner, Christian Geng and Adamantios Gafos and all participating subjects.

6. References

Bates, D., Maechler, M., & Bolker, B. (2011) _lme4: Linear mixed-effects models using S4 classes_, R package version 0.999375-42.


Prosodic Cues and Varying Dialogue Context - Impact on Exhaustivity of Answers
Charlotte Wollermann1, Oliver Niebuhr2, Karin Görs3, and Bernhard Schöder1

1German Linguistics, University of Duisburg-Essen, Germany
2Department of General and Comparative Linguistics, ISFAS, Kiel University, Germany

charlotte.wollermann@uni-due.de, niebuhr@isfas.uni-kiel.de,
kgoers@arcor.de, bernhard.schroeder@uni-due.de

Abstract
The presented study investigates the contribution of dialogue context and of prosody on pragmatic focus interpretation. More particular, we test to what extent specific contextual variation combined with different pitch accent patterns (hat pattern versus dip pattern) biases the interpretation of answers towards exhaustivity or towards non-exhaustivity. For measuring focus interpretation we use pictures which are judged by listeners. From the rankings of the pictures, we infer the listeners' preference for interpretation. The results of our interpretation study suggest that contextual variation in combination with prosodic variation has a significant effect on pragmatic focus interpretation.

Keywords: pragmatic focus, prosody, dip pattern, hat pattern, exhaustivity of answers

1. Introduction
The presented study investigates the contribution of context combined with pitch accent patterns on the exhaustive interpretation of answers to preceding questions in a virtual dialogue scenario. In Section 1 we will present the theoretical background of our work, starting with definitions of focus and of exhaustive interpretation of answers. In this context, we will also discuss the role of prosody for pragmatic focus interpretation. Afterwards, we will introduce a particular type of intonational contrast, i.e. the hat pattern versus dip pattern distinction. In a next step, we relate the concept of focus to the hat pattern vs. dip pattern contrast. Finally, our assumption will be presented.

In Section 2 the material will be described that we use in our study. In Section 3 we present our experimental study, and finally, we summarize our results and discuss them in Section 4.

1.1. Focus interpretation and exhaustivity of answers
Defining the term focus is problematic since the term is used differently in the literature (see Krifka, 2007 and Fisseni, 2011 for an overview). In our approach, we investigate question-answering focus which constitutes a pragmatic focus. The latter is usually applied to that constituent in the answer which corresponds to the interrogative pronoun in the question. According to semantic-pragmatic theories (Rooth, 1992; Groenendijk, Stokhof, 1984), in the context of such a question, pitch accent correlates with focus. If the hearer interprets the (potentially implicit) background question as a mention-all question, an exhaustive interpretation should be triggered. Consider example (1), taken from van Rooij and Schulz (2004: 498).

(1) a. Who called yesterday?
   b. [Peter], called yesterday.

If the hearer takes (1b) to mean that Peter was the only person who called yesterday, the interpretation is exhaustive. But, if s/he infers that Peter was one of the persons who called yesterday – among others –, then the interpretation is non-exhaustive.

The results of Fisseni (2011) suggest that focus accentuation does not suffice to trigger pragmatic focus interpretation. Rather, the expectations of the hearer, the language awareness of the hearer and, in particular, the context also play an important role.

In Wollermann (2012) the following was assumed: if the speaker signals uncertainty (e.g. by a rising intonation) when realizing the focus constituent, the hearer uses the prosodic cue and assumes that the speaker is uncertain with respect to his/her epistemic knowledge. Thus, the interpretation should be biased towards non-exhaustivity. Wollermann's findings show that there was an influence of prosody, but weaker than suggested theoretically. By contrast, the contextual effect was stronger. Based on these findings, a model of focus interpretation was developed (Wollermann, 2012; see also Wollermann et al., 2010). This model accounts for the relative contribution of context and of prosody for pragmatic focus interpretation. Information is processed top-down as well as bottom-up. The expectations of the listener are, for instance, affected by contextual information and processed top-down. The prosodic cues influence focus interpretation in the bottom-up direction. According to this model, the following prediction can be made: the influence of prosodic cues increases when the listeners expectations are less clear.

With respect to written language, the study of Skopeteas and Fanselow (2011) suggests that both question and answer type affect the exhaustive interpretation of answers.

In the current study, we use the same method as Wollermann (2012) for investigating the effect of contextual variation on focus interpretation in German, but this time we combine this variation with the intonational hat vs. dip pattern contrast. Both the context and the intonational pattern constitute together the independent variable.

1.2. Pitch accent patterns: Hat pattern versus dip pattern
According to the Kiel Intonation Model (KIM; Kohler, 1991) it is assumed that concatenation patterns in between two pitch accents are phonologically distinctive. Two basic types of
concatenations are distinguished: dipped or non-dipped. A succession of two pitch accents with a high-level concatenation in between constitutes a *hat pattern*. All other patterns are dip patterns. Ambrazaitis and Niebuhr (2008) provided empirical evidence for this difference by showing that a change from *hat* to *dip pattern* is reflected in the perceptual and functional interpretations of utterances. Niebuhr and Zellers (2012) corroborate this evidence with different stimulus material, but suggest additionally that the pitch accent paradigms in *hat patterns* may not be the same as those in *dip patterns*. Production data of Grice et al. (2009) basically also support the relevance of a *dip*.vs. *no dip* before a non-initial pitch accent. However, the *hat-dip* contrast is not a separate phonological feature in GToBI (Grice, Baumann, 2005; Grice et al., 2005).

1.3. Assumption

It was argued in Ambrazaitis and Niebuhr (2008) that a *hat pattern* has a bracketing function, whereas a *dip pattern* indicates a semantic-pragmatic detachment of two focused arguments of the coordination structure. Therefore, when it comes to listing persons or entities in conversation, *hat patterns* are typically used in combination with complete lists, whereas *dip patterns* are preferably used for incomplete or basically open lists.

We assume on this basis that a *hat pattern* across two focus arguments in combination with a context conveying certainty about the exclusion of alternatives with respect to the focused item would bias the interpretation towards *exhaustivity*. In contrast, we expect that a *dip pattern* combined with a context conveying uncertainty about the exclusion of alternatives with respect to the focused item would trigger a *non-exhaustive interpretation*.

2. Material

The goal of this study was to investigate the effect of the intonational *hat pattern* versus *dip pattern* contrast and also the effect of context on focus interpretation in German. Therefore, question-answer pairs were generated. The answer constituted the utterance with the pragmatic focus and was characterized by a *hat pattern* or a *dip pattern*. In a next step, the question answer pairs were embedded into short dialogues, which were also varied.

The dialogues were all about a party at which groups of students (biologists, physicians, designers etc.) performed different actions. Every action described in the dialogue led to a target question in which one dialogue partner asked for the corresponding agents of that action. The target answer with the relevant focused constituents, i.e. a coordination of two student groups (noun phrases), was then provided by the other dialogue partner. An example is given in (2a) and (2b). The pitch accented syllables of the two focused arguments are underlined.

(2) a. Wer hat denn Karikaturen gemalt?
   Who has drawn caricatures?
   b. [Die Chemikerinnen und Ingenieurinnen], haben Karikaturen gemalt.
   [The chemists and the engineers] have drawn caricatures.

The dialogue frames provided two kinds of contexts.

(1) **Certainty context:** one student group is salient during the dialogue. In addition the target answer ended in a sentence in which the speaker signalled certainty with respect to the performing agents (e.g. I am sure that X were the only ones who did/ performed as…).

(2) **Uncertainty context:** the group introduced at the beginning of the dialogue was not identical with the group mentioned in the target answer. Moreover, the target answer was followed by a sentence in which the speakers signalled uncertainty with respect to the performing agents (e.g. I am not really sure that X were the only ones who did/ performed as…).

As for the intonational cues, the two coordinated noun phrases that functioned as focus constituents in the target answers were either spanned by intonational *hat patterns* or *dip patterns*, each which comparably prominent H* pitch accents on either focus argument (see Figure 1).

It should be mentioned here that we used scripted lab speech in our study, even though this kind of speech lacks spontaneity of everyday speech. However, it is particularly important for us to vary contextual and prosodic cues in a systematic way, and to minimize the occurrence of confounding factors which may come along with spontaneous speech.

![Figure 1: Fc-contour for hat pattern (above) and for dip pattern (below)](image)

3. Experimental study

In this section we present our empirical study. Firstly, we describe our methodological approach. Afterwards, we outline our experimental procedure. Finally, our results are presented.

3.1 Method

Overall, we generated six different dialogues. There were two variants which differed regarding contextual and prosodic cues. For every dialogue, either variant *exh*+ (*exhaustive*) or variant *exh*− (*exhaustive−*) was used (see Table 1). To test whether there is any evidence for our assumption at all, we use the two ‘extreme’ combinations of context and prosody in the current study.

That is, for the variant *exh*+, there was only one student group salient during the dialogue. Furthermore, the coordinated noun phrases in the answer with the pragmatic focus were spanned by a *hat pattern*. Also, a sentence indicating certainty about the exclusion of the alternatives...
followed after the answer.

In contrast, in the case of variant \(\text{exh}^-\), one group of students was introduced at the beginning of the dialogue as usually performing the action under discussion. This group was not identical with the group mentioned in the focus utterance. For the realization of the coordinated noun phrases in the answer with the focus, a \(\text{dip pattern}\) was used. Furthermore, a sentence indicating uncertainty about the exclusion of the alternatives followed after the answer.

**Table 1: Two variants of the stimuli with different contextual and prosodic cues.**

<table>
<thead>
<tr>
<th>Variant</th>
<th>Context</th>
<th>Prosody</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{exh}^+)</td>
<td>only one salient student group, lexical expression of certainty about the exclusion of alternatives</td>
<td>hat pattern</td>
</tr>
<tr>
<td>(\text{exh}^-)</td>
<td>another student group mentioned as possible alternatives, lexical expression of certainty about the exclusion of alternatives</td>
<td>dip pattern</td>
</tr>
</tbody>
</table>

Furthermore, we generated three filler dialogues in order to distract subjects from our actual research question. An example dialogue was generated to make subjects familiar with the task.

For measuring focus interpretation we used the method of Wollermann (2012). Drawings were created that depicted either the exhaustive or the non-exhaustive reading. Subjects referred to these drawings when judging the dialogues. In this way, we wanted to avoid any complex and potentially biasing metalinguistic instructions of our subjects. Moreover, we wanted to avoid that the subjects' linguistic awareness is directed to the goal of our experiment.

As already mentioned in Section 1.3, our hypothesis is the following: we assume that variant \(\text{exh}^+\) biases the interpretation towards exhaustivity, whereas variant \(\text{exh}^-\) biases the interpretation towards non-exhaustivity.

**3.2 Procedure**

Subjects were 18 native speakers of German, 14 females, 4 males, who were on average 23 years old. They were undergraduate students at the University of Duisburg-Essen and participated in the experiment voluntarily at the beginning of a seminar.

Each dialogue was played, and afterwards the subjects had to judge each time on a 5-point scale (with 1 = very bad and 5 = very good), how well the subsequently presented drawing matched with the dialogue. Distractor and example dialogues were also included in the experiment. All in all 9 dialogues were played. Dialogues 1, 2, and 5 were judged against drawings signalling exhaustivity. Dialogues 3, 4, and 6 were judged against drawings illustrating the non-exhaustive drawings.

A Wilcoxon Signed Rank Test was used for statistical analysis. Since we perform several comparisons on our data, we use a Bonferroni correction. Starting from a level of significance of \(p \leq 0.05\), we applied a more conservative significance level, which reflected the number of tested comparisons: 0.05/4, i.e. \(p \leq 0.0125\).

**3.3 Results**

The diagrams in Figure 2 show the results. The upper diagram of Figure 2 shows that our subjects judged dialogues 2 and 5 with a median of 5 each time. In contrast, the median for dialogue 1 was about 4 and hence lower. The Wilcoxon Signed Rank Test showed that this judgement difference is statistically significant. For dialogue 5 vs. dialogue 1, the \(p\)-value was <0.0025; for dialogue 2 vs. dialogue 1, it was <0.00025. So, the subjects' rankings were significantly higher when the drawing representing the exhaustive reading coincided with the dialogue variant \(\text{exh}^+\) (dialogues 2 and 5). This variant is marked by contextual and intonational cues intended to bias the interpretation towards \textit{exhaustivity}. The ranking was lower when there was a mismatch between the information given by the drawing and the information given by the dialogue (dialogue 1).

![Figure 2: x-axis shows dialogues and the respective variant (exh+ or exh-), y-axis shows the median for recipients judgements. Level of significance: \(p = 0.05/4 = 0.0125\), \# \(p \leq 0.0125\), ** \(p \leq 0.0025\), *** \(p \leq 0.00025\).](image)

As regards the bottom panel of Figure 2, we can see the following: dialogues 4 and 6 were ranked with medians of 3.5 and 4 respectively. Dialogue 3 yielded the lowest median of 3. The corresponding Wilcoxon Signed Rank Tests resulted in significant differences between the judgements of these two variants. For dialogue 6 vs. dialogue 3, the \(p\)-value was <0.0025; for dialogue 4 vs. dialogue 3, it was <0.0125. We can conclude from this outcome that judgements were significantly higher when the picture depicting the non-exhaustive reading was presented in combination with the contextual and intonational dialogue variant which was intended to bias the
interpretation towards non-exhaustivity (dialogue 4 and 6). A lower ranking occurred in the case of a mismatch between the information illustrated in the drawing and that conveyed by the contextual and intonational cues (dialogue 3).

4. Discussion

We presented a study dealing with the contribution of varied context combined with different pitch accent patterns (hat pattern versus dip pattern) on the exhaustive interpretation of answers.

Our results suggest that the contextual variation in combination with intonational variation has a significant effect on the exhaustive interpretation of answers.

In the current study, we have not yet tested the two factors context and intonation independently from one another. The information transmitted by the speaker is regarded to be redundant in our approach. As Fisseni et al. (2013) suggest, the transmission of 'optimal' redundant information and its testing by means of empirical data can help to improve communication (for some earlier remarks on the role of redundancy for focus see also Fisseni, 2011: 227).

The question remains open what the relative contribution of context on the one hand, and hat vs. dip pattern on the other hand actually is. However, the intonational effect is clearly consistent with communicative function ascribed to the hat vs. dip contrast by Ambrazaitis and Niebuhr (2008). This also means that our results do not contradict the claim of the Kiel Intonation Model (KIM, Kohler 1991) to represent the intonational difference between hat and dip as a separate feature in intonational phonology, if a correlation between this prosodic distinction and its pragmatic interpretation can be established.

In our future work, we aim at testing all four combinations of the two factors context and prosody. We would like to collect more data by testing these four combinations for each of the six dialogues. Furthermore, for measuring focus interpretation, it would be interesting to use for each stimulus the drawing showing the exhaustive interpretation as well as the drawing showing the non-exhaustive interpretation.

Acknowledgements

We would like to thank Bernhard Fisseni and Nina Jeanette Hofferberth for helpful comments.

References


An exploratory ultrasound study of unreleased plosive consonant sequence in Cantonese

Ivan Yuen
Macquarie University, Australia
ivan.yuen@mq.edu.au

Abstract
A plosive consonant in Cantonese is phonetically unreleased in the coda, but is released in the onset. The acoustic characteristics of the plosive-plosive consonant sequence resembles a singleton stop, leading to a potential difficulty in differentiating the former from the latter. A recent study by Ghosh and Narayanan (2009) showed that closure duration could be used to distinguish incomplete stops in a stop-stop sequence from a singleton stop in American English, possibly attributable to assimilation and coarticulation. This raises the question whether places of articulation will be maintained in the unreleased coda in Cantonese. This study examined the acoustic and articulatory characteristics of the unreleased stop in the hetero-syllabic consonant sequence. A production experiment using ultrasound was conducted on two subjects who took part in a reading aloud task. Target words/characters containing the test stops /h/ or /k/ were combined with other words beginning with either /p/, /t/ or /k/ to generate various homorganic and heterorganic stop-stop sequences. Closure duration of the onset plosive was found significantly shorter than that of the hetero-syllabic stop-stop sequence, but closure duration did not differ across places of articulation. The preliminary analysis of the ultrasound data of one speaker also indicates articulatory anticipation.

Keywords: unreleased stop, articulation, Cantonese

1. Introduction
Cantonese is a tone language with CVC structure, allowing both plosives in the onset and coda positions. A plosive consonant in Cantonese is phonetically unreleased in the coda position, but it is released in the onset position. The acoustic characteristics of a plosive-plosive consonant sequence across two syllables then will resemble a complete single stop production with a closure phase followed by a release phase. Similar phonetic observation has been made in English (Abercrombie 1967; Hardcastle and Roach 1979). This raises the question of how to distinguish a singleton stop from a stop-stop sequence. Examining this question in American English, Ghosh and Narayanan (2009) showed that closure duration could be used to distinguish incomplete stops in a stop-stop sequence from a singleton stop. However it remains difficult to detect the place of articulation in unreleased stops.

In production, Browman and Goldstein (1990) and others have shown that an initial consonant, which is perceived to be either assimilated or deleted, still retains its original place of articulation. In addition, Byrd (1996) has reported less articulatory overlap in an onset consonant cluster than in a coda consonant cluster and a hetero-syllabic consonant sequence in American English. These observations raise further questions about articulatory movements of the hetero-syllabic stop-stop sequence in Cantonese. The goal of this study was to examine how the first consonant in the hetero-syllabic stop-stop sequence was realized, when the second stop consonant was varied.

2. Method
We conducted a sentence-reading experiment using ultrasound (Terasawa). There were four test conditions: (1) onset plosive (i.e. # stop), (2) coda plosive (i.e. stop #), and (3) hetero-syllabic stop-stop sequence (i.e. stop # stop) which was divided into (a) homorganic stop-stop sequence, and (b) heterorganic stop-stop sequence.

The test plosives under investigation were /t/ and /k/. The words containing the target segments were listed below according to the four test conditions:

(1) Onset plosives: # tɪn/, # kɪn/.
(2) Coda plosives: # kɪk̚ # jɪn/, # tɪt̚ # jɪn/, # kɪk̚ # jɪn/.
(3) Homorganic stop-stop sequence: # kɪk̚ # tɪt̚ # jɪn/, # tɪt̚ # jɪn/.
(4) Heterorganic stop-stop sequence: # kɪk̚ # kɪk̚ # tɪt̚ # jɪn/, # tɪt̚ # jɪn/.

Due to the phonotactic constraint in Cantonese, the vowel preceding the onset plosive was long in (1), and the vowel preceding the coda plosive was short in (2) and (3). To minimize the effect of different vowels on consonant articulation, we chose the diphthong /w/ in (1) and the vowel /a/ in (2) and (3), because they all end in /t/. In condition (1), four separate words containing the diphthong /w/ were used to combine with the stimuli containing the test onset plosives.

A list of two-word Cantonese names (Xsame) were then generated through various combinations. These names were then embedded in a carrier sentence: ‘He is also called Mr. Cheung Xsame’. These items were repeated two times.

Two Cantonese-speaking subjects (1F, 1M) were recruited at Macquarie University to take part in the experiment. The mean age was 40.

2.1. Predictions
We were interested in three questions. The first one concerned whether closure duration was used as a cue to distinguish the singleton stop from a stop-stop sequence in Cantonese. On the basis of the findings from Ghosh and Narayanan (2009), we expected closure duration to differentiate them in Cantonese. The second question concerned articulatory closure of the unreleased stops in the coda position. Do speakers retain the articulatory closure of the unreleased stop consonant in the syllable?
coda position, just like the released stop consonant in the onset position? Given the findings from Browman and Goldstein (1990) and others, we predicted the place of articulation distinction to be retained in the unreleased stop consonant. The third question concerned the articulatory realization of the first unreleased coda consonant in the hetero-syllabic /h#k/ and /k#u/ sequences relative to the respective homorganic /h#u/ and /k#k/ counterparts. We predicted co-articulation on the first consonant in the heterorganic sequence.

2.1.1. Acoustic and ultrasound annotation

We used the waveform and spectrogram to identify three time points: the beginning, the middle and the end of the closure interval. The beginning of the closure interval also coincided with the drop-off of F2 in the vowel preceding the closure onset. The end of the closure was identified as the onset of the burst release. Please see Figure 1a for the annotation of a singleton stop consonant in onset position and Figure 1b for that of a stop-stop sequence. The same closure interval was used to identify the segment(s) of interest in the ultrasound recordings. Using the same three time points in the acoustic analysis as the basis, we extracted the three corresponding ultrasound frames for further analysis.

In the acoustic analysis, the closure duration was measured. In the ultrasound analysis, we tracked the tongue contour within the closure interval, using EdgeTrack. Within the interval, three ultrasound frames, representing the beginning, the middle and the end of the closure interval, were analyzed. In each of these frames, three points along the tongue contour were annotated: (a) the lowest point along the contour in the front, (b) the lowest point along the contour in the back, (c) the highest point along the tongue contour. Using these three points, we calculated the intersecting angle (herein called pivot angle) between the line AC and the line BC. (see Figure 2).

![Figure 1a: Annotation of the closure interval of a singleton onset k.](image)

![Figure 2b: Annotation of the closure interval in a homorganic k-k sequence.](image)

2.2. Results

To address the question of whether closure duration served as a cue to distinguish the singleton stop consonant from the stop-stop sequence, we conducted 3 planned paired t-tests with adjusted alpha value of $p = 0.016$.

First, we compared the acoustic closure durations of the singleton stops in both onset (e.g. /tm/ɪn/) and coda (e.g. /tjm/) positions. A significant durational difference was found between the onset and coda stops ($t = 4.15$, df = 13, $p = 0.001$, 2-tailed). The average closure duration of the plosive in the onset position was 67 ms (SD = 20). The average closure duration in the coda position was 105 ms (SD = 36).

We also compared the duration of the singleton onset stop to that of the stop-stop sequence (i.e., /t # k/ and /k # u/) and found a significant durational difference between them ($t = -6.007$, df = 15, $p < 0.0001$, 2-tailed). The average duration of the onset stop consonant was 67 ms (SD = 19) and that of stop-stop sequence was 129 ms (SD = 28). However, when we compared the duration of the singleton coda stop to that of the stop-stop sequence, there was no significant durational difference between them ($t = -1.714$, df = 13, $p = 0.11$, 2-tailed).

To address the question whether the place of articulation of the second stop in the stop-stop sequence would affect the closure duration, we conducted two one-way ANOVAs for the /l/stop sequence and the /k/stop sequence respectively. There were no significant closure durational difference across different places of articulation in the second stop (/l/stop sequence: $F = 1.677$, df = 2, 19, $p = 0.216$; /k/stop sequence: $F = 0.079$, df = 2, 19, $p = 0.925$), as seen in Table 1.

<p>| Table 1: Mean closure durations and standard deviations of the hetero-syllabic stop-stop sequences. |
|-----------------------------------------------------|---------------------------------|</p>
<table>
<thead>
<tr>
<th>Mean closure duration (ms)</th>
<th>Standard deviation</th>
</tr>
</thead>
</table>

462
As predicted, the acoustic closure duration could help differentiate a singleton stop consonant from a stop-stop sequence in Cantonese. This acoustic distinction was not affected by the place of articulation of the second stop consonant in the stop-stop sequence. Interestingly, the acoustic closure duration of the singleton coda stop did not differ from that of the coda stop in a stop-stop sequence, although the closure duration of the singleton onset stop differed from that of the coda stop in the stop-stop sequence. This suggests that the durational difference between the singleton stop and the stop-stop sequence could also be attributed to the durational difference between the onset stop vs. the coda stop.

To address the questions about articulatory closures in the unreleased coda, we preliminarily examined the ultrasound recordings of 1 subject (F).

First, we investigated whether a singleton stop consonant differed articulatorily in the onset and coda positions. Using the pivot angle as the dependent variable, we observed no significant difference between the onset and coda /t/ (t = 1.715, df = 11, p = 0.114). There was also no significant difference between the onset and coda /k/ (t = -0.866, df = 8, p = 0.411). That is, the overall tongue shape during the closure of the stop consonant did not differ between the onset and the coda positions, indicating the presence of articulatory closure of the unreleased stop consonant in the coda position just like that in the onset position.

The overall mean angle for the onset /t/ closure was 132 degrees (SD = 4.4) and that for the coda /t/ closure was 134 degrees (SD = 8.2). The overall mean angle for the onset /k/ closure was 113 degrees (SD = 4.4) and that for the coda /k/ was 116 degrees (SD = 3.2). In short, the pivot angle was more obtuse during the /t/ closure than the /k/ closure, suggesting that the front of the tongue was raised in the former.

Second, addressing the question about co-articulation of the heterorganic stop-stop sequence, we compared the homorganic /t # t/ sequence to the heterorganic /t # k/ sequence. The prediction was that there would be a difference in the pivot angle in the first /t/ if the second consonant /k/ was anticipated in articulation. A paired t-test showed a significant difference between the homorganic and the heterorganic stop-stop sequence (t = 7.479, df = 11, p < 0.0001). Similarly, a separate paired t-test was conducted for the /k # k/ and /k # t/ sequences. We expected a difference in the pivot angle of /k/ between the homorganic and the heterorganic stop-stop sequences, if /h/ was anticipated during the /k/ closure. A significant difference was also found between these two sequences (t = -5.403, df = 11, p < 0.0001).

As predicted, the pivot angle was less obtuse in the heterorganic /t # k/ sequence (120 degrees) than the homorganic one (133 degrees), because the tongue body was raised to anticipate /k/ during the closure of /h/. However, the pivot angle was more obtuse in a heterorganic /k # t/ sequence (126 degrees) than the homorganic one (116 degrees), because the tongue tip was raised to anticipate the second /t/ during the /k/ closure.

3. Discussion and conclusion

In summary, the current study provided some preliminary evidence for the use of the acoustic closure duration in Cantonese to distinguish the singleton onset stop from coda stop. The closure duration, however, did not seem to signal the distinction between a singleton coda stop from a stop-stop sequence. It is also interesting to note that the second stop consonant did not affect the closure duration of the first stop consonant in the stop-stop sequence, suggesting minimal co-articulatory influence from the second stop consonant. Yet, the preliminary ultrasound analysis of one subject showed that the place of articulation of the second stop consonant was anticipated during the closure of the first stop consonant, indicating co-articulation in the heterorganic stop-stop sequence. These preliminary findings seem to suggest that co-articulation and closure duration might not be correlated.

4. References


Gestural overlap within word medial stop-stop sequences in Moroccan Arabic

Chakir Zeroual1,2, Philip Hoole3, Adamantios I. Gafos4 and John H. Esling5

1Faculté Polydisciplinaire de Taza, Morocco; 2Laboratoire de Phonétique et Phonologie CNRS-UMR-7018 & Sorbonne Nouvelle, Paris-France; 3Institute of phonetics and speech processing, Munich, Germany; 4University of Potsdam, Germany; 5University of Victoria, Canada

chakirzeroual@yahoo.fr, hoole@phonetik.uni-muenchen.de, gafos@uni-potsdam.de, esling@uvic.ca

Abstract

This paper deals with the factors behind temporal overlap differences in word medial [bd, db, bg, gb, gd, dg] clusters produced by speakers of Moroccan Arabic (MA). It is argued that certain overlap differences in these MA clusters are related to motor constraints. Specifically, these differences can be attributed to intrinsic physiological properties of the main articulator of each consonant or to constraints in coordinating adjacent consonants due to (presence or absence of) the biomechanical linkage of their articulators. The reported effects weaken the explanatory power of the previously proposed perceptual basis for the place order effect.

Keywords: EMA, gestural overlap, stop-stop clusters, gestural coordination, speech production.

1. Introduction

This paper addresses perceptual and physiological factors behind the temporal overlap differences between [bd, db, bg, gb, gd, dg] produced within Moroccan Arabic (MA) items. Our focus is on stop-stop combinations where no categorical place (nor voicing) assimilation have been previously reported.

Previous studies (e.g. Chitoran et al. 2002) have shown that the amount of intergestural overlap in the C1C2 stop-stop sequences is greater in word medial than word initial position (the so-called Word Position Effect), and when C1 has a more anterior place of articulation than C2 compared to the posterior (posterior) order (the so-called: Place Order Effect). These two effects are generally attributed to perceptual factors. Since word initial position is crucial for lexical access (Marslen-Wilson 1987), less overlap is expected in this position to enhance the recoverability of the stop-stop sequences. A relatively higher degree of overlap within C1C2 stop-stop sequences is expected in the ant-post in the ant-post order, since C1 release can still be perceived only in the first order improving its recoverability.

More recent studies have shown place order effects (POE) even between plosive/non-plosive sequences. Kühnert et al. (2006) reported more overlap in [pl, fl] and [pn, fn] compared to [kl] and [kn] respectively. According to these authors, this POE between stop-liquid and stop-nasal clusters seems to be "due to low-level motor constraints rather than considerations of perceptual recoverability". Indeed in [pl, pn], the tongue tip is free to move during C1 without significantly influencing its articulatory and acoustic properties, while in [kl, kn], anticipation of tongue tip/blade movement for [l, n] is antagonistic with the dorsal articulation of [k]. Notice that in [kp] a low degree of overlap has been reported (Kochetov et al. 2007); this result is not predicted by Kühnert et al.’s (2006) physiological hypothesis, since in [kp] the two consonants involve non-connected articulators. Chitoran et al. (2006), also reported more overlap in ant-post [pl, pr] than in post-ant [kl, kr] respectively. This pattern suggests, according to these authors, that POE "may be a lexically specified pattern" independent of the substance of the phonological forms.

Chitoran et al. (2002) reported an articulatory regularity that they consider as an "unpredicted effect of place combination". They observed "that combinations of labial and coronal stops are the least overlapped" compared to lab-dor and cor-dor. This regularity is also observed and discussed in our present study.

Previous physiological studies have also shown that the overlap can vary across languages. Based on acoustic data, Zsiga (2000) observed that stop-stop clusters produced across word boundary (C#C) are more often released in Russian than in English, suggesting less overlap in the former than in the latter. Recently, Gao et al. (2011) reported more coproduction within [p#k] and especially [p#t] sequences in Taiwanese than in English. These cross-linguistic articulatory differences show that languages exhibit specific coordination patterns. According to Kochetov et al. (2007), language-particular differences in “the degree of overlap may be related to the propensity of a language to assimilate in consonant clusters”.

More precisely, they propose to relate the higher degree of overlap in Korean than in Russian to the presence of a place assimilation only in the former.

Existing cross-linguistic articulatory studies are not exhaustive enough to provide irrefutable evidence for potential relations between cross-linguistic asymmetrical place assimilations and some parallel spatio-temporal patterns. It is also not clear if regressive place assimilation within C1C2 sequences is due to the overlap of C1 by C2 (Browman and Goldstein, 1992; Son et al. 2007), (spatial or temporal) reduction of C1 (Jun, 1996), or a combination of the two factors. Since regressive place assimilations are more common than progressive ones (Steriade, 2001), our articulatory measurements will quantify the amount of C2 anticipation during C1 in stop-stop sequences. We examine whether a coronal C1 is more overlapped than labial and dorsal, since cross-linguistically the former tends to undergo regressive place assimilation more frequently than the latter do (Jun, 2004; Steriade, 2001). We will also check whether dorsal C2 is more anticipated than labial and coronal, since the former triggers regressive place assimilation more than the latter ones (Jun, 2004).

In a recent articulatory study on MA, Gafos et al. (2010) find that heterorganic stop-stop sequences of MA are almost always produced with open-transition or released C1, suggesting less overlap. This characteristic may be related to the absence of categorical place (or voicing) assimilation between radical intervocalic consonants in MA.

In this paper, the main focus is on the potential effect of low-level motor constraints on temporal coordinations. For this reason, we focus our discussions on MA [bd, db, bg, gb, dg, gb], produced word medially where a relatively great overlap and no categorical place (or voicing) assimilation are expected. If some gestural overlap differences are
physiologically based, they should still be present between these MA sequences.

2. Method

2.1. Subjects and stimuli

The stimuli of this study are selected from two large corpora recorded separately, with the 3-dimensional EMA technique (AG500 Carstens Medizinelektronik, Hoole et al. 2010), to test several hypotheses. Only items with intervocalic [bd, db, bg, gb, gd, dg] (table 1) pronounced 5 times by 2 speakers (S4, S5) in [galha__hnaja] (‘he told her (it) _ here’) from a first recording and [bd, db, bg, gb, gd] pronounced 8 times by 3 other speakers (S1, S2, S3) in /3ibi _ hnaja/ (‘bring _ here’) from a second are analyzed here. Our talkers are MA native speakers aged between 29 and 40 years and with no known history of speech or hearing disorders.

The items have the same morphological form /CaCC+a/ where the radical is the active participle and /+a/ 3prs fem sg object suffix. The lexical accent is on the first vowel. We chose aCCa where the movements of lingual and labial gestures are clearly identified and coarticulate relatively weakly with the vowels.

<table>
<thead>
<tr>
<th>Stimuli + glosses</th>
<th>[CC]</th>
<th>Stimuli + glosses</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3abda] 'to pull'</td>
<td>[bd]</td>
<td>[kada} 'to lie'</td>
</tr>
<tr>
<td>[sabga] 'to be ahead'</td>
<td>[bg]</td>
<td>[ragba] 'to appear'</td>
</tr>
<tr>
<td>[fadga] 'to crack'</td>
<td>[dg]</td>
<td>[ragda] 'to sleep'</td>
</tr>
</tbody>
</table>

Table 1: List of stimuli pronounced by our 5 speakers during two EMA recordings. All the items are the same except [kada] pronounced in the first recording by S4 and S5 (5 tokens) instead of [nudba] in the second by S1, S2 and S3 (8 tokens).

2.2. Analysis

This 3-dimensional EMA technique enabled us to track movements of the tongue tip (TTIP), blade (TB), dorsum (TBACK), upper and lower lips (LLIP), and the jaw with sensors placed on these articulators (sample rate 200Hz). The Mview program (developed initially by M. Tiede from Haskins Laboratories) permits to display the spatio-temporal coordinates of the vertical and horizontal movements of each articulator, as well as the evolution of its velocity and acceleration. For each gesture, several landmarks have been identified automatically from the velocity trace (20% threshold) of its opening and closing movements (Fig. 1).

Several spatio-temporal and kinematic measurements were extracted automatically on TTIPOSy, TBACKsy and LLIPsy traces (Fig. 1): (i) Temporal and spatial coordinates at onset, target, maximal constriction, release and offset positions. (ii) The peak velocity and the amplitude of the closing and the opening vertical movements of C1 and C2 gestures.

Based on these parameters, we calculate the degree of gestural overlap within our consonant sequences using the formula given in Fig. 2.

3. Results and discussion

Single factor ANOVA tests run on the data of each subject, show that the degree of overlap varies significantly with the cluster type (table 2). Post-hoc analyses (table 3, Fig. 3) will be used for more detailed comparisons. Additional statistical analyses will also be presented below to quantify the potential contribution of some other factors, especially physiological ones, to these overlap differences.

Table 2. Mean values (and standard deviations) of gestural overlap within MA voiced stop-stop.

This 3-dimensional EMA technique enabled us to track movements of the tongue tip (TTIP), blade (TB), dorsum (TBACK), upper and lower lips (LLIP), and the jaw with sensors placed on these articulators (sample rate 200Hz). The Mview program (developed initially by M. Tiede from Haskins Laboratories) permits to display the spatio-temporal coordinates of the vertical and horizontal movements of each articulator, as well as the evolution of its velocity and acceleration. For each gesture, several landmarks have been identified automatically from the velocity trace (20% threshold) of its opening and closing movements (Fig. 1).

Several spatio-temporal and kinematic measurements were extracted automatically on TTIPOSy, TBACKsy and LLIPsy traces (Fig. 1): (i) Temporal and spatial coordinates at onset, target, maximal constriction, release and offset positions. (ii) The peak velocity and the amplitude of the closing and the opening vertical movements of C1 and C2 gestures.

Based on these parameters, we calculate the degree of gestural overlap within our consonant sequences using the formula given in Fig. 2.

3.1. Place order and place combination effects

A two-way ANOVA over all subjects in a repeated measure model shows that the degree of overlap varies significantly with place order (ant-post, post-ant: [df=1, F=20.44, p=0.0106]) and place combinations (lab-cor, cor-dor, lab-dor; [df=18.89, p<0.0001]), with a significant interaction [df=2, F=4.80, p=0.04].

465
Table 3. Gestural overlap (mean differences) comparisons between place combinations done on the data of each speaker: ***=p<0.001; **=p<0.01; *=p<0.5; ns=not significant).

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.12</td>
<td>1.91</td>
<td>2.47</td>
<td>-1.72</td>
<td>-1.84</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>S2</td>
<td>0.50</td>
<td>2.26</td>
<td>2.49</td>
<td>-1.89</td>
<td>-2.33</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>S3</td>
<td>0.36</td>
<td>0.73</td>
<td>0.24</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>***</td>
<td>ns</td>
<td>ns</td>
<td>***</td>
</tr>
<tr>
<td>S4</td>
<td>0.63</td>
<td>1.04</td>
<td>2.05</td>
<td>-0.29</td>
<td>-2.13</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>S5</td>
<td>0.61</td>
<td>0.19</td>
<td>1.15</td>
<td>0.42</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

Post-hoc analyses of the separate two-way ANOVA tests on the data of each speaker show significantly more overlap in ant-post than post-ant sequences for all subjects (table 4). This result is consistent with previous studies (see introduction).

Table 4. Gestural overlap (mean differences) between voiced stop-stop clusters classified by place order (data of each speaker): ***=p<0.001; **=p<0.01; *=p<0.5; ns=not significant).

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Ant-post vs post-ant</th>
<th>Post-hoc comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1.43</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>S2</td>
<td>1.61</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>S3</td>
<td>0.42</td>
<td>0.0003</td>
</tr>
<tr>
<td>S4</td>
<td>1.24</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>S5</td>
<td>0.72</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

Table 5. Gestural overlap (mean differences) comparisons between voiced stop-stop clusters classified by place combinations (data of each speaker): ***=p<0.001; **=p<0.01; *=p<0.5; ns=not significant).

<table>
<thead>
<tr>
<th>Ant-post</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>lab-cor vs. lab-dor</td>
<td>-0.71</td>
<td>***</td>
<td>-1.00</td>
<td>***</td>
<td>-0.47</td>
</tr>
<tr>
<td>lab-cor vs. dor-cor</td>
<td>-0.66</td>
<td>***</td>
<td>-0.54</td>
<td>***</td>
<td>-0.60</td>
</tr>
<tr>
<td>lab-dor vs. dor-cor</td>
<td>0.05</td>
<td>ns</td>
<td>0.46</td>
<td>ns</td>
<td>-0.13</td>
</tr>
</tbody>
</table>

Figure 3: Degree of overlap differences between ant-post [bd, db, bg, gd] sequences and their post-ant [gb, gd] corresponded pronounced by 5 MA speakers.

3.2. C1 and C2 place effects

Table 6. Mean values of the degree of overlap within C1/C2 stop-stop sequences pronounced by 5 MA speakers and classified by the place of articulation of C1 and C2. Statistical comparisons are also given.

<table>
<thead>
<tr>
<th></th>
<th>[d]</th>
<th>[b]</th>
<th>[g]</th>
<th>[d vs b]</th>
<th>[g vs b]</th>
<th>[g vs d]</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>1.90</td>
<td>1.71</td>
<td>0.50</td>
<td>ns</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>S2</td>
<td>0.96</td>
<td>1.39</td>
<td>0.03</td>
<td>ns</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>S3</td>
<td>0.47</td>
<td>0.75</td>
<td>0.57</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>S4</td>
<td>1.59</td>
<td>1.30</td>
<td>0.50</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>S5</td>
<td>1.11</td>
<td>1.17</td>
<td>1.10</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>0.60</td>
<td>0.23</td>
<td>1.06</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>S4</td>
<td>0.88</td>
<td>0.46</td>
<td>2.05</td>
<td>ns</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>S5</td>
<td>0.83</td>
<td>0.70</td>
<td>1.75</td>
<td>ns</td>
<td>***</td>
<td>***</td>
</tr>
</tbody>
</table>
Separate ANOVA tests, show that the overlap varies significantly with C1 place of articulation for three speakers (S1: [F(2, 45)=10.95, p=0.0001; S2: [F(2,38)=7.78, p=0.0015, and S4: [F(2, 27)=4.03, p=0.03). Post-hoc analyses (Table 6) confirm that for these subjects the overlap of C1 by C2 is significantly greater when C1 is a labial or coronal ([b vs d] not significant) than when it is a velar ([g vs b] and [g vs d] significant). Table 3 also shows that, for almost all our subjects, while [bd] vs [bg] and [d] vs [dg] overlap differences are negative and significant, [gb] vs [gd] are not significant. These results are parallel to cross-linguistic asymmetrical patterns of regressive place assimilations (velar less prone to undergo place assimilation), suggesting that these assimilation patterns and the degree of overlap differences may be connected. Separate ANOVA tests confirm that for all speakers the degree of overlap within C1C2 stop-stop sequences also relies on the C2 place of articulation (p<0.001). Post-hoc analyses (table 6) show that the anticipation of C2 during C1 is more substantial when the former is a velar consonant ([g vs b] and [g vs d] highly different for all subjects) and lower when it is coronal or labial ([b vs d] non-significant for four subjects). These results also agree with patterns of asymmetrical place assimilations (velar more prone to trigger place assimilation). Notice that Gao et al. (2011) also reported that [pk] has in English a greater overlap than [pt] suggesting more anticipation of C2 in [pk] sequence than [pt].

4. Conclusion
Moroccan Arabic shows, within its word medial [bd, db, dg, bg, gb, gd], overlap differences that are in the same direction as the asymmetrical patterns of regressive place assimilations observed cross-linguistically even though this language does not exhibit categorical place assimilation in this context. This result supports the hypothesis that phonological assimilation has a basis in temporal overlap patterns (Ohala, 1990). To our knowledge, evidence for this correlation between cross-linguistic assimilation patterns and overlap patterns is documented here for the first time in a language with open transitions in clusters. Our present study shows that several overlap pattern variations within MA stop-stop clusters are related to motor constraints (biomechanical constraints). These physiological explanations seem to weaken the explanatory power of the previously proposed perceptual basis for the place order effect. More articulatory, acoustic and perceptual research is needed to establish a complete picture of the full interactions between physiologically, perceptual and grammatical factors potentially responsible for the spatio-temporal coordination differences.

5. Acknowledgments
This work was supported by German Research Council grant HO 3271/3-1 to Philip Hoole and by NSF grant 0922437 to Adamantios Gafos.

6. References
Development of lingual coarticulation and articulatory constraints between childhood and adolescence: an ultrasound study

Natalia Zharkova, Robin J. Lickley, William J. Hardcastle

Queen Margaret University, Edinburgh, UK

nzhrarkova@qmu.ac.uk, rlickley@qmu.ac.uk, wjhardcastle@gmail.com

Abstract

This study used ultrasound tongue imaging in order to compare lingual coarticulatory patterns in 13-year-old adolescents and 5-year-old children. Previous articulatory work has reported age-related differences in consonant-specific coarticulation. According to the Degree of Articulatory Constraint model, speech segments with more constrained tongue dorsum have more resistance to lingual coarticulation in adult speech. Bilabial and postalveolar consonants are at opposing ends of a spectrum, with the latter being strongly resistant to vowel-related lingual coarticulation. In this study, tongue shapes for the bilabial stop and the postalveolar fricative were compared across contrasting vowel contexts. For both groups of speakers, there were robust differences between the two consonants, with the bilabial adapting to the vowel influence more than the postalveolar. We conclude that the DAC model can predict some lingual coarticulatory patterns in children as well as in adolescents.

Keywords: lingual coarticulation, articulatory constraints, children, adolescents, development

1. Introduction

Evidence from a number of acoustic and articulatory studies has shown that lingual coarticulation is present in children from an early age (e.g., Sereno et al. 1987; Nittrouer et al. 1989; Katz et al. 1991; Sussman et al. 1992; Zharkova et al. 2011; 2012; Noiray et al. 2013; Song et al. 2013). An influential theory of coarticulation development (Nittrouer et al. 1989) claims that motor planning of speech in young children is carried out at the syllabic level. This idea is based on the observed reduction with age of the extent of anticipatory vowel-on-consonant coarticulation within a consonant-vowel (CV) syllable, as shown by acoustic analysis. This theoretical approach contrasts with the view that children coarticulate less than adults, due to the overly segmental style of speech planning in young children (Kent 1983). Acoustic studies on the topic have produced conflicting results (e.g., no significant differences in lingual coarticulation between children and adults were reported in Katz et al. 1991). There is some support for Nittrouer’s theory from articulatory research (e.g., Zharkova et al. 2011), though previous articulatory work has also reported age-related differences in consonant-specific coarticulation (Katz & Bharadwaj 2001; Zharkova et al. 2012; in press). In the present study, the claim by Nittrouer et al. was explored in the context of an approach to lingual coarticulation based on articulatory constraints. According to the Degree of Articulatory Constraint (DAC) model (Recasens et al. 1997), speech segments with more constrained tongue dorsum have more resistance to lingual coarticulation. Bilabial and postalveolar consonants are at opposing ends of a spectrum, with the latter being strongly resistant to vowel-related lingual coarticulation. Our study used ultrasound tongue imaging to compare lingual coarticulatory patterns in 13-year-old adolescents and 5-year-old children, using two consonants which differ substantially in their DAC properties.

In order to analyse articulatory data from our 5-year-old speakers numerically, we needed a way of quantification that would not be based on a set of tongue curves located in the same coordinate space. The 13-year-olds in our study wore a custom-designed headset (Articulate Instruments Ltd 2008), to stabilise the transducer in relation to the head. This headset was not suitable for the 5-year-old participants. A way to obtain quantitative information in the absence of transducer stabilisation is to use measurements based on a single curve (e.g., Bressmann et al. 2005; Ménard et al. 2012; Stolar & Gick 2013). In order to have a reliable way of comparing between groups, our study required a measure that would provide the same results regardless of whether the head was stabilised or not. To quantify lingual coarticulation in this study, we used the Tongue Constraint Position Index (TCPI), an index of tongue shape developed in Zharkova (2013a). TCPI represents the location of the most “bunched” part of the tongue, with its negative or positive values occurring when the place of maximal excursion of the tongue is further backward or further forward along the tongue contour, respectively, in relation to the middle of the straight line between two ends of the curve. An index value can be obtained from a single tongue curve, which makes it possible to use on the data collected without head stabilisation. The TCPI has been shown to capture vowel-induced lingual coarticulation on consonants (Zharkova 2013b), and to be unaffected by whether participants wear the headset or not (Zharkova & Hewlett 2013).

In both age groups, higher TCPI values were expected in the context of /i/ than in the context of /a/. In the adolescents, more coarticulation was expected in /p/ than in /f/, consistent with the DAC model predictions for adult speech. If more coarticulation was observed in /p/ than in /f/ for the children, then the findings would also be consistent with the DAC model. If the children had more coarticulation than the adolescents, this would provide support for the theory by Nittrouer et al. If the relative difference in amount of coarticulation between the two consonants did not pattern in the same way in the two age groups then we could conclude that the effect of articulatory constraints on lingual coarticulation is not uniform throughout the developmental process from childhood to adolescence.

2. Method

The study used ultrasound tongue movement data (sampling rate 100 Hz) synchronised with the acoustic signal. CV syllables with the consonants /p/ or /f/ and the vowels /a/ or /i/ were produced in the carrier phrase “It’s a ..., Pam” (each target repeated five times) by two groups of speakers of
Scottish Standard English. One group consisted of ten adolescents aged between 13 years 0 months and 13 years 11 months. The other group consisted of ten children aged between 5 years 0 months and 5 years 11 months. The adolescent participants wore a headset to stabilise the ultrasound transducer in relation to the head. For the child participants, the transducer was hand-held by the experimenter. Articulate Assistant Advanced software (Articulate Instruments Ltd 2012) was used to collect and analyse the data. All participants from the younger group were video recorded during the data collection using a separate channel of the multichannel ultrasound system. The video data were collected in two planes: en face and in profile. In the data analysis, all recordings from 5-year-olds were examined in order to ensure that during the target CV sequence the transducer was relatively stable under the chin, and that a midsagittal tongue image was present. The tokens which did not satisfy these conditions were excluded from the analysis (nine tokens in total). Additionally, the /f/ tokens produced by one child (Child 7) were realised as unaspirated dental stops. As the measurements in this study were focused on tongue shape, and the dental stops may have a noticeably different tongue shape from the postalveolar fricative, all /f/ tokens produced by this child were excluded from all analyses.

2.1. Data analysis

In each CV token, an annotation point was placed at mid-consonant (mid-closure for /p/), based on the acoustic data. Cubic splines were fitted to midsagittal tongue curves at each annotation point. The TCPI, which is a ratio, was calculated for each token. The denominator in the ratio equalled half the length of the straight line n between two ends of the tongue curve. The numerator was the distance between the longest perpendicular from n to the tongue curve and the perpendicular to the tongue curve traced from the mid-point of n.

Linear mixed models (LMMs) were performed in R (R Development Core Team 2012), with the lmer software package (Baayen 2008). In order to establish whether the contrasting vowels influenced the tongue shape at mid-consonant, a LMM was carried out for each age group, with vowel as a fixed factor. To find out which consonant was more affected by the following vowels, a LMM was run for each age group with vowel and consonant as fixed factors. In this analysis, a significant interaction between consonant and vowel would mean that the two consonants have different amounts of coarticulation. To establish whether the 5-year-olds coarticulate more than the 13-year-olds, the model included vowel and age as fixed factors. A significant interaction in this model, accompanied by a larger difference between the two vowel contexts in the 5-year-olds than in the 13-year-olds, would be taken as evidence of more coarticulation in the 5-year-olds. Finally, a LMM was run with vowel, consonant and age group as fixed factors. A significant interaction between the three factors would suggest that articulatory constraints on the tongue dorsum affect coarticulatory properties in young children differently to adolescents. In all statistical tests, Speaker was modelled with both random slopes and random intercepts. Following Zharkova et al. (in press), the outcome was deemed significant at the 0.01 level if the F value in the ANOVA exceeded 8.49.

3. Results

Figure 1 shows tongue contours from representative speakers from each age group. It is immediately noticeable that the tongue curves for the 5-year-old child are much more variable in absolute position across repetitions than in the adolescent. This difference was expected, given the recording setup for the 5-year-olds, with the handheld transducer. Despite this large variability in absolute tongue position, the tongue shapes for /p/ in the two vowel contexts are clearly different for the 5-year-old in Figure 1, with the bunched shapes in the context of /i/, and the “hump” located towards the anterior part of the tongue. In the context of /a/, the tongue shapes for the 5-year-old’s /p/ are noticeably flatter than in the context of /i/, reflecting the tongue dorsum lowering for producing the low vowel. Similar patterns are observed for the adolescent. The fact that all the curves for the adolescent are located within one coordinate plane, due to the head-to-transducer stabilisation, lets us see that the absolute tongue root and tongue blade positions for the two vowel contexts are quite different. The shape of the tongue, regardless of the absolute position of the tongue curves, is different across the two vowel contexts, with the “hump” location appearing more anterior in the context of /i/ and more posterior in the context of /a/.

For /f/, for both speakers and in both vowel contexts the “hump” is located towards the anterior part of the tongue, so the tongue shapes for /f/ are resembling those for /p/ in the context of /i/ more than of those for /p/ in the context of /a/.

Mean values of the TCPI for individual speakers are presented in Table 1. For /p/, the difference in tongue shape across the two vowel contexts, noticeable in Figure 1, was observed across speakers and age groups. All adolescent and child speakers had a larger TCPI value for /p/ in the context of /i/ than in the context of /a/. For /f/, consistent with the observation from Figure 1 of rather similar tongue shapes in the two vowel contexts, only seven adolescents and five 5-year-olds had a larger TCPI value in the context of /i/.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Cons.</th>
<th>13-year-olds</th>
<th>5-year-olds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/p/</td>
<td>-0.24</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>/f/</td>
<td>0.25</td>
<td>0.27</td>
</tr>
<tr>
<td>2</td>
<td>/p/</td>
<td>-0.17</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>/f/</td>
<td>0.43</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>/p/</td>
<td>-0.36</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>/f/</td>
<td>0.30</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>/p/</td>
<td>0.20</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>/f/</td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>5</td>
<td>/p/</td>
<td>0.15</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>/f/</td>
<td>0.32</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Figure 1: Tongue curves in the two vowel contexts (solid lines for the context of /a/; dashed lines for the context of /i/).
3.1. Within-group and across-group comparisons

Figure 2 presents TCPI values for each age group, consonant and vowel context. The TCPI values for /p/ in the context of /a/ stand out, for both groups, in that they are visibly smaller than all the other values. This is consistent with the difference in tongue shapes that can be observed in Figure 1.

<table>
<thead>
<tr>
<th></th>
<th>/p/</th>
<th>/i/</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.14 0.17 0.20</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.28 0.33 0.29</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.21 0.23 0.23</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.22 0.24 0.24</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.16 0.20</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: TCPI values (the boxes for /i/ are shaded in grey). The bottom, central and top lines of each box represent the 25th, 50th and 75th percentiles, respectively; the small circles are outliers.

In the analysis of the adolescent data, there was a significant effect of vowel on TCPI ($F = 31.45$), showing evidence of lingual coarticulation at mid-consonant. In the model including both vowel and consonant as fixed factors, there was a significant interaction ($F = 74.72$), with the difference in TCPI values between the context of /a/ and the context of /i/ being larger for the stop (see Figure 2). This interaction provides evidence of the adolescents having adult-like coarticulatory patterns, with the more constrained postalveolar fricative showing less coarticulation than the less constrained bilabial stop. In individual models for the two consonants separately, the effect of the vowel context on TCPI at mid-consonant was significant for /p/ ($F = 61.41$), but not for /i/ ($F = 2.11$).

For the five-year-olds, the vowel context had a significant effect on TCPI at mid-consonant ($F = 15.96$). When both vowel and consonant were included as fixed factors, there was a significant interaction ($F = 28.64$), with the difference in TCPI values between the two vowel contexts being larger for /p/, similarly to the adolescents (see Figure 2). In individual models for the two consonants, again similarly to the adolescents, the effect of the vowel context on TCPI was significant for /p/ ($F = 32.96$), but not for /i/ ($F = 0.07$).

There was no significant interaction in the model including vowel and age group as fixed factors, for both consonants pooled together ($F = 0.95$), or for individual consonants (/p/: $F = 1.85$; /i/: $F = 0.26$). In the model including vowel, consonant and age group as fixed factors, there was no three-way interaction ($F = 1.41$).

4. Discussion

This study showed that for both 13-year-old adolescents and 5-year-old children, contrasting vowels /a/ and /i/ have an effect on the tongue shape at mid-consonant. These results agree with previous findings of lingual coarticulation in young children (Katz et al. 1991; Sussman et al. 1992; Noiray et al. 2013). We have also reported robust differences in coarticulatory patterns across consonants for the adolescents and for the younger children. The similarity of tongue shapes for /i/ across vowel contexts was clear in both groups of speakers, while for /p/, the difference in tongue shapes across vowel contexts was equally clear for both groups.

The difference for the adolescents between the two consonants in the amount of tongue shape adaptation to the following vowel was consistent with the DAC model predictions, with the bilabial stop being significantly more affected than the postalveolar fricative. For the 5-year-olds, the difference between the two consonants in the amount of tongue shape adaptation to the vowel was in the same direction, and also significant. Thus we can conclude that the DAC model, as far as concerning the difference in articulatory constraint on the tongue between bilabials and postalveolars, can predict lingual coarticulatory behavior in 5-year-old children.

One possible reason for the lack of any significant differences between the two age groups is that there was much more variation in absolute tongue positions for the 5-year-olds than for the 13-year-olds, because of the hand-held transducer. In order to find out whether the difference in the recording setup could have been a contributing factor to the across-group results, TCPI was compared across age group and vowel context for /p/, using the data for the 5-year-olds and the data for the same 13-year-old speakers collected without the headset. Those data were collected from the 13-year-olds within a larger project in order to investigate how the presence or absence of head-to-transducer stabilisation may affect quantitative measures of tongue shape and position (see Zharkova & Hewlett 2013). The results of the across age group comparison were the same as the results reported in Section 3.1, with no significant interaction between vowel and age group ($F = 0.33$).

In addition to the lack of significant across-group differences in the extent of lingual coarticulation, there was no significant effect of the vowel context on tongue shape for the postalveolar consonant. This does not necessarily contradict previous studies reporting a decrease of the extent of coarticulation with increasing age, such as an acoustic study by Nitrourer et al. (1996) and an ultrasound study by Zharkova et al. (2011). In the former study, the measurement point was located later in the consonant than in the present study, so at mid-consonant any age-related difference may not have developed yet. The latter study used a different method to ascertain the presence of coarticulation, the nearest neighbour distance method of comparing two sets of tongue curves (Zharkova & Hewlett 2009). It is possible that the TCPI index is not sufficiently robust to detect some coarticulation-related differences that can be captured using measurements based on the data from whole curves and/or measurements comparing...
absolute positions of different curves. For example, a difference for the preadolescent /i/ in the tongue root area between tongue curves in the two vowel contexts (as indicated in Figure 1) could be quantified by a measure specifically targeting the tongue root. This difference would also be likely to show in an analysis based on comparing sets of tongue curves which uses information from the whole of the visible tongue contour (e.g. the nearest neighbour distance method, or SSAnova, Davidson 2006). Indeed, Zharkova (2013c) has compared /i/ curves from these preadolescent speakers using the nearest neighbour method, and reported a significant difference between the curve sets for /i/ in the two vowel contexts. While such measures have a number of advantages over single curve based measures (they are not based on the visible ends of the tongue curve, and they make use of the absolute location of a set of curves), they are not applicable to the data from multiple tongue curves which are not located within the same coordinate space, hence in the present study we used a method that could provide us with robust results on tongue shape in the absence of head-to-transducer stabilisation. We are currently investigating whether other single curve based measures, which can potentially capture more relevant aspects of the tongue shape, can be reliably used to analyse lingual articulation regardless of whether the speaker is wearing a transducer stabilising headset.

5. Acknowledgements

This work was supported by a grant from the Economic and Social Research Council (ES/K002597/1).

6. References


471
How do voiceless fricatives contribute to intended intonation? A comparison of whispered speech, semi-whispered and normal speech.

Marzena Żygis, Daniel Pape, Luis M.T. Jesus, Marek Jaskula

1 Centre for General Linguistics (ZAS) & Humboldt University, Berlin, Germany
2 EEETA, University of Aveiro, Portugal
3 ESSUA, University of Aveiro, Portugal
4 West Pomeranian University of Technology, Szczecin, Poland

zygis@zas.gwu-berlin.de, danielpape@ua.pt, lmt@ua.pt, Marek.Jaskula@zut.edu.pl

Abstract

The overall goal of the paper is to contribute to a better understanding of segment-prosody interaction in different speech modes. In particular, the paper shows that depending on intonation patterns, utterance-final voiceless fricatives in Polish show different properties. In polar questions with rising intonation, the frequency of the highest spectral peak is higher than in statements produced with falling intonation. Voiceless fricatives in questions display higher Centre of Gravity accompanied by higher standard deviation. Furthermore, the fricatives are produced with higher mean intensities in questions than statements.

Differences in fricatives are more pronounced when the sentences are whispered. For instance, in contrast to normal speech where skewness and kurtosis do not show differences, in whispered speech both spectral moments display lower values in statements than in questions. In addition, spectral slopes (m1, m2) are steeper in questions than in statements.

In summary, the paper shows that voiceless fricatives are prone to intonation differences and follow its patterns by means of varying spectral cues inherent to these sounds.

Keywords: voiceless sibilants, intonation, Polish

1. Introduction

Since the fundamental frequency (F0) is the main correlate of intonation and it is only found in voiced segments, voiceless segments are generally not investigated when intonation is investigated. Even more, they are tried to be avoided in experimental stimuli as they are treated as triggers of micro-prosodic perturbations. Despite a common consensus about the perturbing function of voiceless segments in F0 investigations, very little is known of whether the voiceless segments remain immune to intonation or whether they contribute to intonation patterns by their inherent acoustic cues. As have been more recently shown by Niebuhr, Lill & Neuschulz (2011) and Niebuhr (2012) for German, sibilants change their properties depending on intonation patterns; cf. also Grabe (1998) on the role of voiceless segments in production of intonation. This paper investigates the interaction of voiceless fricatives and different intonation patterns in Polish.

It remains unclear to what extent changes in sibilants are found in other speech modes as well, such as whispered and semi-whispered speech, where F0 is completely or partly absent, accordingly. Previous studies about whispered speech concentrated on vocalic segments (e.g. Li & Xu 2005, Heeren & van Heuven 2011), whereas the present paper concentrates on properties of consonantal segments.

2. Experimental evidence

The aim of the experiment is two-fold. First, it will be examined if voiceless clusters in utterance-final positions characterized by different intonation contours display spectral differences in phonated speech. Second, the study is aimed to explore to what extent possible changes in sibilant properties are triggered by different intonation patterns when varying the speech mode, i.e., comparing whispered and semi-whispered speech to phonated speech.

2.1. Experimental design

Our material includes eight monosyllabic words ending in a cluster consisting of a voiceless retroflex fricative followed by a retroflex affricate. Each item was presented in a polar question (‘Widzi ten pl[ta]?’ ‘Does he see the coat?’) and statement (‘Widzi ten pl[ta]’ ‘He sees the coat’). The polar question was expected to be produced with a rising intonation and the statement with a falling intonation. All questions and statements described above were produced in three different speech modes: normal, whispered and semi-whispered and repeated 3 times. The sentences were read by 16 native speakers of Standard Polish (8 male).

All recordings were conducted in the sound-proof room at the Electrical Engineering Department of the West Pomeranian University of Technology in Szczecin using a TLM103 Neumann microphone (20cm distance from lips) connected to a ProTools system with a Digi 001 interface (sample rate 44100 Hz). The items were analysed with Praat (version 5.3.57) and MATLAB (version R2007b). In the following we report results only on fricative sibilants which are based on measurements of 2304 items (8 words × 2 intonation types: rising & falling × 3 speech modes: normal & whispered & semi-whispered × 3 repetitions × 16 speakers).

2.2. Methods

Using multitaper spectra we investigated the following acoustic parameters: frequency of the highest spectral peak in the range from 20Hz to 11kHz, frequency of the highest peak in the range of 2 to 4kHz, mean intensity of the friction noise, spectral Center of Gravity (COG), the COG’s standard deviation, skewness, kurtosis, and the spectral regression lines slopes (low frequency slope m1 and high frequency slope m2); see Jesus & Shadle (2002) and Żygris, Pape & Jesus (2012) for all parameter details. Furthermore, for reference we also measured the difference in F0 between the onset and offset of the preceding vowel.

Regarding statistical analysis, linear mixed effects models were employed for the investigated variables, which were studied as effects of INTONATION TYPE (rising, falling) and
SPEECH MODE (normal, semi-whispered, whispered) as well as their interaction (INTONATION*SPEECH MODE) and GENDER (female, male). In addition, speaker-specific random slopes for INTONATION TYPE and SPEECH MODE were included into the model and ITEM and SPEAKER were taken as random effects. All analyses were conducted in the R environment software (version 3.0.2).

2.3. Results

The results show that differences between questions and statements are clearly encoded in the acoustic characteristics of fricatives.

Before presenting the results in detail, let us stress that regarding normal speech the F0 difference between the vowel offset and onset pointed to raising F0 in questions (female=142 Hz, male=81 Hz) and falling F0 in statements (female=-15 Hz, male=-17 Hz) which confirmed our initial assumption about differences in intonational patterns (t=10.726, p<.0001); for other results concerning acoustic properties of vowels cf. Žygis et al. (to appear).

Figure 1 presents multitaper spectra obtained at the acoustic midpoint of the fricative for all three speech modes, where the black lines represent the statement condition and the lighter colours illustrate questions, cf. also Jesus & Shadle (2002).

![Multitaper spectra of 1145 fricative spectra](image1.png)

Figure 1: Multitaper spectra of 1145 fricative spectra (8 male speakers) obtained at the acoustic midpoint for all speech modes.

As the results reveal, in all three speech modes questions were produced with the higher spectral peak towards higher frequencies when compared to statements (whispered t=2.737 p<.01, semi-whispered t=3.254 p<.001, normal: t=4.885 p<.0001). However, since the spectra obtained for the range 20Hz to11kHz were highly variable, as can be inferred from Figure 1, and the first spectral peak is known to be related to place of articulation of fricatives, we additionally analysed the highest spectral peak in the range from 2 to 4kHz. Figure 2 presents the results for all three speech modes.

![Boxplots for the highest frequency peak in the range from 2 to 4kHz across three speech modes](image2.png)

Figure 2: Boxplots for the highest frequency peak in the range from 2 to 4kHz across three speech modes.

The highest spectral peak found in the range from 2-4kHz was significantly higher in questions than in statements across all three speech modes but, as indicated by t-values, the difference was greatest in whispered speech (t=8.165 p<.0001), followed by normal speech (t=7.050 p<.0001) and semi-whispered speech (t=4.264 p<.0001). Furthermore, regarding spectral slopes m1 and m2, significant differences were found for questions as compared to statements predominantly in whispered speech. In particular, both spectral slopes were significantly steeper for questions than for statements (m1: t=3.349 p<.0001 and m2: t=8.58 p<.0001). In addition, m2 values were significantly higher for questions in semi-whispered speech (t=-2.19 p<.05). Figure 3 provides an illustration of the spectral lines.

![Multitaper spectra (mean plots over all items and all speakers) at the fricative midpoint in whispered, semi-whispered and normal speech mode.](image3.png)

Figure 3: Multitaper spectra (mean plots over all items and all speakers) at the fricative midpoint in whispered, semi-whispered and normal speech mode. Black solid lines correspond to the question and lighter colour to the statement condition. Dotted lines are the spectral regression lines m1 and m2.

Furthermore, the mean intensity of frication noise was higher in questions than in statements, as illustrated in Figure 4. This conclusion holds true for all three speech modes. Again, the difference was most pronounced in whispered speech (t=17.01 p<.0001) followed by normal (t=15.76 p<.0001) and semi-whispered speech mode (t=5.29 p<.0001).

![Boxplots for the mean of frication intensity across three speech modes](image4.png)

Figure 4: Boxplots for the mean of frication intensity across three speech modes.

Spectral moments were also significantly different in questions than in answers varying however across speech modes. Figure 5 presents Centre of Gravity values found at the midpoint of frication.

![Boxplots for the Center of Gravity](image5.png)
Figure 5: Boxplots for Centre of Gravity across three speech modes. The results show that questions were produced with significantly higher COG values than answers in all three speech modes. Again, the highest difference was found in whispered speech (t=5.316 p<.0001) in comparison to normal (t=4.99 p <.0001) and semi-whispered speech mode (t=2.976 p<.001). It should also be noted that COG values were generally higher in normal speech as opposed to whispered and semi-whispered speech mode.

In addition, we also observed a time-dynamic pattern for the COG, being systematically higher at fricative midpoint in comparison to both fricative onset and fricative offset for all speech modes. Again, questions showed significantly higher COG values compared to statements at all three measurement points. Figure 6 illustrates the dynamic COG pattern.

Figure 6: COG values (in Hz) at fricative onset, midpoint and offset in whispered, semi-whispered and normal speech mode. The plots show questions (triangles) and statements (circles).

The other three spectral moments, i.e. standard deviation (SD) of the Centre of Gravity, skewness and kurtosis were significantly different only in whispered speech (and partly in semi-whispered speech) when questions as opposed to statements were produced.

Figure 7 shows the results for standard deviation values (i.e. the second spectral moment) at the acoustic midpoint of the fricative. The SD values were significantly higher for questions than for statements in whispered speech (t=5.923 p<.0001) and semi-whispered speech (t=1.786 p<.05). In phonated speech no significant differences regarding SD were found. In a similar vein, other spectral moments, i.e. kurtosis and skewness were not significantly different in questions and statements in phonated speech mode.

Figure 7: Boxplots for SD across all three speech modes.

Skewness and kurtosis differed significantly in questions as opposed to statements, however exclusively in whispered speech. The results are illustrated by Figure 8 and Figure 9.

Figure 8: Boxplots for skewness across all three speech modes.

The results revealed that as far as kurtosis and skewness is concerned, the difference in questions and statements was exclusively found for whispered speech. Both skewness and kurtosis values were lower in questions as opposed to statements (skewness: t=-3.949 p<.001, kurtosis: t=-6.020 p<.0001).

3. Discussion and conclusion

Our results show that differences between questions and statements are encoded in utterance-final consonants
properties. In particular, the frequency of the highest spectral peak of the frication noise (both for the range from 20Hz to 11kHz and from 2 to 4kHz) was higher in questions than in statements across all three speech modes. Sibilants were produced with higher Centre of Gravity accompanied by higher standard deviation in questions as opposed to statements. Questions were also produced with a higher intensity of the frication noise than statements. These results apply to all three speech modes. Further differences were found only in whispered speech mode where the mass of the spectral distribution moved towards higher frequencies in questions as compared to statements. The spectra of whispered questions were also characterised by a broader peak in comparison to whispered statements. Furthermore, spectral slopes were significantly steeper for questions than for statements in whispered speech (m2 was falling steeper in semi-whispered speech).

These results indicate that differences in spectral properties of utterance-final sibilants are most pronounced when questions and statements are produced in whispered speech mode, where F0 is completely absent. This in turn suggests that consonant properties might play a more important role in the production of intended intonation for whispered speech as compared to other speech modes. Whereas the role of vocalic elements for the perception of intonation in whispered speech was examined before, cf. e.g. Heeren & van Heuven (2009), it remains to be investigated to what extent sibilants contribute to perception of intonation across different speech modes, cf. also Niebuhr’s (2008) study on the role of stop aspiration for discriminating different meaning in German phonated speech.

Regarding the semi-whispered speech mode, our results show that some spectral properties of frication are similar to sibilants produced in whispered speech whereas others mirror properties of sibilants produced in normal speech. For instance, Centre of Gravity and its time-dependent dynamics extracted from sibilants produced in semi-whispered speech mode resembles sibilants produced in whispered speech, whereas other spectral moments such as skewness and kurtosis are more similar to those in sibilants produced in normal speech mode.

The conclusions regarding spectral differences of sibilants in dependence on intonation patterns are in line with results provided for German by Niebuhr, Lill and Neuschultz (2011) and Niebuhr (2012). The fact that two typologically different languages, Polish and German, show a similar kind of effects sheds light on segment-prosody interaction from a more cross-linguistic perspective.

Furthermore, the results have an impact on acoustic investigation of sibilants in general, as they show that differing intonation patterns can seriously affect spectral properties (and therefore all spectral measures). We believe that it is therefore indispensable to consider intonation patterns when investigating sibilants.

Finally, the results reveal a mutual interaction between segments and intonation (cf. also Kohler 2012) which might have consequences not only for re-considering the role of voiceless segments for production and perception of intonation patterns but also contribute to a discussion and verification of speech planning models. These models oscillate between approaches which treat segments and intonation separately but differ in the assumptions about how their interaction proceeds, e.g. ‘prosody first’ (Keating & Shattuck-Hufnagel 2002) versus ‘prosody last’ approach (Levitt 1989) and approaches against segment-prosody division according to which segments and prosody are intrinsically coherent (cf. e.g. Xu and Fang 2012).

4. Acknowledgements

This research has been supported by Bundesministerium für Bildung und Forschung (BMBF, Germany) Grant Nr. 01UG14111 to Marzenna Żygis. The study was also funded by Portuguese national funds through FCT - Foundation for Science and Technology, in the context of the project PEst-OE/EEI/UI0127/2014 to IEETA (Luis Jesus and Daniel Pape), and the post-doctoral fellowship from FCT (Portugal) SFRH/BPD/48002/2008 to Daniel Pape.

5. References


# Index of Authors

<table>
<thead>
<tr>
<th>A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Abugharbieh, R. –</td>
<td>174</td>
</tr>
<tr>
<td>Adda-Decker, M. –</td>
<td>5, 126</td>
</tr>
<tr>
<td>Agius, S. –</td>
<td>14</td>
</tr>
<tr>
<td>Agnew, Z. –</td>
<td>202</td>
</tr>
<tr>
<td>Al Kork, S. K. –</td>
<td>5</td>
</tr>
<tr>
<td>Alshangiti, W. –</td>
<td>9</td>
</tr>
<tr>
<td>Amelot, A. –</td>
<td>5, 57, 126</td>
</tr>
<tr>
<td>Aron, M. –</td>
<td>245</td>
</tr>
<tr>
<td>Auris, B. –</td>
<td>186</td>
</tr>
<tr>
<td>Auszmann, A. –</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bóna, J. –</td>
<td>41, 150</td>
</tr>
<tr>
<td>Baayen, R. H. –</td>
<td>325</td>
</tr>
<tr>
<td>Babel, M. –</td>
<td>13</td>
</tr>
<tr>
<td>Badin, P. –</td>
<td>17, 445</td>
</tr>
<tr>
<td>Baider, F. –</td>
<td>273</td>
</tr>
<tr>
<td>Baker, B. –</td>
<td>33</td>
</tr>
<tr>
<td>Barbe, M. T. –</td>
<td>25</td>
</tr>
<tr>
<td>Bassily, M. –</td>
<td>170</td>
</tr>
<tr>
<td>Baum, S. R. –</td>
<td>49</td>
</tr>
<tr>
<td>Baumann, S. –</td>
<td>21</td>
</tr>
<tr>
<td>Beaufremp, D. –</td>
<td>375</td>
</tr>
<tr>
<td>Bechet, M. –</td>
<td>45</td>
</tr>
<tr>
<td>Becker, J. –</td>
<td>25</td>
</tr>
<tr>
<td>Beke, A. –</td>
<td>329, 281</td>
</tr>
<tr>
<td>Berger, M.-O. –</td>
<td>245</td>
</tr>
<tr>
<td>Best, C. T. –</td>
<td>33, 94, 237</td>
</tr>
<tr>
<td>Birkholz, P. –</td>
<td>37, 336</td>
</tr>
<tr>
<td>Bombien, L. –</td>
<td>198</td>
</tr>
<tr>
<td>Bouarrouou, F. –</td>
<td>45</td>
</tr>
<tr>
<td>Bouhake, S. –</td>
<td>138</td>
</tr>
<tr>
<td>Bourguignon, N. J. –</td>
<td>49</td>
</tr>
<tr>
<td>Braun, A. –</td>
<td>413</td>
</tr>
<tr>
<td>Braun, B. –</td>
<td>387, 433</td>
</tr>
<tr>
<td>Brehm, A. –</td>
<td>53</td>
</tr>
<tr>
<td>Brkan, A. –</td>
<td>57</td>
</tr>
<tr>
<td>Brunner, J. –</td>
<td>61</td>
</tr>
<tr>
<td>Buchman, L. –</td>
<td>5</td>
</tr>
<tr>
<td>Bundgaard-Nielsen, R. L. –</td>
<td>33, 237</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Côté, D. –</td>
<td>277</td>
</tr>
<tr>
<td>Cadot, M. –</td>
<td>249</td>
</tr>
<tr>
<td>Cangemi, F. –</td>
<td>65</td>
</tr>
<tr>
<td>Carlson, E. –</td>
<td>253</td>
</tr>
<tr>
<td>Carter, P. –</td>
<td>69</td>
</tr>
<tr>
<td>Cederbaum, J. –</td>
<td>332</td>
</tr>
<tr>
<td>Chawah, P. –</td>
<td>5</td>
</tr>
<tr>
<td>Chen, W. –</td>
<td>395</td>
</tr>
<tr>
<td>Chen, Y. –</td>
<td>134</td>
</tr>
<tr>
<td>Chetcuti, F. –</td>
<td>1</td>
</tr>
<tr>
<td>Coleman, J. –</td>
<td>340</td>
</tr>
<tr>
<td>Corley, M. –</td>
<td>98</td>
</tr>
<tr>
<td>Crépel, S. –</td>
<td>17</td>
</tr>
<tr>
<td>Crevier-Buchman, L. –</td>
<td>126</td>
</tr>
<tr>
<td>Cummins, F. –</td>
<td>73</td>
</tr>
<tr>
<td>Cunha, C. –</td>
<td>78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakhoul, E. –</td>
<td>395</td>
</tr>
<tr>
<td>Dang, J. –</td>
<td>82, 194</td>
</tr>
<tr>
<td>De Looze, C. –</td>
<td>320</td>
</tr>
<tr>
<td>del Mar Vanrell, M. –</td>
<td>122, 417</td>
</tr>
<tr>
<td>Delvaux, V. –</td>
<td>86</td>
</tr>
<tr>
<td>Deme, A. –</td>
<td>90</td>
</tr>
<tr>
<td>Denby, B. –</td>
<td>5</td>
</tr>
<tr>
<td>Derrick, D. –</td>
<td>94, 395</td>
</tr>
<tr>
<td>Drake, E. –</td>
<td>98</td>
</tr>
<tr>
<td>Dreyfus, G. –</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Egger, S. –</td>
<td>387</td>
</tr>
<tr>
<td>Erickson, D. –</td>
<td>102</td>
</tr>
<tr>
<td>Esling, J. H. –</td>
<td>464</td>
</tr>
<tr>
<td>Evans, B. G –</td>
<td>9</td>
</tr>
</tbody>
</table>