Proceedings of Fonetik 2013

Edited by
Robert Eklund

Linköping University

< This page intentionally left blank>

Proceedings of Fonetik 2013

The XXVIth Annual Phonetics Meeting 12–13 June 2013, Linköping University Linköping, Sweden

Studies in Language and Culture
no. 21

Robert Eklund, editor

Linköping University

Conference website: www.liu.se/ikk/fonetik2013
Proceedings also available at: http://roberteklund.info/conferences/fonetik2013

Cover design and photographs by Robert Eklund
Photo of Claes-Christian Elert taken by Eva Strangert on the occasion of his 80th birthday

Proceedings of Fonetik 2013, the XXVIth Swedish Phonetics Conference
held at Linköping University, 12–13 June 2013
Studies in Language and Culture, no. 21
Editor: Robert Eklund
Department of Culture and Communication
Linköping University
SE-581 83 Linköping, Sweden

ISBN 978-91-7519-582-7
eISBN 978-91-7519-579-7
ISSN 1403-2570

Printed by LiU-Tryck, Linköping, Sweden, 2013

This conference is dedicated to Professor Claes-Christian Elert on the occasion of his 90th birthday, and for his outstanding contributions to Swedish phonetics

Preface

This volume contains the contributions to Fonetik 2013, the XXVIth Swedish Phonetics Conference, held at Linköping University, Sweden, 12–13 June 2013, and was organized as a collaboration between the Department of Culture and Communication, Linköping University, and the Department of Clinical and Experimental Medicine, Division of Speech and Language Pathology, Linköping University.

The papers appear in alphabetical order, sorted by first author's last name. An author index is found on page 89.

Only a limited number of copies of this publication have been printed for distribution among the authors and those attending the conference. For access to electronic versions of the contributions and the entire proceedings, the conference homepage is found at:

http://www.liu.se/ikk/fonetik2013

The conference proceedings are also available at:

http://roberteklund.info/conferences/fonetik2013

The organizers would like to thank all contributors to the proceedings. We would also like to extend our thanks to *Fonetikstiftelsen* for generous financial support. Finally, we would thank the following people, whose help is very appreciated: Gunilla Christiansen, Carin Franzén, Nigel Musk, Anna F. Söderström and Linnea Björk Timm.

Linköping, May 2013

Robert Eklund, Agnese Grisle, Jan Anward, Charlotta Plejert, Christina Samuelsson and Simon Sundström

Previous Swedish Phonetics Conferences (from 1986)

I	1986	Uppsala University
II	1988	Lund University
III	1989	KTH Stockholm
IV	1990	Umeå University (Lövånger)
V	1991	Stockholm University
VI	1992	Chalmers and Göteborg University
VII	1993	Uppsala University
VIII	1994	Lund University (Höör)
—	1995	(XIIIth ICPhS in Stockholm)
IX	1996	KTH Stockholm (Nässlingen)
X	1997	Umeå University
XI	1998	Stockholm University
XII	1999	Göteborg University
XIII	2000	Skövde University
XIV	2001	Lund University
XV	2002	KTH Stockholm
XVI	2003	Umeå University (Lövånger)
XVII	2004	Stockholm University
XVIII	2005	Göteborg University
XIX	2006	Lund University
XX	2007	KTH Stockholm
XXI	2008	University of Gothenburg
XXII	2009	Stockholm University
XXIII	2010	Lund University
XXIV	2011	KTH Stockholm
XXV	2012	University of Gothenburg

Table of contents

Extracting and analysing co-speech head gestures from motion-capture data

Simon Alexanderson, David House & Jonas Beskow
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

Abstract

This paper reports on a method developed for extracting and analyzing head gestures taken from motion capture data of spontaneous dialogue in Swedish. The head gestures were extracted automatically and then manually classified using a 3D player which displays time-synced audio and 3D point data of the motion capture markers together with animated characters. Prosodic features were extracted from syllables co-occurring with a subset of the classified gestures. The head gestures show considerable variation in temporal synchronization with the syllables, while the syllables generally show greater intensity, higher F0, and greater F0 range when compared to the mean across the entire dialogue.

Introduction

Along with intonation there are a wealth of visual gestures, including head, facial and body movements, which co-occur with speech adding emphasis and prominence to portions of the utterance and contributing to the flow of the dialogue. Beat gestures such as rapid hand movement as described by McNeill (2005) are particularly interesting in this regard as they coincide and appear to be synchronized with prosodic and intonational peaks related to prominence. They also share the same prominence-lending function as the focal or sentence accent, but they can be repetitive marking the stress or rhythmical structure of an utterance.

In terms of timing and synchronization, there are many similarities between intonational gestures and visual gestures produced in accompaniment with speech. Both intonation and visual gestures are free to vary across the vowels and consonants of the segments. In intonation, however, this variation is restricted by the specific patterns used by a language to signal meaning in spoken interaction.

If we wish to study the timing of gestures in the same way as we approach timing in intonation, we currently lack an established methodology to extract and analyze gestures, especially gestures occurring in spontaneous dialogue. The Spontal corpus of Swedish dialogue provides a rich database as a point of departure for testing gesture extraction and analysis methodology. The database, containing more than 60 hours of unrestricted conversation in over 120 dialogues between pairs of speakers is comprised of high-quality audio and video recordings (high definition) and motion capture for body and head movements for all recordings (Edlund et al., 2010).

The progression and timing of the motion of a head nod can be described in much the same terms as an intonational excursion. However, many speakers move their heads extensively while speaking, and manual annotation of head-gestures in spontaneous dialogue involves a number of difficulties. Among the sources of disagreement are segment boundaries and location of maximum extent. Gestures may be multifunctional and involve simultaneous rotations around several axes. In this study we present a semi-automatic approach to head-gesture annotation, in which the main goal is to test its viability and potential for annotation of gesture data on a large scale, such as is represented by the Spontal database.

Method

To overcome some of the difficulties of head-gesture annotation we are developing and testing a semi-automatic annotation procedure consisting of two steps. First an automatic head-gesture segmentation algorithm is applied to the motion capture data and then the segments are manually classified by the annotators.

In addition to gesture annotation we also processed the audio files of the speakers and generated pitch and intensity data, talk vs. no-talk segmentation and syllable segmentation. This was done to be able to investigate the relationship between the head-gestures and prosodic features.

In this exploratory stage we chose one of the dialogues from the Spontal database where the

speakers demonstrated a relatively large number and variety of head-movements. The participants were a male and female who did not know each other.

Automatic segmentation of head nods

A simplistic segmentation approach was used for head-nod segmentation. The head orientations were calculated from three markers attached to the headbands of the speakers and expressed in an Euler angle form. We then calculated the angular velocity of the pitch component as the basis for the segmentation. During a head nod the angular velocity follows an oscillatory movement during a limited time period, and we use its local extreme values as segment boundaries. A segment is defined as a maximum velocity followed by a minimum or vice versa. Two thresholds are used in this process. The first enforces the peak velocity to be over a specified value, thus prohibiting small movements caused by noise to be interpreted as nods, and the second enforces the nod segments to be shorter than a specified duration. Using the velocity peaks as segment boundaries has some desirable features. During head-nods there are rapid changes in angular velocity causing clear detectable spikes in the data. It also naturally splits repeated head-nods into a consecutive sequence of down-up (nod) and up-down (jerk) segments, which fits well with the MUMIN multimodal annotation scheme proposed by Allwood et al. (2007).

For our data, the minimum peak velocity threshold was empirically set to a value of 0.0015 radians/s, which was the lowest value before noise in the data would be manifested as segments. The maximum segment duration threshold was set to 1000 ms.

In the current study we were interested in gestures synchronized with speech and especially gestures with beat-function produced in companion with stressed syllables. In order to narrow our search space we discarded all segments occurring while the subject was not speaking and further all nods in the up-down order, leaving all down-up nods occurring during speech as our candidate gestures. The segmentation of talk vs. no-talk was performed with an automatic speech activity detection algorithm (Heldner et al., 2011).

Manual classification

After running the automatic processing, the resulting segments were examined and manually classified by two annotators. To make the classification, the annotators viewed each segment in a specially designed 3D player which plays time-synchronized audio and displays 3D point data of the motion capture markers together with animated characters following the 3D marker motion. As expected, the segments from the automatic process did not only contain unambiguous beat gestures, but also gestures with other functions co-occurring with speech (McClave, 2000). Such other functions were feedback, confirmation, word or phrase intensification and listing of lexical items. Moreover, some of the extracted gestures did not appear to co-occur with a stressed syllable.

Therefore, an annotation scheme was devised with two main queries: Q1, “Is there a clear nod in synchrony with a stressed syllable?” and Q2, “Is the nod multifunctional?” If the answer to the first query is positive the second query is also answered. This scheme resulted in three categories: 1. No clear nod in synchrony (no sync), 2. A clear nod with a beat function (beat-function), and 3. A clear nod which is multifunctional (multi-function).

Prosodic features

The pitch and intensity curves were extracted from the audio signals of the speakers using the SNACK toolkit (Sjölander & Beskow, 2000). Also syllable boundaries and nuclei were derived by applying Mermelstien’s convex-hull algorithm (Mermelstien, 1975).

After gesture- and prosodic feature extraction was performed, we determined which syllable was closest in time to the maximum rotation of the nod. The time difference between each gesture and the start and nucleus of its closest syllable was then calculated. Also pitch and intensity properties of the closest syllable were compared with mean values across the total dialogue.

Results

The automatic segmentation algorithm was applied on the 20 minute dialog, extracting 64 nod-segments for speaker 1 (male) and 150 segments from speaker 2 (female). The manual classification by the two annotators resulted in a 69% and 65% agreement for the head-nods of speaker 1 and speaker 2 respectively. As is displayed in *Table 1* the annotators showed greatest agreement for the category with no

syllable-synchronization. Less agreement was obtained from the categories beat-function and multi-function. Annotator 1 perceived more of the nods having beat-function while annotator 2 perceived more having multi-function.

Table 1: Results of the manual annotation showing class and agreement. The first number in each pair is the male speaker, the second is the female.

Class	Annotator 1	Annotator 2	Agreement
No sync	23/44	15/41	13/29
Beat-function	9/74	11/66	4/50
Multi-function	32/32	38/43	27/18
Total	64/150	64/150	44/97

Figure 1 shows the durations of the segments in the different categories for speaker 1 and speaker 2 for those gestures for which the annotators agreed.

Note that the segment length is the part of the nod between the peak velocities of the downwards and upwards phase as described earlier. The results show a tendency for the multi-functional nods to be shorter than those with a beat function. The nods classified as non-synchronous showed greater temporal variation than the other categories.

The subset of gestures annotated as having a beat function for the female speaker was analyzed in terms of timing related to its closest syllable. Only the gestures from the female speaker were analyzed due to the small number of beat gestures annotated for the male speaker. *Figure 2* shows the time difference between two different anchor-points of the syllable (onset and nucleus) and three different phases of the nod: peak velocity of the downward phase (p1), max rotation (p2) and peak velocity of the upward phase (p3). The timing relationship between the gesture and the syllable does not seem to be influenced by the choice of syllable anchor-point. The timing relationships show a considerable amount of variation regarding the question of gesture synchronization with the syllable.

When compared with mean values across the total dialogue, the syllables closest to the annotated beat nods generally showed greater integrated intensity, higher F0 at the nucleus, and greater F0 range as shown in *Table 2*.

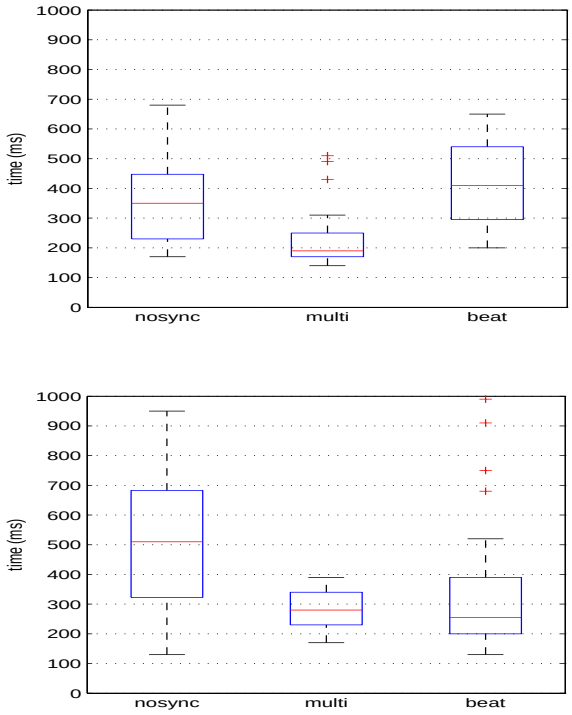


Figure 1: Durations of the agreed nods in the classes for speaker 1 (top) and speaker 2 (bottom).

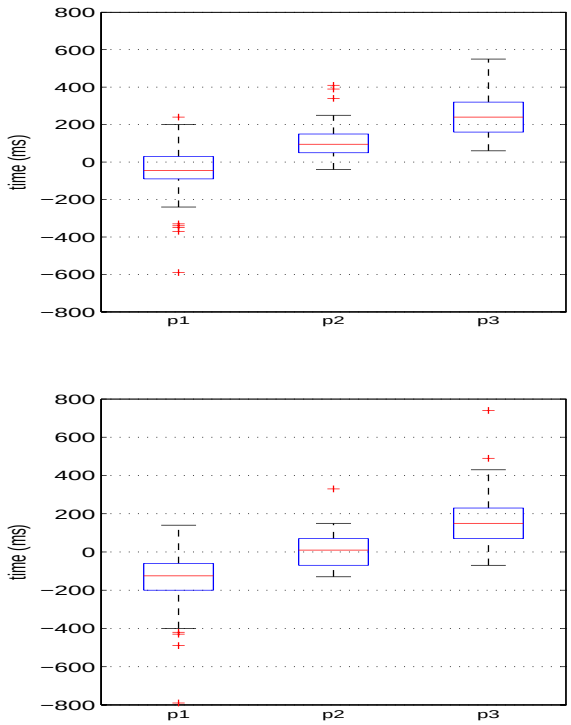


Figure 2. Timing of syllable anchor-points, onset (top) and nucleus (bottom), with respect to three different phases of the nod: peak velocity of the downward phase (p1), max rotation (p2) and peak velocity of the upward phase (p3).

Table 2. Comparison of syllable features

	integrated intensity (dBs)	F0 at nucleus (Hz)	F0 range (Hz)
Mean closest syllable	13.2	217	77
Mean across the total dialogue	11.1	180	67

Discussion

The goal of this study was primarily to develop and test a new method for extracting and annotating gesture data. While head gestures, and in particular relatively small gestures, have been problematic for manual annotation schemes, the semi-automatic method tested here shows promise for the selection and fast annotation of multimodal data.

This process is a starting point for further work in the field of automatic recognition and classification of multimodal communication. Given the fact that non-verbal and verbal communication are tightly coupled, the motion data may provide important and robust features for machine-learning techniques. In this study we started an investigation along this path by analyzing features for prominence detection and their coupling to beat gestures. This method may also prove useful in analyzing features related to other communicative functions such as feedback and turn-taking.

While the results concerning the analysis of the characteristics of the head nods and their timing must be seen as quite preliminary due to the small sample and very limited classification categories, the timing results are consistent with those results reported in Leonard and Cummins (2011) for hand and arm beat gestures. More available data and the development of automatic methods and tools should better enable us to compare and evaluate results such as these.

Acknowledgements

The work reported here is carried out within the projects: “Timing of intonation and gestures in spoken communication,” (P12-0634:1) funded by the Bank of Sweden Tercentenary Foundation, and “Large-scale massively multimodal modelling of non-verbal behaviour in spontaneous dialogue,” (VR 2010-4646) funded by the Swedish Research Council. A longer version of this paper will be presented at the Tilburg Gesture Research Meeting in June 2013.

References

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta & P. Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41:273−287.

Edlund, J., J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson & D. House. 2010. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valetta, Malta, 2992−2995.

Heldner, M., J. Edlund, A. Hjalmarsson & K. Laskowski. 2011. Very short utterances and timing in turn-taking. In: *Proceedings of Interspeech 2011*. Florence, Italy, 2837−2840.

Leonard, T. & F. Cummins. 2011. The temporal relation between beat gestures and speech. *Language and Cognitive Processes* 26: 1457−1471.

McClave, E. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32:855−878.

McNeil, D. 2005. *Gesture and thought.* Chicago: The University of Chicago Press.

Mermelstien, P. 1975. Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America* 58:880−883.

Sjölander, K. & Beskow, J. 2000. WaveSurfer – an open source speech tool. In: B. Yuan, T. Huang & X. Tang (eds.), *Proceedings of ICSLP 2000, 6th International Conference on Spoken Language Processing*. Beijing, China, 464−467.

Alternative questions and sentence intonation in Greek

Antonis Botinis, Anthi Chaida & Marianna Georgouli
Laboratory of Phonetics and Computational Linguistics, University of Athens, Greece

Abstract

This is an experimental study of intonation as a function of alternative questions in Greek. The results of a production experiment indicate that: (1) compound sentences with alternative question functions are produced with tonal prominence on the disjunctive element, (2) the clause on the left has fairly regular tonal inflections associated with stressed syllables whereas (3) the clause on the right does not exhibit much tonal variability. Thus compound alternative questions have a distinct tonal structure with no apparent similarities with any other type of questions in Greek.

Introduction

This study is an investigation of sentence prosody and alternative questions in Greek. Alternative questions, much like polar (yes/no) ones, lack any morphological and/or syntactic markers in Greek. Earlier research has indicated that the characteristic tonal structure of simple polar questions is a tonal prominence in the vicinity of the right sentence edge whereas complex sentences with main and subordinate syntactic structures are associated with distinct tonal structures (Chaida, 2007, 2010).

The main question addressed in this study is whether alternative questions have similar tonal structures to other types of questions or whether they have distinct tonal structures. In relation to the main question, further questions are addressed with reference to the main tonal characteristics of alternative questions as well as local and global tonal structures in Greek.

In addition to alternative questions, the tonal structures of simple and compound sentences with statement and question functions have been investigated. In this context, comparisons between sentence functions and sentence complexity related to alternative questions will be outlined. Very little is known about the intonation of alternative questions reported in the international literature, including functional constituency and related tonal structures.

Experimental methodology

Four sets of sentence pairs were constructed. Each pair consisted of one statement and one question. The first 2 sets were simple sentences whereas the last 2 ones were compound sentences with coordinative or alternative functions (*Table 1*).

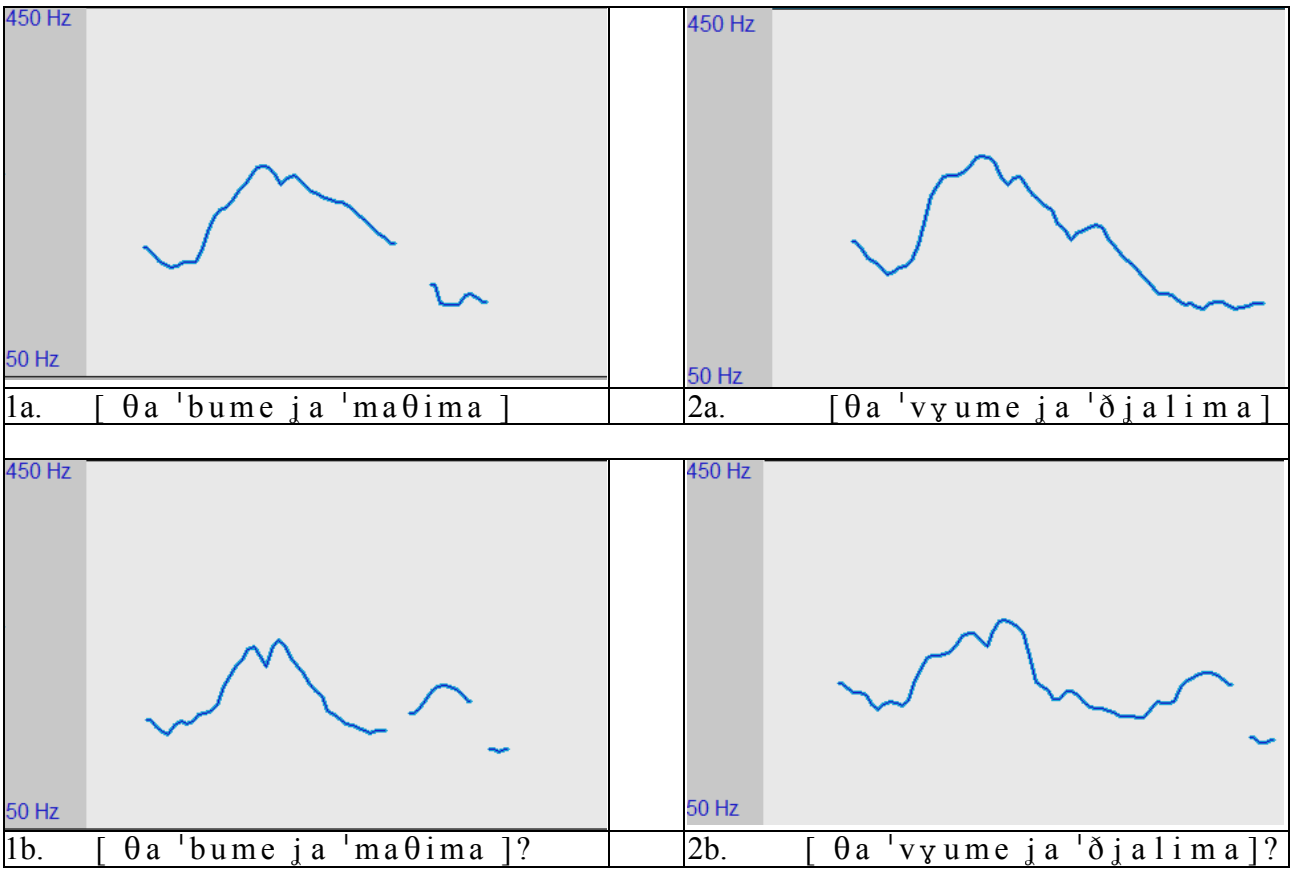
Ten female speakers in their twenties, with standard Athenian pronunciation, produced the speech material at the phonetics studio of the Laboratory of Phonetics and Computational Linguistics at Athens University. The speech material was written in standard Greek orthography and was read from separate cards, each card containing only one sentence. The speakers' productions were recorded directly onto a computer disc and analysed with the Praat software package. Tonal (F0) measurements were made at the beginning and at the middle of each syllable of the total speech material.

Table 1. Speech material: simple statements and questions (1–2) as well as compound coordinative and alternative statements and questions (3–4).

1a.	[θa ˈbume ʝa ˈmaθima] (We go in for lesson.)
1b.	[θa ˈbume ʝa ˈmaθima]? (We go in for lesson?)
2a.	[θa ˈvɣume ʝa ˈðʝalima] (We go out for break.)
2b.	[θa ˈvɣume ʝa ˈðʝalima]? (We go out for break?)
3a.	[θa ˈbume ʝa ˈmaθima ce θa ˈvɣume ʝa ˈðʝalima] (We go in for lesson and we go out for break.)
3b	[θa ˈbume ʝa ˈmaθima ce θa ˈvɣume ʝa ˈðʝalima]? (We go in for lesson and we go out for break?)
4a	[θa ˈbume ʝa ˈmaθima i θa ˈvɣume ʝa ˈðʝalima] (We go in for lesson or we go out for break.)
4b	[θa ˈbume ʝa ˈmaθima i θa ˈvɣume ʝa ˈðʝalima]? (We go in for lesson or we go out for break?)

Results

The results are presented in *Figures 1–5*. *Figures 1–4* show raw tonal curves of one female speaker and *Figure 5* average values of 10 female speakers.



Figures 1-2. Intonation exemplification of simple statements over (1a–2a) vs. questions under (1b–2b).

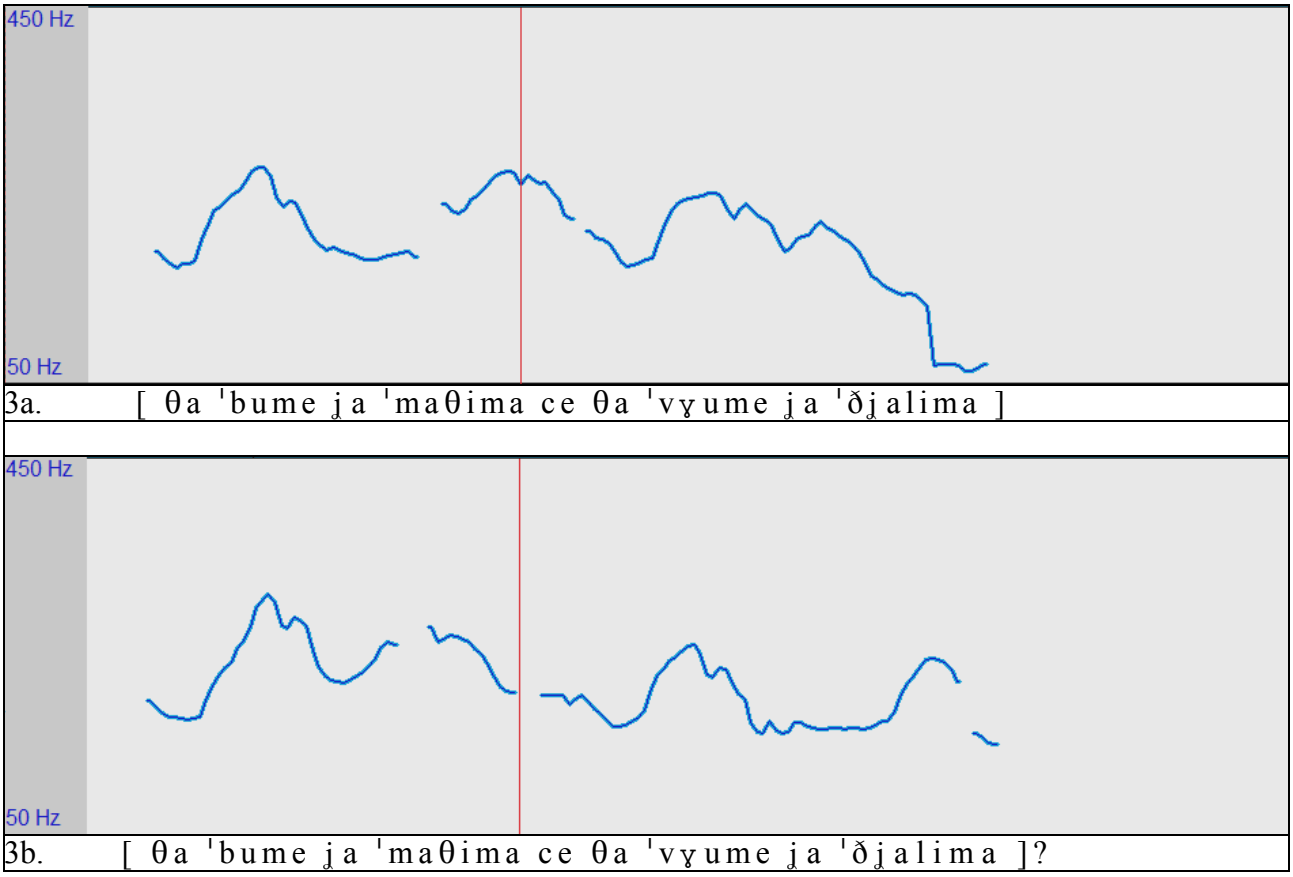


Figure 3. Intonation exemplification of coordinated statements over (3a) vs. questions under (3b).

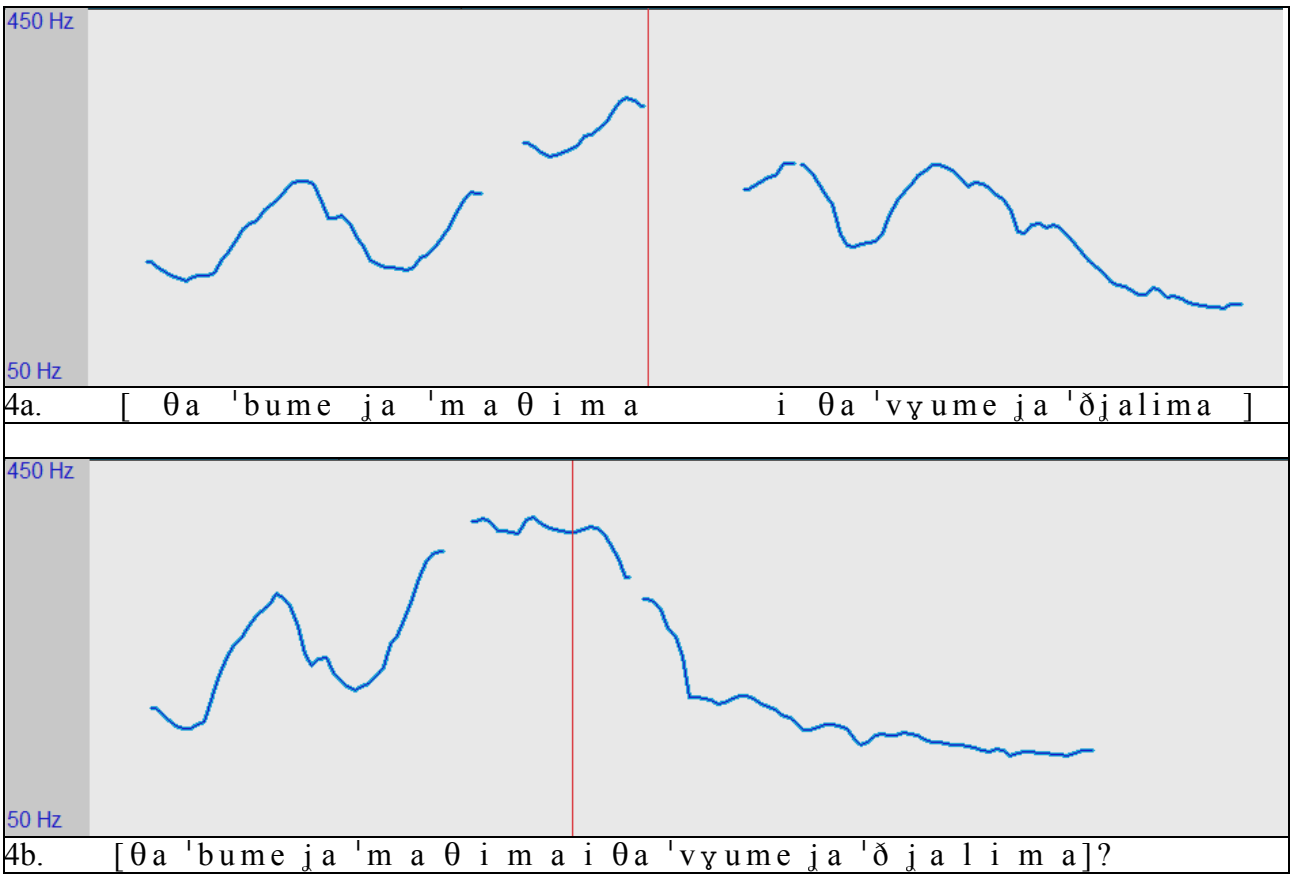


Figure 4. Intonation exemplification of alternative statements over (4a) vs. questions under (4b).

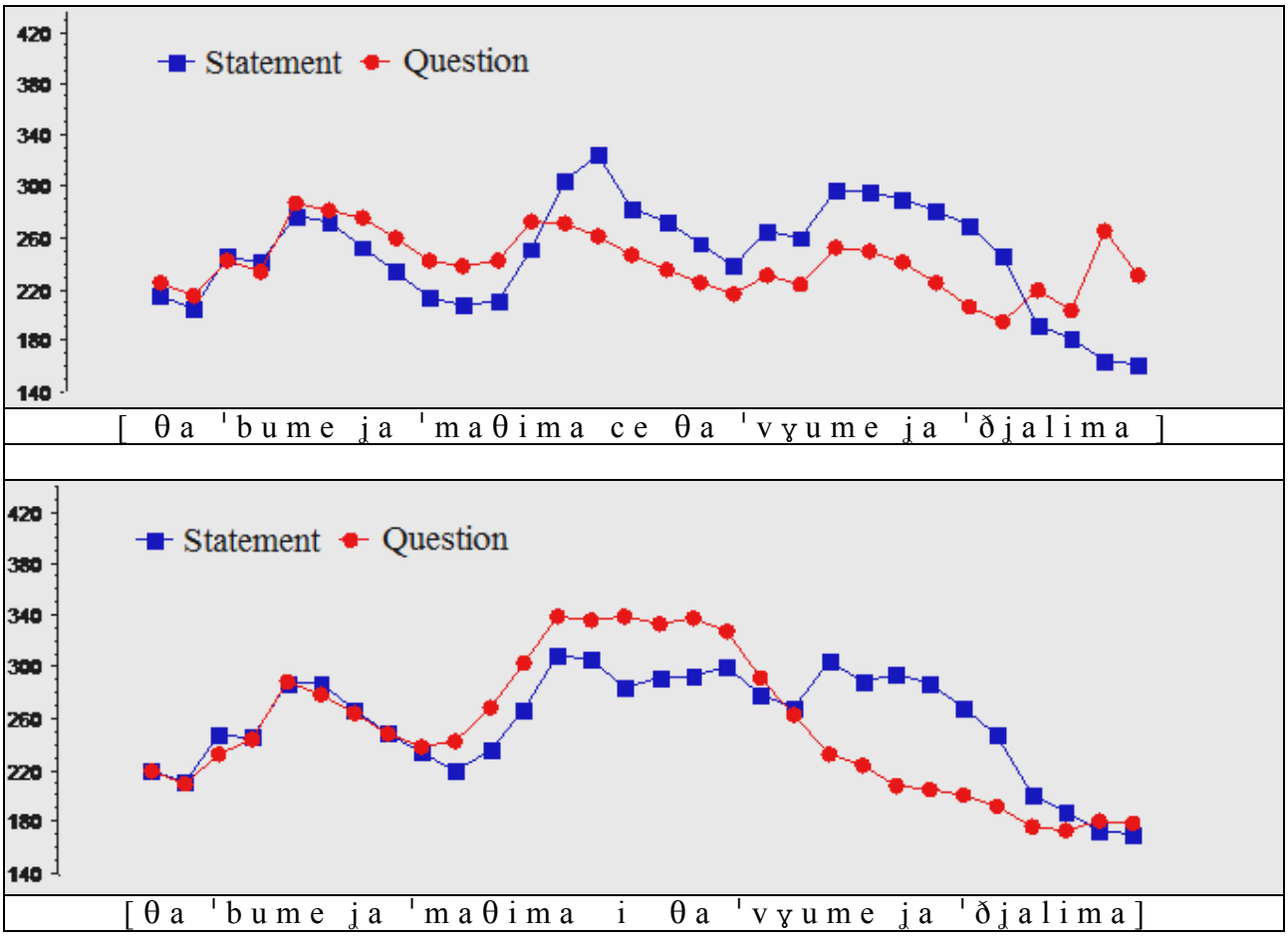


Figure 5. Coordinated statements vs. questions over and alternative statements vs. questions under.

Not accounting for local tonal perturbations, i.e. microprosodic effects, simple statements have a one tonal peak structure whereas simple polar questions have a two tonal peak structure (*Figures 1–2*). The initial rise of the first tonal peak in both statements and polar questions is associated with the respective stressed syllables whereas the second tonal peak in polar questions is taking place within the last two unstressed syllables of the test sentences.

In the following *Figure (3)*, the first clauses of the compound sentences are associated with two tonal peaks. However, in *3a* (statement) a late-rise is associated with the second tonal peak whereas, in *3b* (polar question) the respective tonal rise is an early one. On the other hand, the second tonal peak is associated with a late-fall into the next clause vs. an early fall within the first clause, for statements and polar questions, respectively. In the second clauses, each clause has a fairly similar tonal structure to the respective simple statement and polar question one.

Figure 4 shows tonal structures of compound sentences. In alternative statements (*4a*), a specific tonal event is associated with the alternative disjunctive element [i], which has a rise-fall tonal pattern and thus a tonal peak. In addition, a prosodic phrasing between the two clauses is evident. Apart from the latter differences, alternative statements (*4a*) and coordinative sentences (*3a*) have fairly similar tonal structures, especially with reference to the clauses on the right.

In alternative questions (*4b*), the respective tonal structures show major differences in comparison to both simple and compound sentences under examination. First, a "tonal hat" is formed, including the last stress group of the clause on the left as well as the disjunctive element [i]. The tonal rise is correlated with a stressed syllable on the left clause whereas the tonal fall is concluded by the beginning of the stress group on the right clause. Furthermore, hardly any tonal inflection after the tonal hat is evident.

The quantitative results of this study are shown in *Figure 5*. As outlined above, polar questions are characterised with a tonal rise-fall at the right edge of the compound sentences (*over*) whereas alternative questions are missing the respective tonal structure (*under*). Instead, a tonal hat pattern associated with the disjunctive element is formed, followed by a tonal flattening to the end of the sentence.

Discussion

The results of the present study indicate that the intonation of compound alternative questions in Greek is very dissimilar to other types of question intonation we have been studying (e.g. Chaida, 2007). Its peculiarity is mainly related to the disjunctive element, which is correlated with a prominent tonal structure (i.e. nucleus) followed by a steep fall and a tonal flattening to the end of the speech material. Thus, alternative questions have a fairly similar tonal structure at global level to focus productions (Botinis, 1989; Themistocleous, 2011; Nikolaenkova, 2013).

Greek is a "stress language" and local tonal inflections have no lexical function in the way "accent" or "tone" languages, such as Swedish or Chinese, may have (Bruce, 1977; Botinis, Granström & Möbius, 2001; Xu, 2011). However, (pitch) accents are as a rule associated with stressed syllables and prosodic boundaries with other tonal patterns. A basic question is the derivation of surface tonal representations. In accordance with the results of this study, alternative questions and different sentence modalities trigger specific tonal structures and tonal combinations with functional distinctions over sentence domains.

Acknowledgements

Thanks to Special Research Account of Athens University for a travel grant to Fonetik 2013 conference.

References

Botinis, A. 1989. *Stress and Prosodic Structure in Greek*. Lund: Lund University Press.

Botinis, A., B. Granström & B. Möbius. 2001. Developments and paradigms in intonation research. *Speech Communication* 33:263–296.

Bruce, G. 1977. Swedish Word Accents in Sentence Perspective. Lund: Gleerup.

Chaida, A. 2007. Tonal structures of complex sentences in Greek. *Proc. 8th Intern. Conference of Greek Linguistics*, Ioannina, Greece, 61–69.

Chaida, A. 2010. *Production and Perception of Intonation and Sentence Types in Greek.* PhD thesis, University of Athens.

Nikolaenkova, O. 2013. *Multifactor Analysis of Greek and Russian Sentence* (in Greek). PhD thesis, University of Athens.

Themistocleous, C. 2011. *Prosody and Information Structure in Athenian and Cypriot Greek* (in Greek). PhD thesis, University of Athens.

Xu, Y. 2011. Speech prosody: A methodological review. *Journal of Speech Sciences* 1:85–115.

Perception of focus and word order variability in Greek

Anthi Chaida, Olga Nikolaenkova & Antonis Botinis
Laboratory of Phonetics and Computational Linguistics, University of Athens, Greece

Abstract

This is an experimental study of prosodic focus and syntactic word order interactions with reference to agent/recipient semantic relations in Greek. In accordance with one perception experiment, the results in morphology neutral context indicate: (1) listeners do not perceive agent and recipient semantic distinctions and (2) prosody and focus applications do not thus have any perception effect whereas (3) word order has a major interference effect.

Introduction

The present study investigates the perception of semantic relations with reference to agent and recipient specifications. In Greek, the syntactic elements S(ubject), V(erb) and O(bject) may in principle appear in any word order position. However, these syntactic elements may be dislocated at the left edge of a sentence, as a syntactic effect of focus application (Botinis et al., 2005; Nikolaenkova, 2013, 2011).

In addition to focus and word order, morphology and in particular casus declensions may have a syntactic function and related semantic distinctions in Greek. In the sentence, e.g. *[to a'ɣori ma'loni ti ʝi'neka]* "the boy is scolding the woman", the subject on the left edge is a masculine nominative and the object on the right edge is a feminine accusative. A focus application on the subject has usually only prosodic correlates (*to **a'ɣori** ma'loni ti ʝi'neka)* whereas a focus application on the object may have either prosodic correlates *(to a'ɣori ma'loni ti ʝ **i'neka**)* or both prosodic and syntactic correlates *(ti **ʝi'neka** ma'loni to a'ɣori)*. If the object of the latter sentence were the subject, the corresponding casus would be in nominative, i.e. *(i **ʝi'neka** ma'loni to a'ɣori)*. Thus, in Greek, casus declension is a major syntactic correlate whereas word order position shows a large variability and may be correlated with focus applications.

Either subject and object or both may be neutral with reference to casus. Thus, the sentence *[to a'ɣori ma'loni to ko'ritsi]* "the boy is scolding the girl", appears with casus neutralization as the nominative and accusative of neutral nouns have no morphological distinction. The object *[to ko'ritsi]* may move on the left edge when in focus, i.e. *[to **ko'ritsi** ma'loni to a'ɣori]*, whereas *[to ko'ritsi]* may also be a subject in focus, i.e. *[to **ko'ritsi** ma'loni to a'ɣori]*. In the former case the object *[to ko'ritsi*] has recipient function whereas in the latter case the subject *[to ko'ritsi*] has an agent function. In accordance with these examples, a major issue is raised with reference to agent vs. recipient relations and respective distinctions in speech perception.

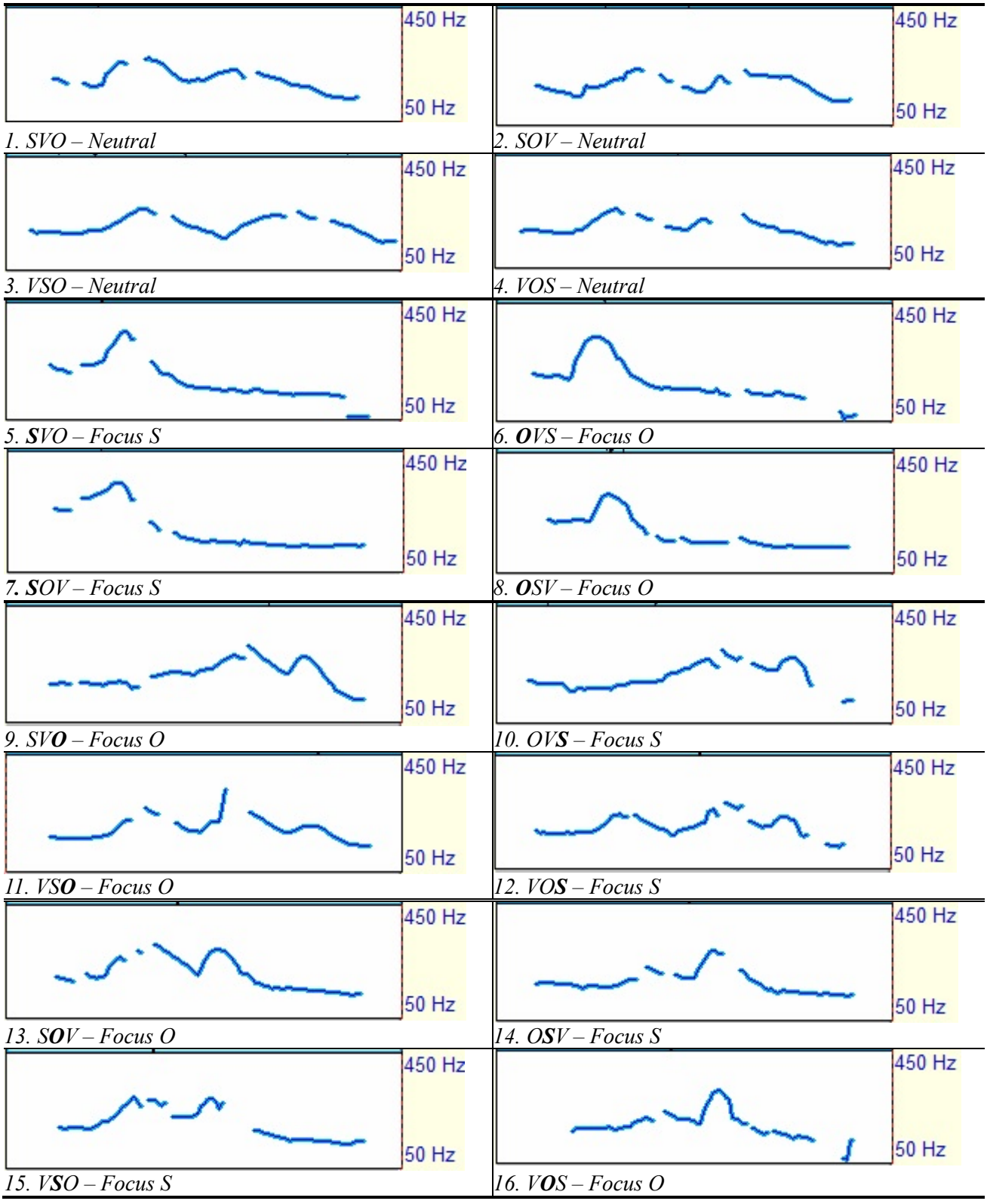
Experimental methodology

A set of 6 test sentences was designed, with all possible syntactic variability (SVO, SOV, OSV, OVS, VSO, VOS) crossed with 3 prosodic focus productions (focus on subject, focus on object, and neutral). The reference test sentence was *[to a'ɣori ma'loni to ko'ritsi]* (the boy is scolding the girl) with the above word order and focus application variability. The speech material was recorded at the Laboratory of Phonetics and Computational Linguistics of Athens University by 2 female speakers in their thirties, with standard Athenian pronunciation.

A set of 72 stimuli (6 syntactic structures × 3 focus productions × 2 nouns × 2 speakers) was organized in a random order and 10 native listeners participated in a close perception test. For each stimulus, listeners were asked to identify the agent. The stimuli were presented on a Matlab program (Mathworks, 2011) over Direct Sound headphones, and were played once the "Play Next" button was pressed. A prompt appeared on screen containing the two possible response buttons, which contained the words *"το κορίτσι"* (the girl) and *"το αγόρι"* (the boy), written in standard Greek orthography. Participants could press only one of them. A "Replay" button was also included in the prompt, which permitted three (3) repetitions maximum.

Results

Figures 1–16 show raw tonal curves of selected productions. The results of the perception experiment are presented in *Table 1* and *Figure 17*. Statistical analysis was carried out using the SPSS 19.0 (SPSS Inc., 2009) software package.



1. SVO – Neutral | *2. SOV – Neutral*

3. VSO – Neutral | *4. VOS – Neutral*

*5. **S**VO – Focus S* | *6. **O**VS – Focus O*

*7. **S**OV – Focus S* | *8. **O**SV – Focus O*

*9. SV**O** – Focus O* | *10. OV**S** – Focus S*

*11. VS**O** – Focus O* | *12. VO**S** – Focus S*

*13. S**O**V – Focus O* | *14. O**S**V – Focus S*

*15. V**S**O – Focus S* | *16. V**O**S – Focus O*

Figures 1–16. Raw tonal curves of selected productions with different syntactic and focal conditions produced by a female speaker

As shown in *Figures 1–4*, the neutral productions have a regular tonal structure, according to which stressed syllables of lexical words are as a rule associated with local tonal commands which are aligned with respective stress group boundaries. The final stress groups have however a suppressed tonal structure, as a result of utterance finality. Focus productions (*Figures 5–16*), on the other hand, are associated with extended tonal variability: (a) speech material in focus has a local tonal range expansion, (b) speech material out of focus

undergoes deaccentuation and (c) speech material out of focus undergoes major tonal compression. These three ways may operate simultaneously or in combinations in variable linguistic domains. As noted in many previous production studies (e.g. Botinis et al., 2000; Botinis et al., 2005; Chaida, 2010; Nikolaenkova, 2010, 2013), focus productions seem to have constant tonal correlates which operate independently from syntactic, although both tonal and syntactic structures may function complementary with reciprocal reinforcement for focus structures and focus distinctions. Thus, regardless semantic interpretations, the basic tonal characteristics of focus are the same with respect to focus position, i.e. there are great similarities between the following groups of utterances: *(i)* SVO-Focus S, OVS-Focus O, SOV-Focus S, OSV-Focus O (*Figures 5–8*), with focus on the 1st word; *(ii)* SVO-Focus O, OVS-Focus S, VSO-Focus O, VOS-Focus S (*Figures 9–12*), with focus on the 3rd word; *(iii)* SOV-Focus O, OSV-Focus S, VSO-Focus S, VOS-Focus O (*Figures 13–16*) with focus on the second word.

With regards to perception, generalized Estimating Equations (GEE) analysis was used, in order to conduct the equivalent of a repeated measures ANOVA. The logistic linking function was used and each of the two repetitions (one per speaker) was treated as a repeated measurement. Focus syntax, and speaker were within-subjects factors; participant and stimulus were a between-subjects factor.

Table 1. Results for the identification of the agent in each utterance, with respect to syntactic structure and prosodic focus.

Syntax - Focus	Identification (n)	Identification (%)	SD
SVO - FOCUS S	35/40	88%	.335
SVO - FOCUS O	40/40	100%	.000
SOV - FOCUS S	34/40	85%	.362
SOV - FOCUS O	40/40	100%	.000
OSV - FOCUS S	4/40	10%	.304
OSV - FOCUS O	17/40	43%	.501
OVS - FOCUS S	0/40	0%	.000
OVS - FOCUS O	5/40	13%	.335
VSO - FOCUS S	22/40	55%	.504
VSO - FOCUS O	32/40	80%	.405
VOS - FOCUS S	12/40	30%	.464
VOS - FOCUS O	19/40	48%	.506

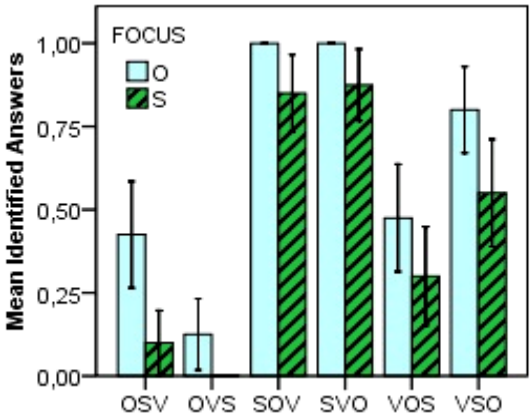


Figure 17. Mean results for the intended identification of the agent in each utterance, with respect to syntactic structure and prosodic focus (placed either on the object or on the subject).

In general, syntax (Wald χ^2 (5) = 10673,843, $p < .0001$) had a highly significant main effect, with SVO and SOV structures prevailing in identification (94% for SVO, 93% for SOV, 68% for VSO, 39% for VOS, 26% for OSV, 6% for OVS).

Focus (Wald χ^2 (1) = 5076,080) had a highly significant main effect, with a higher identification of utterances with focus on the object (64% for Focus O, 45% for Focus S).

Furthermore, a highly significant main effect of the interaction between syntax and focus was noted (Wald χ^2 (5) = 4299,466, $p < .0001$).

The highest identification results regarding the agent correspond to all syntactic structures, where the agent is placed in the beginning of the sentence, regardless focus placement, i.e. 100% for SVO-Focus O, 100% for SOV-Focus O, 88% for SVO-Focus S, and 85% for SOV-Focus S. The results for SVO-Focus O and SOV-Focus O were highly significant regarding the interaction between syntax (word order) and focus (Wald χ^2 (1) = 1384,470, $p < .0001$, and Wald χ^2 (1) = 1445,804, $p < .0001$, respectively). On the other hand, there were very low identification rates for OVS-Focus O (13%), OSV-Focus S (10%), and OVS-Focus S (0%).

Neutral productions were 100% identified and are thus excluded from the *Table* and *Figure* above.

Discussion

The results of the present study indicate that syntactic structure, and word order in particular, has a major interference effect on the perception of agent vs. recipient semantic relations and respective specifications.

It is evident that the identification of the agent is highly correlated with the first word of an utterance, which is in accordance with the very high identification rates of respective productions, i.e. 100% for SVO-Focus O, 100% for SOV-Focus O, 88% for SVO-Focus S, and 85% for SOV-Focus S. The placement of the subject and thus the specification of the agent in the beginning of a sentence is the most neutral syntactic structure in Greek. On the other hand, an SVO canonical syntactic structure is the norm taught in schools and grammar textbooks as a rule, which constitutes a significant educational and social interference.

The dominance of the SVO syntactic structure in neutral contexts in Greek has been proposed, in accordance with a series of results in earlier experimental studies (Nikolaenkova, 2013). In a written experiment, students of linguistics at Athens University were asked to set up simple sentences consisting of a wide set of different words with variable morphology, scattered in a random order in a piece of paper. The results of this written experiment suggested an overwhelming SVO syntactic structure. In an oral experiment, another group of students were presented with animated pictures on a computer screen in an agent vs. recipient context and were asked to read aloud what they were seeing with a simple sentence. Much like the results of the written experiment, the results of the oral experiment also suggested an overwhelming SVO syntactic structure.

In contrast to high identification rates with reference to S-leading syntactic structures and respective agent correlations, other word order structures resulted in low identification rates, such as OVS-Focus O (13%), OSV-Focus S (10%), and OVS-Focus S (0%). These results indicate that there is a strong perception bias of word order and agent semantic correlations. The results also suggest that when other linguistic factors, in the first place morphology and prosody, are neutralized, the interpretation of spoken utterances is mainly based on word order syntactic structures.

Focus productions have a variety of language structure correlates, including prosodic, morphological and syntactic ones. This is in accordance with the high functional load and semantic distinctions which focus distinctions are associated with. On the other hand, focus applications have very high identification rates in a variety of different linguistic contexts, including interactions with different sentence types (Chaida, 2010; Nikolaenkova, 2013).

A dislocation of the Object and thus the recipient at sentence beginning as a result of focus application has different perception effects in different contexts. If the Object is casus marked, high identification rates are achieved (Nikolaenkova, 2013). If, however, the Object is morphologically neutral, i.e. neutral nouns in Greek, the dislocation of the Object results to semantic ambiguity, according to which the neutral S-leading syntactic structure prevails. Thus, there is no perception distinction between Object in focus vs. Subject in focus at sentence beginning.

The results of this study indicate that prosodic focus, syntax and morphology have a variety of interactions with reference to semantic relations and the interpretation of spoken utterances in Greek. In particular, morphological variability may be a decisive factor with distinctive functions in syntactic structures and semantic relations across different linguistic contexts.

Acknowledgements

Thanks to the University of Athens for a travel grant to Fonetik 2013 conference. Thanks also to Elina Nirgianaki and Marios Fourakis for comments and much useful feedback.

References

Botinis, A., Y. Kostopoulos, O. Nikolaenkova & Ch. Themistocleous. 2005. Syntactic and tonal correlates of focus in Greek and Russian. In: A. Eriksson & J. Lindh (eds.), *XVIII National Phonetics Conference FONETIK 2005 Proceedings*, 99–103.

Chaida, A. 2010. *Production and Perception of Intonation and Sentence Types in Greek.* PhD thesis, University of Athens.

Nikolaenkova, O. 2013. *Multifactor analysis of Greek and Russian sentence*. PhD Thesis. Athens: University of Athens.

Nikolaenkova, O., A. Chaida & A. Bakakou-Orphanou. 2011. Focus perception of syntax and intonation in Greek. *7th AISV National Conference*, Università del Salento, Lecce.

Nikolaenkova, O. 2010. Perception of tonal focus in Greek. In: A. Botinis (ed.), *3rd ISCA Workshop ExLing* 2010 *Proceedings,* 121–124.

Audience response system based annotation of speech

Jens Edlund [1], Samer Al Moubayed [1], Christina Tånnander [2] & Joakim Gustafson [1]
[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2] Swedish Agency for Accessible Media, MTM, Stockholm, Sweden

Abstract

Manual annotators are often used to label speech. The task is associated with high costs and with great time consumption. We suggest to reach an increased throughput while maintaining a high measure of experimental control by borrowing from the Audience Response Systems used in the film and television industries, and demonstrate a cost-efficient setup for rapid, plenary annotation of phenomena occurring in recorded speech together with some results from studies we have undertaken to quantify the temporal precision and reliability of such annotations.

Introduction

We present a cost-efficient and robust setup for plenary perception experiments based on existing consumer market technologies. In an experiment in which 26 subjects divided in 4 groups react to sets of auditory stimuli by pressing buttons, we show that humans annotating as a group, in real time, can achieve high precision for salient acoustic events, making this a feasible alternative to expert annotations for many tasks. The experimental results validate the technique and quantify the expected precision and reliability of such plenary annotations on different tasks.

Over the past decades, corpus studies of speech and spoken interaction have been increasingly common, for purposes ranging from basic research to analyses undertaken as a starting point in efforts to build humanlike spoken dialogue systems (Edlund et al., 2008). This has led to a dramatic increase in numbers of data collections of human interactions and in the sheer amounts of data that is captured per hour of interaction in the resulting corpora, as exemplified by massively multimodal corpora such as the AMI Meeting Corpus comprised of a large number of video and audio channels as well as projector output (McCowan et al., 2005) or our own D64 and Spontal corpora (Edlund et al., 2010; Oertel et al., 2010) combining multiple video and audio channels with motion capture data.

While these corpora are useful, the task of annotating them is daunting. It is becoming near impossible to produce annotations in the traditional manner, with one or more highly skilled, highly trained experts spending several times real time or more for each annotation type – even a simple task such as correcting utterance segmentations that are already largely correct requires 1–3 time real time (Goldman, 2011). For many types of annotation, we may also raise a question related to ecological validity: why should it be so hard to label what people perceive effortlessly in each everyday conversation?

This rapidly growing demand for annotation has led to an increasing interest and use of crowdsourcing services such as Amazon Mechanical Turk. But although Mechanical Turk annotations are reported to be cheap, fast and good enough (Novotney & Callison-Burch, 2010), crowdsourcing means that we relinquish control over the situation in which the annotation is made.

To achieve increased throughput while maintaining a high measure of experimental control, we merge the ideas behind the Rapid Prosody Annotation pioneered by Jenifer Cole (see Mo, 2010) with a technical solution borrowed from the Audience Response Systems used in the film and television industries with the goal of having laymen annotators annotate some events in real time. Apart from being fast and cost effective, it places annotators in a situation similar how they perceive speech on an everyday basis.

We have previously investigated the use web based ARS-based methods for perception experiments (Edlund et al., 2012) and evaluation of children's speech (Strömbergsson & Tånnander, submitted). The present system has been used for speech synthesis evaluation (Tånnander et al., submitted), and is currently being tested for annotation of the type of speech events that are sometimes describes as disfluencies in the literature (Edlund et al., submitted).

ARS systems are used for screenings of new films and television series. They are typically expensive, and their software proprietary. In order to make an ARS-based system that is affordable for academic research, we built our system (see *Figure 1*) of Xbox 360 Controllers for Windows. These devices use low-latency wireless communication over a receiver which supports four controllers, but we have been able to robustly use more than one receiver per computer, allowing us to use more frame synchronized controllers simultaneously.

Figure 1. The system: Xbox 360 Controllers and Receivers in a custom made portable case.

To capture the controller states, we developed a Java library based on the DirectInput Microsoft API. The software automatically captures all controllers connected to the system by querying them for the state of their buttons, triggers and pads at a specified frame rate. The capture software can read the state of all input components on the controller (analogue and digital).

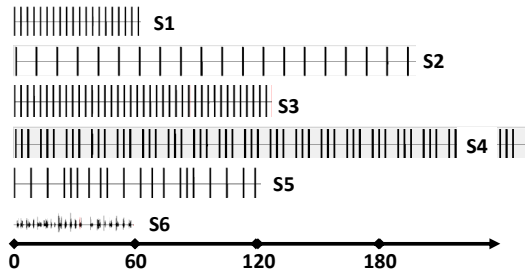


Figure 2. The waveforms of the six stimuli sets S1 – S6 (S6 is not included in the study).

Method

The purpose of this study to provide baseline statistics that show what can be expected of such a system. In order to achieve this goal, we ran a series of tests with stimuli that are significantly clearer than most speech phenomena, allowing us to measure with accuracy the temporal precision of the system (for such stimuli).

We used 4 groups of subjects, with 8, 8, 7 and 3 participants, respectively, for a total of 26 subjects. The subjects within a group took the test simultaneously, in full view of each other. All subjects were computer science students participating as part of a class. 10 of the subjects were female, 16 male.

Stimuli series S1 through S5 (*Figure 2*) all consist of 1 second long beeps at 170 Hz, but their temporal patterns vary, as do the task of the annotators. The first series S1 consists of 20 beeps spaced evenly, so that their onset occur regularly every 3.6 seconds. S2 also contains 20 evenly spaced beeps, but at a larger interval: 11.9 seconds. S3 contains 40 beeps, spaced evenly as in S1. S4 contains 60 beeps, presented in groups of three within which the beeps are spaced evenly at 3.5 seconds, as in S1. Between groups, there is an 8 second spacing corresponding to the duration an extra beep would have taken up: 3.5 seconds for the first spacing, 1 second for the beep, and 3.5 seconds for the second spacing. S5 holds 20 beeps spaced irregularly, at random intervals of up to 10 seconds.

The subjects were presented with the stimuli sets one after the other. For each set, they were told what to expect (i.e. if the clicks would be regular or irregular, and roughly at what intervals). They were also instructed as to what they should react to (click on). For S1, S2, and S5, they we simply asked to click as close to beep onset as possible. For S3, they were told to click as close to every other beep onset as possible, starting with a click of their own choice (the first or the second). For S4, they were asked to click where the left out fourth beep in every series of three *should have been*. In other words, they were asked to click at something that was lacking in the stimuli, rather than present.

Results

There were no signs of equipment failure or code malfunction. In group 2, there were 615 instances of double clicks within less than 10 ms. The latter of each of these was removed. Once this was done, between-group differences were negligible.

Overall, the subjects performed well and did what they were asked to do: subjects produced on average 1 click per occasion.

All clicks could easily and automatically be related to the stimuli by virtue of appearing shortly after stimuli onset, and long before the next stimuli onset.

For all stimuli series except S4, there is a very clear effect of the onset of the stimuli series. The average response time to the first stimuli in a series is 2–4 times larger than that of any of the remaining 19. We therefore treat the first stimuli in each series as an outlier and remove it from further analysis.

Table 1. Average response times (ms) for all subjects over stimuli types, with standard deviation, counts, and significance at 4 % level versus the rest of all stimuli types.

	Mean	Stddev.	N
S1	525	162	498
S2	615	260	501
S3	528	226	518
S4	696	1276	518
S5	592	130	495

Table 1 shows the mean response times to stimuli types. The differences are significant (one-way ANOVA, $p < 0.0001$).

For purposes of using our system for annotation of speech and interaction phenomena, we need an analysis of response time distribution that allows us to predict where, in a continuous data stream, the unknown event that caused a group of subjects to click is most likely to be found. Instead of histograms, we perform this analyses using Kernel Density Estimation (KDE) estimations, which produces an analysis that is similar to a histogram, but produces a continuous, smooth curve.

The stimuli sets S1 and S3 (both with a predictable 3.5 second interval, with the task of clicking only every second beep for S3) show similar distributions. S2 (a predictable 11.7 second interval) has a wider distribution, S4 has a very different distribution spanning well over 2 seconds. S5 shows the most narrow distribution of all.

In the remainder of this analysis, we use reaction time estimates acquired by finding the peak of these curves, rather than using averages. Two sets of reaction times are used: for descriptions, we base reaction time estimates on all subjects. For error estimation, we use reaction time estimates based only on group 1 (8 subjects), which is then held out from the testing.

Our ultimate goal is to use subjects' clicks to localize events in streaming data. As a first step, we return to KDE: we build an estimate over an entire stimuli set by adding a narrow width (0.5 s) Gaussian for each click of each subject relative to the start of the session. The resulting curve for S1 along with the wave form of the stimuli is shown in *Figure 3*.

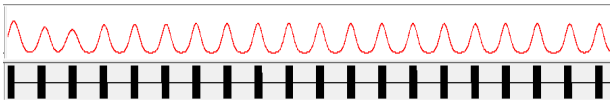


Figure 3. Overview of the waveform of S1 (below) with the KDE estimate based on all all subjects (above).

We estimate the onset of a trigger by finding a peak P in the KDE curve for a stimulus and deducting the RT estimate from that stimulus from the peak position.

Table 2. The mean error in ms (the mean of the absolute difference between actual trigger positions and estimated trigger positions) for each stimuli set.

	Mean	Std. dev.	N	Sig.
S1	23	33	19	-
S2	9	11	19	**
S3	28	25	39	-
S4	61	45	19	**
S5	25	22	19	-

Table 2 shows the errors for trigger estimates when both RT estimates and KDE curves are based on all participants. Overall differences were tested with a one-way ANOVA ($p < 0.001$). Differences between each set and all other sets were tested with pairwise *t*-tests and are reported in the table.

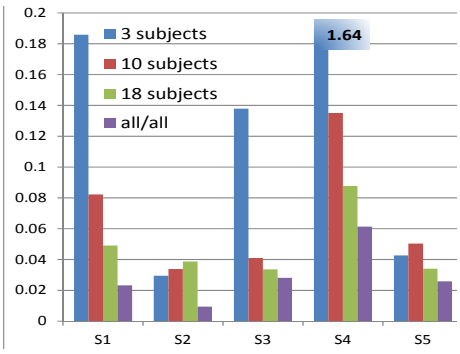


Figure 4. Average errors for S1-S6 for trigger estimates based on the clicks of 3, 10 and 18 subjects and RT estimates based on 8 different subjects, plus estimates based on clicks and RT times from all subjects. Note that two bars are truncated.

In order to investigate the effect of the number of subjects used on the reliability of the trigger estimates, we use one group (group 1 of 8 subjects) to estimate RT and then use group 4 (G_4, 3 subjects), group 3 and 4 (G_{34}, 10 subjects) and group 2, 3, and 4 (G_{234}, 18 subjects) to estimate trigger times for each trigger in each stimuli set.

Figure 4 shows the mean error for each stimulus set for trigger estimates based on 3, 10, and 18 subjects (all disjoint from G1, the subjects used to estimate RT for each stimuli type), and also for all 28 subjects, using RT estimates from the same 28 subjects.

One-way ANOVAs within each stimulus set as well as within all stimuli all show a significant overall effect of number of subjects ($p < 0.0001$ in all cases).

Future work

We will attempt to add more game controllers to the system, in order to be able to get more out of one single, controlled test series. Optimally, we would like to be able to run groups of 16 or 24 subjects. Further technical development includes making use of the feedback systems provided in the controllers: they are able to vibrate and have a small number of LEDs that could be used for this.

Regarding actual annotation, we are in the process of annotating language phenomena such as filler pauses, hesitations and repairs. We will follow up on these annotations and compare them, from efficiency and from an accuracy point of view, to traditional expert annotation.

Acknowledgements

This work was funded by the *GetHomeSafe* project (EU 7th Framework STREP 288667) and by the Swedish Research Council (VR) project *Introducing interactional phenomena in speech synthesis* (2009-4291).

References

Edlund, J., C. Hjalmarsson & C. Tånnander. 2012. Unconventional methods in perception experiments. In *Proc. of Nordic Prosody XI*. Tartu, Estonia.

Edlund, J., J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson & D. House. 2010. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta, Malta, 2992–2995.

Edlund, J., J. Gustafson, M. Heldner & A. Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication, 50*(8–9):630–645.

Edlund, J., S. Strömbergsson & J. Gustafson. Submitted. Audience response system-based annotation of conversational speech phenomena. Submitted to *Proc. of DiSS 2013*. Stockholm.

Goldman, J.-P. 2011. EasyAlign: a friendly automatic phonetic alignment tool under Praat. In: *Proceedings of Interspeech 2011*. Florence, Italy, Ses1-S3:2.

McCowan, I., J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma & P. Wellner. 2005. The AMI Meeting Corpus. In *Proc. of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research*. Wageningen, Netherlands.

Mo, Y. 2010. *Prosody production and perception with conversational speech*. Doctoral dissertation, University of Illinois at Urbana-Champaign.

Novotney, S. & C. Callison-Burch. 2010. Cheap, fast and good enough: automatic speech recognition with non-expert transcription. In: *Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 10)*, 207–205.

Oertel, C., F. Cummins, N. Campbell, J. Edlund & P. Wagner. 2010. D64: A corpus of richly recorded conversational interaction. In: M. Kipp, J.-C. Martin, P. Paggio & D. Heylen (eds.), *Proceedings of LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, Valetta, Malta, 27–30.

Strömbergsson, S. & C. Tånnander. Submitted. Correlates to intelligibility in deviant child speech – comparing clinical evaluations to audience response system-based evaluations by untrained listeners. Submitted to *Proc. of Interspeech 2013*. Lyon, France.

Tånnander, C., J. Edlund & J. Gustafson. Submitted. Audience response system-based evaluation of speech synthesis. Submitted to *Proc. of Interspeech 2013*. Lyon, France.

Does production facilitate discrimination? An infant mismatch negativity study

Isabelle Edström[1], Lisa Gustavsson[2], Petter Kallionen[2], Marie Markelius[1, 2], Andrea Strandberg[1], Nina Strömberg[1] & Katarina Svensson[1]
[1]Department of Clinical Science, Intervention and Technology, Karolinska Institute, Stockholm
[2]Department of Linguistics, Stockholm University, Stockholm

Abstract

MMN is an ERP component that reflects pre-attentive discrimination between a recurring standard sound and a deviating sound. MMN is frequently used in infant studies focused on speech development since its elicitation does not require the attention of the child. The general ability of infants to discriminate speech sounds is gradually specialized towards discrimination of phonetic contrasts in their mother tongue. The aim of the present study was to examine if an MMN response is elicited by naturally varying speech stimuli (/ti/ and /ki/) and if this response is stronger for the speech sound that infants typically produce at this age (/t/). An EEG experiment with an oddball paradigm was designed. Participants were 19 infants (9-mo). An MMN-like negative response to deviants compared to standards was found, however it was not statistically significant. No significant interaction effect was found for MMN and type of deviant stimulus. Variation in the standard stimuli may have contributed to the lack of effect. It is also possible that the infants already were equally competent in discriminating both speech sounds, which may account for the small difference between the deviant waveforms.

Introduction

Mismatch negativity (MMN: Näätänen, Gaillard & Mäntysalo, 1978) is an ERP-component frequently studied with EEG-recordings. MMN is elicited by discriminable changes in an otherwise uniform stream of stimuli (Näätänen, Gaillard & Mäntysalo, 1978) and is usually studied in oddball paradigms (first used by Squires, Squires & Hillyard, 1975). The auditory MMN response is elicited by changes in frequency or amplitude for sinusoidal sounds (Sams et al., 1985; Ford & Hillyard, 1981) and to changes in complex sounds, such as speech sounds (Pakarinen et al., 2013). The MMN response is the result of a comparison process in the brain, between a memory trace formed by the recurring stimuli and a deviating sound (Näätänen, 1990). For adults, the MMN latency is typically 100–200 ms (Näätänen, Gaillard & Mäntysalo, 1978, 1978; Näätänen & Alho, 1997) whereas it is longer in infants (250–450 ms) (Cheour et al., 1998; Conboy & Kuhl, 2011). Its distribution is generally strongest at fronto-central sites (Čeponiene et al., 2004).

MMN is particularly appropriate for infant studies considering it is elicited by unattended stimuli (Näätänen, 1991), in comparison to other behavioural methods such as head turning techniques and high-amplitude-sucking.

Late difference negativity (LDN) is another ERP component that has been linked to sound change processing in MMN studies (Čeponiene et al., 2004; Alho et al., 1992), with a peak latency of 350–500 ms (Korpilahti et al., 2001).

Typical speech development in children follows a certain pattern where dental and bilabial sounds (e.g. /t/, /d/, /b/ and /p/) are produced early. Velar speech sounds (e.g. /k/ and /g/) are typically produced later (e.g. Locke, 1983; Lohmander, Olsson & Flynn, 2011; McCune & Vihman, 2001). Newborns' competence to discriminate phonetic contrasts of their mother tongue improves during their first year at the cost of a more general ability to discriminate phonetic contrasts from virtually any language (Kuhl et al., 2006). EEG studies have revealed that the MMN response to non-native speech sounds is attenuated and the response to native speech sounds increases by the time the infant is 12 months old (Cheour et al., 1998).

The present study will focus on 9-month-old infants' discriminatory ability by measuring MMN in response to naturally produced speech sounds. The aim of the study is to see if there is an auditory MMN elicited in response to dental and velar consonants respectively and further to examine whether there is a difference between the MMN waveform for the two deviating stimuli. It is plausible that the MMN response would be stronger for when the dental consonant serves as a deviant, considering

9-month-old infants typically can produce dental sounds. The idea would be that production facilitates discrimination.

Method

Participants

The participants were 19 healthy infants (10 female; m=9 months, 14 days; sd=9.16 days). All participants' mother tongue was Swedish and four came from bilingual homes (Italian, Finnish, Japanese, and Serbian).

Letters with information about the study were sent to 200 randomly selected families accessed from the National Swedish address register, based on the infant's date of birth (pre-term infants were not excluded). Informed consent was obtained from all caregivers and the study was conducted with the approval of the local ethical committee (Dnr 2011/955-31/1). The participants received a diploma for their participation.

Stimuli and experimental design

The Swedish CV syllables /ti/ and /ki/ were pronounced by a native Swedish speaking female and recorded in an anechoic chamber. The syllables were produced in a standard sentence *(Jag sa /x/ till de andra)* multiple times with deliberately varied stress patterns in order to acquire naturally varying speech sounds. The speech sounds were manually edited in Wavesurfer (ver. 8.4.2.9) and 18 stimuli were chosen for the experiment (9 /ti/; 9 /ki/). An oddball paradigm design was used with two blocks of stimuli. The speech sound /ti/ served as standard (p = 80%) and /ki/ as deviant stimuli (p = 20%) in one of the blocks (dev/ki/). /ki/ served as standard (p = 80%) and /ti/ as deviant (p = 20%) in the other block (dev/ti/). The order in which the blocks were presented was randomized for every subject, with an interstimulus interval (ISI) of 600 ms. The deviant stimuli occurred at pseudorandom order but was always preceded by at least two standard stimuli. Each block was initiated with 12 standard stimuli and consisted of 60 deviant and 240 standard stimuli. The total length of the experiment is 12.4 min.

Procedure

The participants were seated in the caregivers' lap on a chair placed approximately 70 cm from a computer screen in a sound-attenuated room. A children's TV-program (In The Night Garden, BBC) was used as a source of distraction during the experiment. Loudspeakers (NuForce S-1) were placed on each side of the screen, through which the auditory stimuli was presented at approximately 60–70 dBA. E-Prime 2.0 was used to run the experiment.

A HydroCel Geodesic Sensor Net (128 channels) was used. During recording vertex reference was used and impedance was held below 50kΩ for most channels. EGI hardware and software (Electrical Geodesics, Inc.) were used during recording.

Processing of data

Analysis of data was performed with Net Station 4.2, using a bandpass filter set to 1–40 Hz (FRI). Epochs of 800 ms including a 200 ms pre-stimulus period were separately averaged for the standards and deviants respectively. Segments and channels were rejected when exceeding 200 μV in the relevant channels and data was re-referenced to mastoids. The mean voltage of the pre-stimulus (200 ms) period served as baseline. A grand average file was created for all subjects, after omitting four participants due to noisy data.

Based on visual inspection a time-window of 150–300 ms was chosen. The MMN parameters were computed from the difference waveform by subtracting the grand average standard-stimulus ERP from the grand average deviant-stimulus ERP.

Statistical analysis

A multivariate ANOVA (2 × 2, repeated measures) was performed on the average data for the frontal electrode Fz (11), with the variables deviant stimuli, standard stimuli and the two blocks (dev/ti/, dev/ki/), respectively.

Results

The deviant stimuli elicited an MMN response (peak 150–300 ms), and a later negativity at 500–600 ms, measured at electrode Fz (see *Figure 1*). Another early ERP that was elicited (peak 80–100 ms) can be seen for the average waveform for the deviant stimuli, as seen in *Figure 1*. There is a small observable difference in the average waveform depending on block type (dev/ki/ vs. dev/ti/) at 150–300 ms, measured at electrode Fz, as seen in *Figure 2*.

The ANOVA (2x2) showed an insignificant MMN effect ($F_{1,14}$ = 2.615, p < s0.05, ns.), and an insignificant interaction between the two blocks ($F_{1,14}$= 0.171, p < 0.05, ns.), see *Figure 3*.

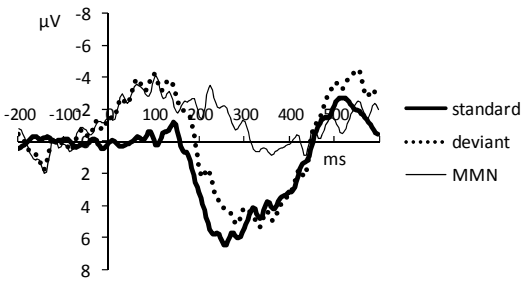


Figure 1. Auditory ERPs measured at frontal (Fz). The average waveform to the standard stimuli (thick line), deviant stimuli (dotted line), and the MMN shown as a difference wave (thin line). The time window is one epoch (800 ms).

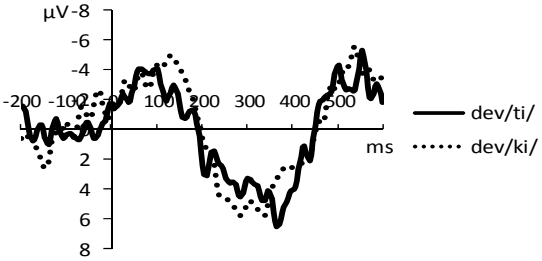


Figure 2. Auditory ERPs measured at frontal (Fz). The average waveform to the deviant stimuli from the blocks dev/ti (thick line) and dev/ki/ (dotted line). The time window is one epoch (800 ms).

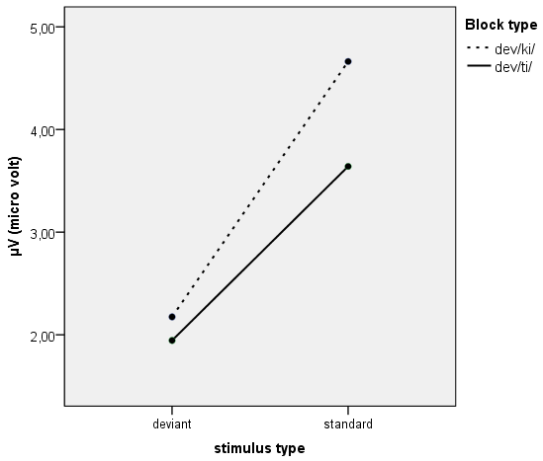


Figure 3. There is an observable difference in amplitude (µV) depending on stimulus type (standard vs. deviant) for both blocks, however not statistically significant. There was no observable interaction effect between MMN and block type.

Discussion

The aim of the present study was to examine the discriminatory ability for dental and velar sounds in 9-month-old infants. The primary purpose was to see if an MMN response was elicited for the speech sounds /ti/ and /ki/. The results indicate a negativity similar to the MMN component, with a latency of 150–300 ms. However, the ANOVA indicated no significant effect of stimulus type (standard vs. deviant). The insignificant result might be related to the fact that the number of participants was relatively low as well as some of them being omitted due to noisy data. A prerequisite of MMN elicitation is that a memory trace of the standard stimuli has been formed. It is likely that this neural representation is weaker when using varying complex sounds, as was the case in this study. Furthermore, previous studies have suggested that MMN amplitude is higher for tone stimuli than for more complex linguistic stimuli (Korpilahti et al., 2001). Nevertheless, it is of importance to study the MMN response to naturally produced speech sounds with varying characteristics since it is more applicable to non-experimental contexts.

Moreover, this study aimed to analyse whether there is a difference in amplitude between the two waveforms elicited by the deviant stimuli. The results showed no significant interaction effect with MMN. The average waveform representing the block dev/ti/ is slightly more negative, for the time window 150–300 ms. Even if this is in line with the hypothesis of the study, the results are not significant, and hence cannot confirm it. The infant's native language specialization during its first 12 months results in a greater competence in discriminating mother tongue contrasts (Cheour et al., 1998). The infants in this study are 9 months old, and it is plausible that, contrary to our hypothesis, they are already equally competent in discriminating the two speech sounds /t/ and /k/.

Due to the participants' young age, movement artefacts were difficult to avoid completely, which is the case in any infant study.

The negativity with a 500–600 latency of has the characteristics of the ERP-component LDN, which has been suggested to reflect automatic stimulus change detection, primarily in children, (Alho et al., 1992), however the function of LDN is still unclear. Other studies have suggested that the LDN reflects a later stage in sound processing (Čeponiene et al., 2004). This response however has not been further statistically analysed in this study. The early negative component with a peak around 80–100 ms is apparent in the deviant

but not the standard waveform. This response has previously been related to the neuronal refractory period, not associated with the MMN (Walker et al., 2001).

Studying infants' discriminatory ability by measuring MMN provides interesting insights in language development. Variation in the speech stimuli is essential when studying phonetic contrasts rather than acoustic differences. Possibly the present study pushed this variability too far.

References

Alho, K., D.L. Woods, A. Algazi & R. Näätänen. 1992. Intermodal selective attention. II. Effects of attentional load on processing of auditory and visual stimuli. *Electroencephalography and Clinical Neurophysiology* 82:356–368.

Čeponiene, R., T. Lepistö, M. Soininen, E. Aronen, P. Alku & R. Näätänen. 2004. Event-related potentials associated with discrimination versus novelty detection in children. *Psychophysiology* 41:130–141.

Cheour, M., R. Čeponiene, A. Lehtokoski, A. Luuk, J. Allik, K. Alho & R. Näätänen. 1998. Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience* 1(5): 351–353.

Conboy, B. T. & P. K. Kuhl. 2011. Impact of second-language experience in infancy: brain measures of first- and second-language speech perception. *Developmental Science* 14(2):242–248.

Ford, J.M. & S.A. Hillyard. 1981. Event-related potentials (ERPs) to interruptions of a steady rhythm. *Psychophysiology* 18(3):322–330.

Korpilahti, P., C. M. Krause, I. Holopainen & A. H. Lang. 2001. Early and Late Mismatch Negativity Elicited by Words and Speech-Like Stimuli in Children. *Brain and Language* 76:332–339.

Kuhl, P., E. Stevens, A. Hayashi, T. Deguchi, S. Kiritani & P. Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science* 9(2):F13–F21.

Locke, J. 1983. *Phonological acquisition and change*. New York: Academic Press. (With a foreword by Michael Studdert-Kennedy.)

Lohmander, A., M. Olsson & T. Flynn. 2011. Early consonant production in Swedish infants with and without unilateral cleft lip and palate and two-stage palatal repair. *Cleft Palate–Craniofacial Journal* 48(3): 271–285.

McCune, L. & M. M. Vihman. 2001. Early phonetic and lexical development: a productivity approach. *Journal of Speech, Language, and Hearing Research* 44:670–684.

Näätänen, R. 1991. Mismatch negativity outside strong attentional focus: a commentary on Woldorff et al. *Psychophysiology* 28(4):478–484.

Näätänen, R. 1990. The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences* 13:201–288.

Näätänen, R. & K. Alho. 1997. Mismatch negativity—the measure for central sound representation accuracy. *Audiology & Neuro-otology* 2(5):341–353.

Näätänen, R., A. W. K. Gaillard & S. Mäntysalo. 1978. Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica* 42(4):313–329.

Pakarinen S., T. Teinonen, A. Shestakova, M.S. Kwon, T. Kujala, H. Hämäläinen, R. Näätänen, M. Huotilainen. 2013. Fast parametric evaluation of central speech-sound processing with mismatch negativity (MMN). *International Journal of Psychophysiology* 87(1):103–110.

Sams, M., P. Paavilainen, K. Alho & R. Näätänen. 1985. Auditory frequency discrimination and event-related potentials. *Electroencephalography and Clinical Neurophysiology* 62(6):437–448.

Squires, N.K., K.C. Squires & S.A. Hillyard. 1975. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology* 38 (4):387–401.

Walker, L.J., M. Carpenter, C.R. Downs, J.L. Cranford, A. Stuart, D. Pravica. 2001. Possible neuronal refractory or recovery artifacts associated with recording the mismatch negativity response. *Journal of the American Academy of Audiology*. 12(7):348–356.

An acoustic comparison of voice characteristics in 'kulning', head and modal registers

Robert Eklund[1], *Anita McAllister*[2] *& Fanny Pehrson*[2]
[1] *Department of Culture and Communication, Linköping University, Linköping, Sweden*
[2] *Faculty of Health Sciences, IKE/Speech and Language Pathology, Linköping University, Linköping, Sweden*

Abstract

The Swedish folk singing style 'kulning' is surprisingly understudied, despite its almost mythical status in Swedish folklore. While some physiological–productive aspects of kulning have been treated in previous work, acoustic properties are still much lacking description. This paper compares kulning, head ('falsetto') and modal voice from an acoustic perspective.

Introduction

Throughout history, long-distance calls have been created at several different locations where there has been a need of making oneself heard over long distances. Examples include e.g. yodeling (Luchsinger, 1942) and whistled languages (Busnel & Classe, 1976), and such long-distance calls have been used both for human–human communication and for human–animal calling.

Kulning is the most common term (see Rosenberg, 2003:8, for an extensive listing of alternative words) for a specific type of cattle or herding calls used in some parts of Sweden (e.g. Dalarna, Härjedalen and Jämtland) and is used to call cows or goats gracing freely in the mountains when it is time for milking. The tradition of cattle calls dates far back in history and was most common in the province of Dalarna, Sweden, where young women looked after the live stock during summer in small mountain farms, away from the homestead. Kulning generally has no lyrics and consists of vowel-heavy syllables that feel comfortable to the singer. The singing technique is often used by women, and is high-pitched, without vibrato to make the sound carry over long distances. Despite its well-nigh mythical status in Swedish folklore, kulning has received surprisingly little attention from a research point of view. This paper aims at remedy this situation by looking at some of the acoustic properties of kulning, and comparing them to head ('falsetto') and modal register singing.

Previous research

Kulning is mentioned in both Moberg (1955:38 et passim) and Ling (1978), but mainly in passing. For example, Ling (1978:22) states that kulning is not really "singing" in a traditional sense but is more like some kind of falsetto-like calling in very high registers, and that it requires a tightened larynx, while Moberg (1955:37) points out that it is normally sung on vowels, without lyrics in the traditional sense.

Johnson (1984, 1986:216–259) reports that kulning production is characterized by a strong correlation between frequency and amplitude in higher registers (not so much so in lower registers) and that, contrary to classical singing, the larynx moves with the frequency, and is raised considerably when high notes are produced (up to +39 mm). Jaw opening is also correlated with high frequency (in line with classical singing). The vocal tract length is varied with up to 37 mm, compared to 22 mm in normal singing: Also, the pharynx is tightened, even to the point of making optical glottography impossible. Johnson (ibid.) also reports SPL values up to 105 dB (without mentioning any reference values). The results presented in Johnson are largely repeated in Rosenberg (2003:24).

As for the acoustic properties of kulning, Uttman (2002) studied partials spectra of kulning songs obtained from outdoor recordings, and reported strong partials up to the 16–18 kHz register, compared to ~6 kHz in normal folk singing.

Data collection

Data from one female subject (the third author) were collected in two settings: a normal room approximately 5.5 × 7 meters, and in an anechoic chamber (AC), slightly smaller in size.

Recordings were calibrated at 88–90 dBa using a sustained vowel and a sound-level meter (Brüel and Kjaer 2215). The results were announced on the recording. Sound was doubly recorded, but the recordings used in this paper were made using a professional Audiotechnica

AT813 cardoid-pattern, condenser mono microphone that fed into a high-definition video camera (Canon HG-10).

The third author (FP) is educated in kulning at Musikkonservatoriet in Falun and Malungs Folkhögskola, and by Agneta Stolpe and Ann-Sofi Nilsson. Data consisted of FP singing three versions of a cattle call (cattle call from Äppelbo in a traditional arrangement by Agneta Stolpe, Vallslinga från Äppelbo) in each of the two rooms described above. Each song was sung in three ways: (1) kulning voice; (2) head register (sometimes incorrectly referred to as "falsetto"); and (3) modal voice (chest register). Each version of the song was initiated by giving the start pitch using a pitch pipe. The starting tone had the same F_0 independently of room condition. The duration of the song in all of the different versions/singing techniques was approximately one minute.

The recording sessions were unproblematic, although FP reported that singing in an indoors setting, without the characteristic outdoors echo, was a new experience which possibly affected the performance. This was especially true for the anechoic chamber, of course. In both rooms, and for the above reason, FP reported having problems producing a really loud tone, something which was somewhat exacerbated by the fact that FP recently had recovered from a light cold.

Analysis

For analysis, the following post-processing was carried out. Data were resampled to 44.1 kHz, 16 bit, mono. The six different versions were carefully excised from the audio files so as to omit all extraneous sound (like the authors discussing recordings settings). From these files, the first high-pitched [ɥ] was excised. The frequency was around 670 Hz, corresponding to (a somewhat sharp) E5.

The fundamental frequency mode value is the most frequent value occurring in the song. The values were obtained from the entire extracted song in head, modal and kulning registers and for both settings, and F_0 was extracted using Soundswell. The files were all low pass filtered at 1000 Hz, with a maximum frequency 1100 Hz and with a high pass filter set at 40 Hz. As mentioned above, while F_0 mode values were obtained from the modal/chest register and the two conditions, due to the frequency difference of one octave (the modal version being sung one octave lower than the head and kulning versions) no comparisons between modal and kulning/head versions were made. Modal register singing comparisons with kulning/head register singing were only made based on the two different room conditions.

Additional analyses were carried out using Cool Edit Pro 2.0, Cool Edit 2000, WaveSurfer 1.8.8p4 and Soundswell.

Results

Fundamental frequency

First, fundamental frequency in the two settings was examined. The results show a clear effect on F_0 mode value as a function of the two room conditions; see *Figure 1*.

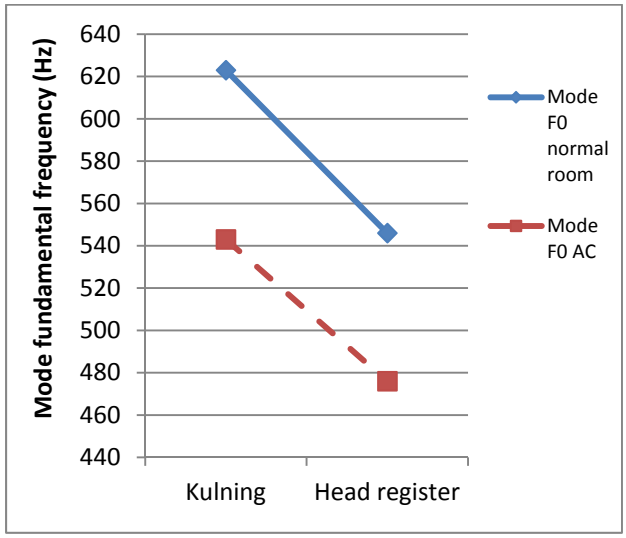


Figure 1. Fundamental frequency mode values for kulning and head register singing in the normal room and in the anechoic room (AC).

The observed effect was 80 and 70 Hz for kulning and head register, respectively, showing that F_0 mode value for kulning was higher in both recording conditions. The observed room difference was 67–77 Hz higher in kulning for the normal room and the anechoic chamber, respectively. The room effect for modal/chest register showed the opposite pattern: here a higher F_0 mode value of 20Hz for the anechoic room was observed.

Our general observation that kulning is higher in frequency than other singing more or less repeats the results reported in Johnson (1986:253) who compared kulning and normal folk singing, and observed roughly similar frequencies for one vowel/tone, [ɥ], and higher frequency in kulning for another [ɥ] vowel/tone.

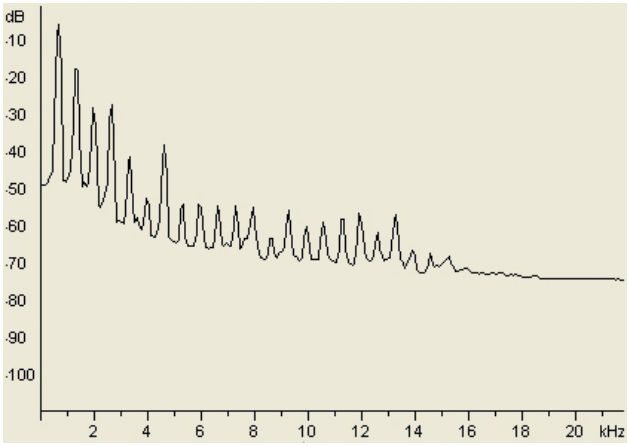


Figure 2: Kulning [ɬ] produced in normal room, LTAS/FFT/Hamming analysis.

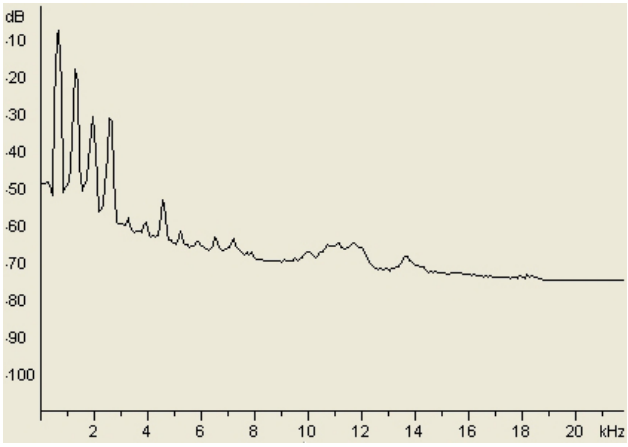


Figure 3: Head [ɬ] produced in normal room. LTAS/FFT/Hamming analysis.

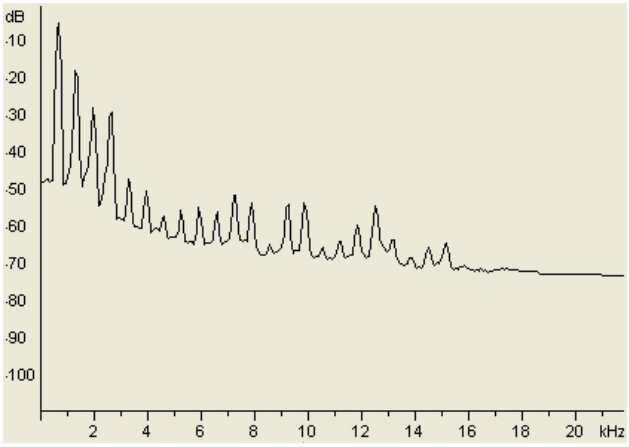


Figure 4: Kulning [ɬ] produced in echo-free room, LTAS/FFT/Hamming analysis.

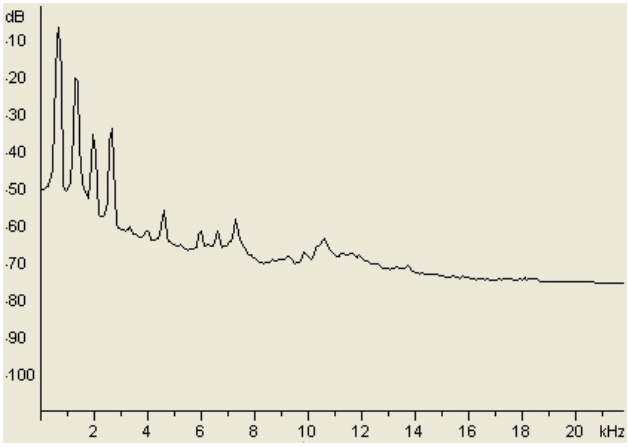


Figure 5: Head [ɬ] produced in echo-free room. LTAS/FFT/Hamming analysis.

Long-Term Average Spectrum analysis

The first [ɬ] in the song was then extracted in the kulning and head register versions (for two room conditions) and a Long-Term Average Spectral (LTAS) analysis (using Fast Fourier Transform/Hamming window) was performed. The results are presented in *Figures 2–5* above.

As can be seen, the partials are much more distinguishable in kulning than in head voice, under both recording conditions, and while only four peaks are clearly visible in head register, and more or less disappear above 5 kHz, partials in kulning register are easily observed up to 15 kHz before fading out.

Discussion

First, it should be mentioned that both recording settings described in this paper are somewhat counterproductive from a kulning point of view, where the basic concept is crucially based on an outdoor setting, far from the echo-attenuating feedback encountered here. FP's previous experience of kulning has, quite naturally, been outdoors or at least in large rooms with good acoustics and feedback properties, which means that both recording conditions here were new and somewhat artificial to her, which likely affected the performance. Naturally, especially the anechoic chamber was perceived as difficult to accommodate to. This is likely the underlying cause behind the observed lower F_0 mode values observed for both kulning and head register singing. Interestingly, however, this effect was reversed in modal/chest register, where a higher F_0 mode value was observed in the anechoic room. A possible explanation for this could be that this was the last recording and that the singer experienced a warm-up effect after performing the other songs.

With regard to singing technique, F_0 mode value was higher for kulning compared to head register, independently of room setting, despite making sure that all recordings used the same starting tone.

The spectral comparison revealed a marked difference between kulning and head registers in that kulning exhibited distinct partials up to approximately 15 kHz, while only four partials were clearly visible in head register. Head register also showed a clear decline in palpable partials at around 5 kHz. A partial (pun intended) explanation for this could be ascribed to loudness differences between the two effects, but no such analysis was carried in this pilot paper of limited space. Given that reference dB values were carefully recorded, future studies could easily encompass a deeper study of this phenomenon.

It is interesting to compare our finding of clearly marked partials in kulning to Uttman (2002). She, too, reported very high partials for all four singers studied, with a variation between the singers ranging from 12 kHz (for the singer Maria Röjås) to 18 kHz (for the singer Lena Willemark). The other two singers (Agneta Stolpe and Kerstin Sonnbäck) studied had their highest (visible) partials at 16 kHz. Since Uttman used phonograms, there is no way (or at least very difficult) of knowing details about the recording conditions (e.g. microphone specifications), but while our own data are "clean" and controlled, Uttman's data are more ecologically valid. Ideally, future studies of kulning should use controlled outdoors recordings sessions, including dB reference values, coupled with detailed technical specifications. Outdoor recordings would also enable analysis of sound transmission characteristics for different singing styles in different outdoor environments (cf. Marten & Marler, 1977)

Future studies should/could also include analyses of glottal properties, like the closed quotient and the crest-factor. The closed quotient can be related to pressed phonation and the crest factor is related to closure speed. Both measures are related to vocal loudness. Measures on loudness should be included in future comparisons of kulning and head register, as already mentioned.

Finally, it is interesting to note that yodelling has been devoted some recent interest (Echternach, Marki & Richer, 2011; Echternach & Richter, 2010; Schlöminger-Their et al., 2009). From an acoustic perspective, comparisons between kulning and yodelling could be of potential interest given the similar rationales for the two singing styles.

Acknowledgements

The authors would like to thank speech therapy student Cecilia Eriksson whose inquisitive inquiry inspired, initiated and instigated the present study. Thanks to Susanne Schötz and Inger Lundeborg Hammarström for comments and last-minute proof-reading.

References

Busnel, R.-G. & A. Classe. 1976. *Whistled languages* (Communication and Cybernetics 13). Berlin: Springer.

Echternach, M., M. Marki & B. Richter. 2011. Vocal tract configurations in yodelling—prospective comparison of two Swiss yodeller and two non-yodeller subjects. *Logopedics Phoniatrics Vocology* 36:109–113.

Echternach, M. & B. Richter. 2010. Vocal perfection in yodelling—pitch stabilities and transition time. *Logopedics Phoniatrics Vocology* 35:6–12.

Johnson, A. 1986. *Sången i skogen: Studier kring den svenska fäbodmusiken*. PhD thesis, Department of Musicology, Uppsala University.

Johnson, A. 1984. Voice Physiology and Ethnomusicology: Physiological and acoustical Studies of the Swedish Herding Song. In: D. Christensen (ed.), *Yearbook for Traditional Music* 16, 42–66.

Ling, J. 1978. *Svensk folkmusik*. Stockholm: Prisma.

Luchsinger, R. 1942. Untersuchungen über die Klangfarbe der menschlichen Stimme. *Archiv für Sprach- und Stimmphysiologie und Sprach- und Stimmheiligkunde*, 1–39.

Marten, K. & P. Marler. 1977. Sound Transmission and Its Significance for Animal Vocalization. *Behavioral Ecology and Sociobiology* 2:271–290.

Moberg, C.-A. 1955. Om vallåtar. En studie i de svenska fäbodarnas musikaliska organisation. *Svensk Tidskrift för Musikforskning*, 1–27.

Rosenberg, S. 2003. *Kulning. Musiken och metoden*. Stockholm: Udda Toner. 1950.

Schlömicher-Thier, J., D. G. Miller, H. Noé & C. T. Herbst. 2009. *Yodelling – acoustic and physiological properties*. Poster at the 38th Annual Symposium of the Voice Foundation, Philadelphia, PA, USA, 3–7 July 2009.

Uttman, M. T. 2002. Eine Untersuchung der Teiltonspektren bei Kulning und Lockruftechniken anhand von Beispielen aus Schweden und Finnland. *STM-Online* 5. http://musikforskning.se/stmonline/vol_5/ Also appears in *Systematische Musikwissenschaft* VI/4 1998; published in 1999 by ASCO Art & Science, Bratislava.

A comparative acoustic analysis of purring in juvenile, subadult and adult cheetahs

Robert Eklund[1] *& Gustav Peters*[2]
[1] *Department of Culture and Communication, Linköping University, Linköping, Sweden*
[2] *Forschungsmuseum Alexander Koenig, Bonn, Germany*

Abstract

Previous studies of cheetah purring have described purring in adult cheetahs. This paper extends the cheetah purring research to include juvenile and subadult cheetahs and analyzes purring data from cheetahs in ages ranging from 7 months to 7 years, and with weights ranging from 18 kilos to over 70 kilos. Results show that while there is considerable variation across most parameters analysed (amplitude, phase duration, cycles per phase and fundamental frequency), mainly attributable to degree of relaxation/agitation, previously reported observations that ingressive phases tend to be lower in frequency are largely confirmed, with one notable exception.

Introduction

Despite the fact that the purring domestic cat (*Felis catus*, Linneaus 1758) has been a companion of humans for around 10,000 years (Driscoll et al., 2009), and the fact that the prominent purrer, the cheetah (*Acinonyx jubatus*, Schreber 1776), also has been kept as a pet animal for thousands of years, it is still not known exactly how purring is produced.

Eklund, Peters and Duthie (2010) compared purring in the cheetah and the domestic cat, while Eklund et al. (2012) compared purring in four adult cheetahs.

However, the papers mentioned above studied purring in adult cheetahs. The present paper extends the previous studies by including analyses of juvenile/subadult cheetahs, ranging in age from around 7 months to 7 years, with a weight range of 18 kilos to more than 70 kilos.

The cheetah

The cheetah (*Acinonyx jubatus*) is probably best known for being the fastest land animal in the world with an estimated top speed of circa 112 km/h (Sunquist & Sunquist, 2002:23). Contrary to a widespread misconception that the cheetah "is not a cat", it is a full-fledged felid, most closely related to the puma (*Puma concolor*) and the jaguarundi (*P. yaguarondi*) (O'Brien & Johnson, 2007:70).

The cheetah is of roughly the same size as the leopard (*Panthera pardus*) – with which it is often confused – but is of a lighter and more slender build, has a smaller head and smaller teeth. The cheetah is distinguished by dark tear-marks in the facial fur running down its eyes, towards the muzzle.

Sexual dimorphism is not very pronounced in the cheetah: a male cheetah weighs 29–65 kg, and is 172–224 cm nose-to-tail with a shoulder height of 74–94 cm; a female cheetah weighs 21–63 kg, and is 170–236 cm nose-to-tail with a height of 67–84 cm (Hunter & Hamman, 2003:141). Although the cheetah is a relatively large carnivore, there are no records of a wild cheetah ever killing a human being (Hunter & Hamman, 2003:17).

Purring

The term 'purring' has been used liberally in the mammal vocalization literature, and an exhaustive review is given in Peters (2002). Using a definition of purring that continuous sound production must alternate between pulmonic egressive and ingressive airstream (and usually go on for minutes), Peters (2002) reached the conclusion that only "purring cats" (Felidae) and two species of genets (Viverridae *sensu stricto*), *Genetta tigrina*, and likely also *Genetta genetta*, had been documented to purr. For further discussion see Eklund, Peters and Duthie (2010).

Data collection and processing

Data were collected at the Dell Cheetah Centre, in Parys, South Africa, on 30 December 2011. Recordings were made of the two males Finley (F) and Mufasa (M), and from the two sisters Tippi (T) and Jade (J) – daughters of the previously studied male cheetah Caine (see Eklund, Peters & Duthie, 2010).

All five cheetahs were recorded in their enclosures and purring was elicited by Pieter Kemp or the first author. While F and M were quietly resting, the two sisters were playful and agitated, which occasionally led to a number of

passages with very short exhalation–inhalation sequences. Finally, film clips from December 2009 of the juvenile (male) cheetah Parker (P) were analysed. Parker, too, was also active during the recording, rather than resting calmly on the ground. Biographical information about all cheetahs are found in *Table 1*.

Equipment

The equipment used was a Canon HG-10 HD camcorder with a clip-on DM50 electret stereo condenser shotgun microphone with a frequency range of 150–15,000 Hz, and a sensitivity of –40 dB. The position of the microphone varied, partly due to the cheetahs moving, but was mostly directed towards the muzzle of the cheetahs, where the sound emanates (see e.g., Eklund, Peters & Duthie, 2010). Photos from the data collection are given in *Plate 1* and *Plate 2*.

Data post-processing

Audio tracks (44.1 kHz, 16 bit, mono) were excerpted with TMPGEnc 4.0 Xpress.

Analysis tools

The sound files were analyzed with Cool Edit 2000 and Cool Edit Pro 2.1. Cycles per phase were counted manually from the waveform. Statistics were calculated with SPSS 12.0.1.

Egressive–ingressive identification

For most of the data, egressive and ingressive phases were identified according to the method described in Eklund, Peters and Duthie (2010), i.e. with the first author keeping his hand on the side of the chest of the cheetah to monitor breathing, while uttering the words "in" and "out" in synchronization with the cheetah's breathing/purring. When this method was not possible (during playful stretches), the phases were identified by the first author from a combination of visual inspection of the waveform and sound characteristics – based on the distinctive sound differences between egressive and ingressive purring. Finally, in difficult cases, the original video was consulted for visual confirmation where it could be seen whether or not the ribcage of the cheetah was expanding or collapsing.

Parts of the recordings where analysis was difficult or not possible for some reason, e.g. talkover, bird chirps or other background noise, were discarded.

Results

The results are presented in *Table 1*.

Amplitude

Previous studies have reported louder egressive phases in both cheetahs and domestic cats (Eklund, Peters & Duthie, 2010; Schötz & Eklund, 2011; Eklund et al., 2012). This general pattern was observed in this study, as shown in *Figure 1*, where a long stretch of purring (from M) shows egressive–ingressive phases with clearly stronger egressive phases.

However, while Eklund et al. (2012) reported louder egressive phases in all four cheetahs analysed, one cheetah (Aisha) did produce a number of louder ingressive phases, which they attributed to the fact that she was in an agitated state. This was confirmed in the present study, and a less consistent and varying pattern was observed in P, T and J, all of whom were moving about, playing or licking the author's hand. Often there were no discernible amplitude differences between egressive and ingressive phases, and in some cases ingressive phases were clearly louder than egressive phases, as is shown in *Figure 2*.

Phase durations

Previous studies have not reported any consistent pattern in phase durations. Eklund, Peters and Duthie (2010) observed longer egressive phases while Schötz and Eklund (2011) reported longer ingressive phases. Eklund et al. (2012) reported longer egressive phases in two of the cheetahs studied, and longer ingressive phases in the two other cheetahs studied. This inconsistent pattern was also observed in the present study. Three of the cheetahs (P, T, J) exhibited no significant differences between egressive and ingressive phases, while M had significantly longer egressive phases and F significantly longer ingressive phases. It should be noted, however, that the F data set is very small ($N = 12$), and that far-reaching conclusions based on this data set should be subject to extreme caution. On average, shorter phase durations were observed in the three young cheetahs. This can likely partly be explained by the fact that they were in an agitated state and moving about rather than resting peacefully. However, all young cheetahs were capable of long purrs – although no phases longer than 3 s were observed.

Plate 1. Pieter Kemp and first author recording Mufasa. Photo by Miriam Oldenburg.

Plate 2. First author recording Jade and Tippi. Photo by Miriam Oldenburg.

Table 1. Summary Table. For all five cheetahs results are given for duration, cycles per phase and fundamental frequency. Results are presented independently for egressive and ingressive phases and for the two combined, and statistical tests are performed on differences between egressive and ingressive phonation.

	Mufasa (M)		**Finley (M)**		**Parker (M)**		**Tippi (F)**		**Jade (F)**	
Age	7 years		6 years		11 months		7 months		7 months	
Weight (kilos)	> 70		50		25		18–20		18–20	
Phonation type	Ingr	Egr	Ingr	Egr	Ingr	Egr	Ingr	Egr	Ingr	Egr
No. phases analysed	38	38	6	6	21	21	42	43	24	25
Mean duration (ms)	2174	2438	2763	1662	1003	970	1045	1063	685	590
Mean duration egr+ingr (ms)	2306		2212		986		1054		637	
Standard deviation	385.5	534.5	398.9	268.7	413.6	406.6	406.4	359.9	376.1	243.3
Maximal duration	3300	3640	3270	2100	1700	1710	2000	1790	2100	2100
Minimal duration	1280	1200	2360	1360	100	280	300	460	300	160
Δ *t* test (paired-samples, two-tailed)	$p = 0.014$		$p = 0.004$		$p = 0.074$		$p = 0.807$		$p = 0.168$	
Δ Wilcoxon (two related samples)	$p = 0.018$		$p = 0.028$		$p = 0.068$		$p = 0.726$		$p = 0.094$	
Mean no. cycles/phase	49.3	49.1	63.7	37.7	20.3	21.5	25.3	28.0	19.3	18.4
Mean no. cycles/phase egr+ingr	49.2		50.7		20.9		26.7		18.9	
Standard deviation	10.5	12.1	6.7	3.5	6.7	9.1	10.5	10.6	11.8	7.4
Maximal no. phases/cycle	69	77	75	43	35	38	50	50	67	34
Minimal no. cycles/phase	24	23	57	35	7	3	7	12	10	8
Δ *t* test (paired-samples, two-tailed)	$p = 0.921$		$p = 0.001$		$p = 0.562$		$p = 0.182$		$p = 0.576$	
Δ Wilcoxon (two related samples)	$p = 0.959$		$p = 0.028$		$p = 0.456$		$p = 0.268$		$p = 0.471$	
Mean fundamental frequency (Hz)	22.6	20.1	23.2	22.2	19.6	22.7	24.2	26.1	28.3	30.8
Mean frequency egr+ingr (Hz)	21.3		22.8		21.1		25.2		29.6	
Standard deviation	2.25	2.25	1.9	2.14	2.38	1.58	3.1	3.1	4.45	7.26
Highest fundamental frequency	25.7	21.9	24.9	26.5	23.4	25.7	32.9	35.0	37.5	49.0
Lowest fundamental frequency	18.7	11.2	20.5	19.7	16.4	20.0	20.0	21.5	22.5	24.1
Δ *t* test (paired-samples, two-tailed)	$p < 0.001$		$p = 0.048$		$p < 0.001$		$p = 0.002$		$p = 0.072$	
Δ Wilcoxon (two related samples)	$p < 0.001$		$p = 0.053$		$p < 0.001$		$p < 0.001$		$p = 0.113$	

Cycles per phase

Once again, this study repeats previous studies (ibid.) that have failed to observe any consistent differences as regards numbers of cycles per phase. This is perhaps not very surprising, given that cycles per phase can be expected to be closely linked to phase durations in general. While there is a strongly significant difference observed in F, this is likely attributable to the very limited data set, and no strong conclusions should be drawn based on this observation.

Fundamental frequency

Previous studies on the cheetah have in general reported lower fundamental frequency (F_0) patterns in ingressive phases (Eklund et al., 2012; Frazer Sissom, Rice & Peters, 1991). In the present study, two cheetahs (P, T) had significantly lower F_0 in ingressive phases, while M showed the opposite pattern. F and J exhibited no strong tendency in either direction. The observation that one cheetah (M) has significantly lower F_0 in egressive phases shows

that this parameter, too, is subject to individual variation. While some "high" frequencies were observed in the playing cheetahs, P produced some very low F_0 values showing that body size seemingly does not play an important role, shown in studies on the domestic cat (ibid.).

Discussion and conclusions

The present study extends on previous studies on purring in the cheetah by including subadult cheetahs in the data studied. Young cheetahs seemingly exhibit the same characteristics as do adult cheetahs, with the possible exception of very long phase durations, and no phases with a duration of 3 seconds or more were observed.

While individual variation is observed in both the present study and previous reports, actual value ranges are basically the same, with very low F_0 produced by all animals, regardless of body size and/or age. However, there is still a tendency for ingressive phases to have lower F_0, even if this study has found an exception to this general trend.

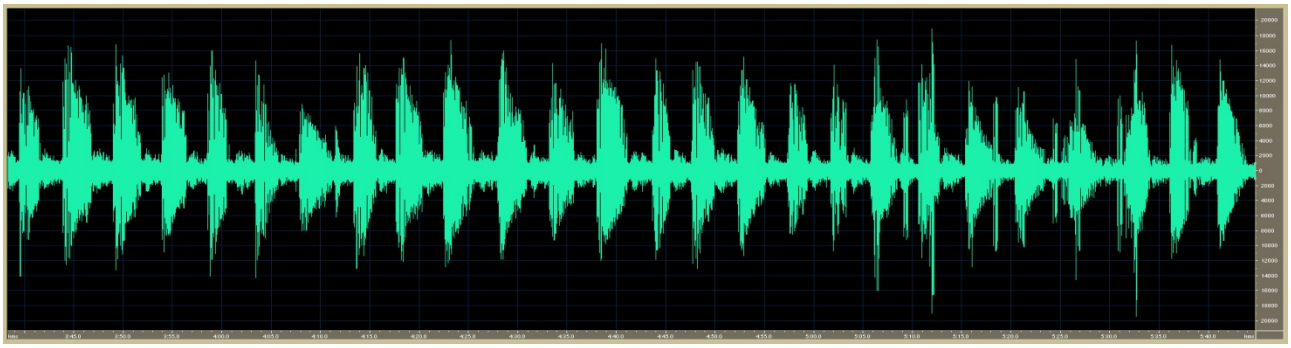


Figure 1. Mufasa purring amplitude pattern. Egressive phases clearly exhibit consistently higher amplitude. Window duration = 2 minutes 34 seconds.

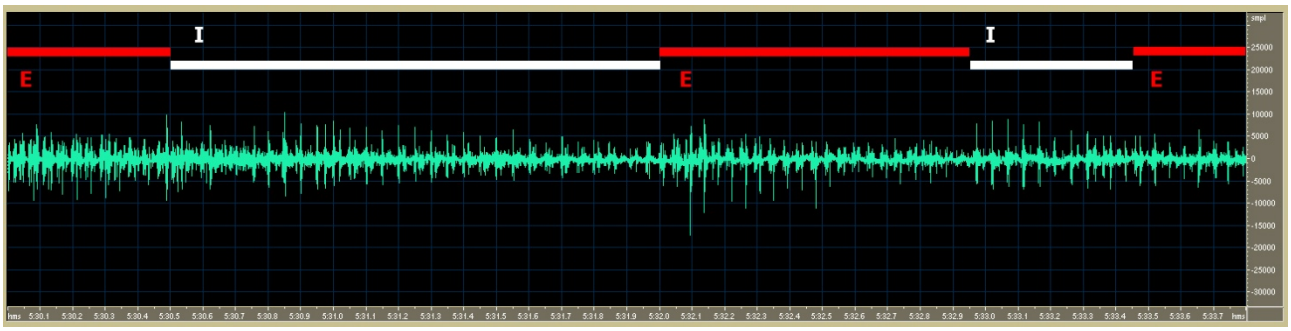


Figure 2. Tippi purring durations. Egressive phases (E) red/upper bar. Ingressive phases (I) white/lower bar. Durations: E: 480 ms – I: 1515 ms – E: 960 ms – I:480 ms – E: 340 ms.

Notes

Sound files/film clips found at http://purring.org

Acknowledgements

Deepest thanks to Pieter and Estelle Kemp at the Dell Cheetah Centre in Parys, South Africa (http://www.dccafrica.co.za) for making the data available and for outstanding hospitality. Thanks to Lizzie Duthie, Leante Liebenberg and Lu Mari Groenewald for help with the recordings. The first author also thanks Miriam Oldenburg for volunteering companionship over the years and for still photography.

References

Driscoll, C. A., J. Clutton-Brock, A. C. Kitchener & S. J. O'Brien. 2009. The taming of the domestic cat. *Scientific American*, June 2009, 68–75.

Eklund, R., G. Peters, R. Weise & S. Munro. 2012. A comparative acoustic analysis of purring in four cheetahs. In: *Proceedings from FONETIK 2012*, Gothenburg, May 30–June 1, 2012, 41–44.

Eklund, R., G. Peters & E. D. Duthie. 2010. An acoustic analysis of purring in the cheetah (*Acinonyx jubatus*) and in the domestic cat (*Felis catus*). In: *Proceedings of Fonetik 2010*, Lund University, 17–22.

Frazer Sissom, D. E., D. A. Rice & G. Peters. 1991. How cats purr. *Journal of Zoology* 223:67–78.

Hunter, L. & C. Hamman. 2003. *Cheetah*. Cape Town, South Africa: Struik Publishers.

O'Brien, S. J. & W. E. Johnson. 2007. The Evolution of Cats. *Scientific American*, July 2007, 68–75.

Peters, G. 2002. Purring and similar vocalizations in mammals. *Mammal Review* 32(4):245–271.

Schötz, S. & R. Eklund. 2011. A comparative acoustic analysis of purring in four cats. In: *Quarterly Progress and Status Report TMH-QPSR, Volume 51, 2011. Proceedings from Fonetik 2011*, 9–12.

Sunquist, M. & F. Sunquist. 2002. *Wild Cats of the World*. Chicago: University of Chicago Press.

Vocal development in two young cochlear implant users: Preliminary results

Anne M. Frank[1] *& Wim A. van Dommelen* [2]
[1]*Statped sørøst, Department of Hearing Impairment, Oslo, Norway*
[2]*Department of Language and Communication Studies, NTNU, Trondheim, Norway*

Abstract

This study investigates prelinguistic vocal development in two young cochlear implant users during their first eight months of hearing. Audio recordings of two girls' babbling were made and analyzed both auditorily and instrumentally. Results show for both changes in production from simple to more complex syllable structures. For both girls, the pattern of preferred places of consonant articulation changes over time and the duration of CV-dyads decreases. The data allow the conclusion that their developmental patterns are similar to those found for normal-hearing infants.

Introduction

During the first 2 years of life infants' prelinguistic vocalizations gradually become more complex and speech-like. Progress in vocal development is considered crucial for the acquisition of a phonological system: from producing early isolated consonantal and vocalic elements the infant must learn to combine vowels and consonants into adult-like syllables that can function as the phonetic building blocks of words (Oller, 2000). Childhood hearing loss can have a severe impact on early vocal development. This includes late onset of adult-like syllable production and smaller consonant and vowel inventories (Oller, 2000). It is, however, shown that prelingually deaf children who receive cochlear implants (CI) at an early age make significant progress in vocal development during the first year of device use. Main findings in a study made by Ertmer, Young and Nathani (2007) showed that milestones in vocal development were achieved within fewer months of hearing experience than observed by normal-hearing infants. Wie (2010) found that children receiving CIs between 5 and 18 months had expressive and receptive language skills within normal range. However, a large variation in rate of development has been identified (Wie, 2010; Ertmer et al., 2002) and also atypically developmental patterns in young CI recipients have been found (Ertmer, Young & Nathani, 2007). These findings indicate that some children remain at various levels of development longer than normal-hearing children and show limited gains in phonetic and syllabic inventories. This variability in performance across children may be due to a combination of several factors, including length of auditory deprivation and age in which a child received CIs.

Additional research is needed to gain a better understanding of factors that influence postimplantation vocal development. The present paper describes prelinguistic vocal development in two young CI users who received bilateral cochlear implants at 5.5 and 7 months of age, respectively. The intention was to explore to what extent these two children show developmental patterns comparable to those of normal-hearing infants.

Method

Subjects

Subjects involved in this study were two girls (G1, G2) both identified at birth with profound hearing loss. G1 received bilateral cochlear implants (CI) at the age of 5.5 months and G2 received bilateral cochlear implants at the age of 7 months. The CI devices in both children were activated within 1 month post-surgery.

Recordings

Both audio and video recordings were made using a Roland HR-9 and a Canon Legria HF G10, respectively, in the children's home environments by their respective parents. They were instructed to interact with the children in their usual manner.

In total, five recordings were made of each of the two subjects at intervals of 2 months (±3 weeks). Each recording comprised 30–40 minutes, in most cases resulting from several recording sessions spread out over a week. G1 got her CIs activated at the age of 6.5 months and a few days later the first recording was made. The first recording of G2 was made at the age of seven months, 1 week before she received her implants.

Analysis

For this investigation only the audio material was evaluated. Audio recordings were stored as wav files with a sampling frequency of 44.1 kHz and 16-bit quantization. Acoustic analysis was performed using Praat (Boersma & Weenink, 2012) and involved auditory evaluation combined with visual inspection of waveform and spectrogram.

Material used for analysis comprised 50 utterances selected from each of the five recordings. For the present purposes, an *utterance* is defined as a single vocalization or a group of vocalizations separated by articulatory pauses.

First step in analysis was the division of an utterance into syllable-like entities. A *syllable* was defined as a vocalic or (in some cases) consonantal nucleus, or could comprise a combination of consonant and vowel.

Consonants were classified according to the usual criteria of place and manner of articulation, and voicing. Due to reasons of space, this paper will deal only with place of articulation. Since the present babbling material precludes conventional fine-grained division of the place dimension, four broader categories were established:

1) bilabial
2) dental/alveolar/palatal (*front*)
3) velar/uvular (*back*)
4) glottal

In addition, vowel quality was analyzed following the usual criteria of degree of opening, backness and lip rounding. Results will be presented in future publications.

Results

Syllable structure

Table 1 displays the percentage of each syllable type produced by G1 and G2 during their first eight months of implant use. Both girls show similar developmental patterns of increasing syllable shape complexity. G1s dominant form of syllables shifted from simple V syllables (59%) in recording 1 to CV syllables (69%) in recording 4. Simple V and C syllables (39% and 27% respectively) dominated G2 production in the first recording, whereas CV structures accounted for 77% of the syllables in recording 5.

Consonants

Place of articulation

Table 2 presents an overview of the places of articulation used in the girls' production of consonantal elements in the five recordings. In spite of sometimes substantial differences between the two girls, similar developments can be observed. For both, the most frequently used place of articulation in the first recording (pre-implant for G2) is bilabial (42% and 63%, respectively). Also, for both of them the use of this place shows a decreasing trend, percentages for recording 5 being 20% and 25%, respectively. Further, a relatively consistent increase in the use of the front region emerged from the data: from 6% in recording 1 to 38% in recording 5 for G1; correspondingly from 2% to 40% for G2. Also, the amount of back (velar/uvular) articulations increases over time, although especially for G1 less clear than the increased use of the front region (G1: 27%–36%; G2: 7%–15%). Finally, both girls show a tendency of reduced use of glottal productions. The trend is clear for G1 (reducing from 24% to 7%) and less consistent for G2 (from 28% to 20%).

CV duration

In this section we shall investigate temporal aspects of the girls' vocalizations. To that aim we selected syllable-like CV units from the recordings as a measure of development towards the production of syllables in adult speech. Recall that CV-dyads represented the most frequent type of vocalization for both girls (around 50% of all productions pooled across recordings 1–5; *Table 1*).

Mean CV durations presented in *Figure 1* show similar trends for G1 and G2. The general trend goes from longer durations at the time of the first recording (mean 551 ms) to shorter durations in the last recording (mean 379 ms). The only striking deviation from this pattern is the mean value of 1097 ms measured in recording 2 productions, corresponding to a hearing age of 2 and 3 months, respectively, for G1 and G2. Although G2 has only one CV unit among the analyzed recording 2 productions (corresponding to 1%; *Table 1*), it has a typical duration (1262 ms).

Table 1. Syllable structure in G1's and G2's vocalizations. V=single vowel; C=single consonant; CV=syllable-like CV-dyad. Numbers represent occurrences in % in recordings (Rec) 1–5. mo=months.

		Rec 1	Rec 2	Rec 3	Rec 4	Rec 5	Rec 1-5
	hearing age	1 week	2 mo	4 mo	6 mo	8 mo	
G1	**V**	59	47	31	17	40	36
	C	9	9	2	2	4	5
	CV	30	31	64	69	47	49
	Other	2	13	4	12	9	7
	hearing age	w/o CI	3 mo	4 mo	7 mo	8 mo	
G2	**V**	39	0	18	13	11	16
	C	27	75	10	1	2	17
	CV	21	1	63	57	77	52
	Other	13	24	9	29	11	16

Table 2. Place of articulation in G1's and G2's consonantal productions. Numbers represent occurrences in % in recordings (Rec) 1–5. mo=months; front=dental/alveolar/palatal; back=velar/uvular.

		Rec 1	Rec 2	Rec 3	Rec 4	Rec 5	Rec 1-5
	hearing age	1 week	2 mo	4 mo	6 mo	8 mo	
G1	**bilabial**	42	39	64	36	20	40
	front	6	9	3	38	38	23
	back	27	18	5	26	36	23
	glottal	24	33	27	1	7	14
	hearing age	w/o CI	3 mo	4 mo	7 mo	8 mo	
G2	**bilabial**	63	90	33	11	25	38
	front	2	0	23	48	40	28
	back	7	3	4	33	15	14
	glottal	28	8	40	8	20	20

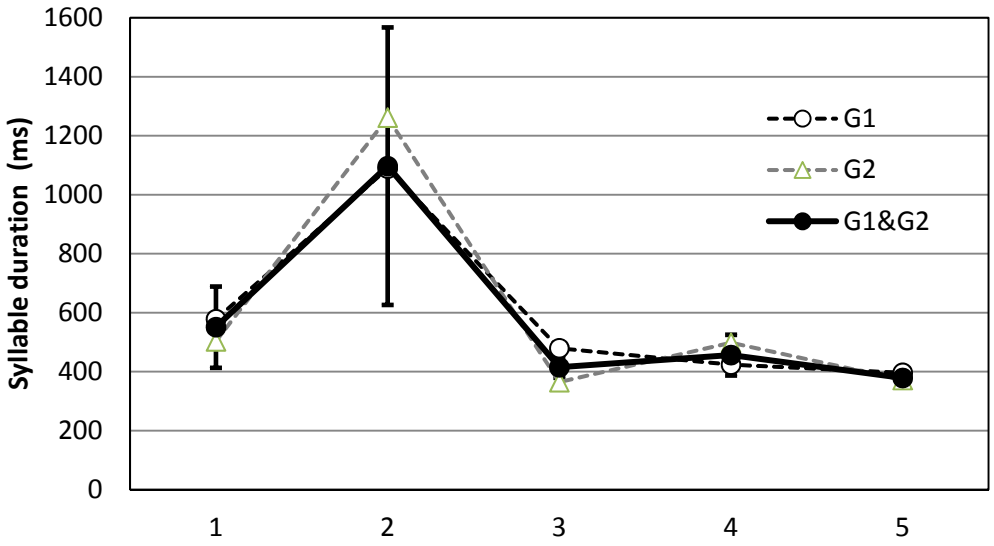


Figure 1. CV durations in ms for G1 and G2 separately and pooled for recordings 1–5. Vertical bars indicate 95% confidence intervals for pooled durations.

The internal structure of this token is a long bilabial nasal (1066 ms) followed by a much shorter vowel (196 ms). In contrast, mean duration of the consonantal element in G1's CV tokens is only 192 ms, while the vowel is relatively long (mean of 879 ms).

To study variation of individual CV durations, in *Figure 1* the 95% confidence interval was chosen as a measure. The results show that the pattern found for mean CV durations in recordings 1–5, i.e. a general downward trend with the exception of recording 2, is also present in the variation of durations. Generally, the width of the confidence interval is relatively small, decreasing from 276 ms for recording 1 to 58 ms for recording 5. The corresponding width calculated for recording 2 is substantially larger (941 ms). This value reflects the wide range of CV durations for this condition, varying between 237 ms and 3874 ms.

Discussion

The most important conclusion that can be drawn from the data collected thus far is that the girls' prelinguistic development is similar to patterns observed for normal-hearing infants. During their first eight months of hearing both girls made substantial progress. Vocalizations developed from predominantly simple V and C nuclei to CV syllables as the most frequent shape. Further, a typical trait is the increasing use of the front region of the mouth as place of articulation. As to temporal organization, CV durations appeared to become gradually shorter. This whole picture emerging from the data is in line with findings reported by Oller (2000) and Frank (2009) for normal-hearing children.

Generally, the patterns found for the present infants were rather similar, in particular with regard to CV syllable durations. It should be noted, however, that we are dealing with two case studies here and that similarities are in all probability accidental. Syllable duration in eight normal-hearing infants in Frank (2009) gradually decreased from the age of 6 months to 24 months. Considerable between-subject variation was found, however, especially at ages of 6, 8, and even 20 months.

Apart from between-subject variation the data revealed also considerable within-subject variation. For many categories, tendencies of increase or decrease were not monotonic but characterized by larger up-and-down changes. For example, for G1 the clear predominance of CV shaped syllables at the hearing age of six months was diminished two months later. Another example is the most frequent use of bilabial place of articulation, which was not found at the time of the first recordings but in the third recording of G1 and the second recording of G2. Also for this kind of within-subject variation it can be noted that it is not atypical but also found in children without hearing impairment.

Conclusion

The conclusion seems justified that young cochlear implant recipients can achieve prelinguistic vocal development comparable to that of normal-hearing infants.

Acknowledgements

We would like to thank the parents and the children involved in this study for their pleasant and smooth cooperation.

References

Boersma, P. & D. Weenink 2012. Praat: doing phonetics by computer ([Computer program]. Version 5.3.15, retrieved 10 May 2012 from http://www.praat.org/.

Ertmer, D. J., N. M. Young, K. Grohne, J. A. Mellon, C. Johnson, K. Corbett & K. Saindon 2002. Vocal development in young children with cochlear implants: Profiles and implications for intervention. *Language, Speech & Hearing Services in Schools* 33:184–195.

Ertmer, D. J., N. M. Young & S. Nathani 2007. Profiles of vocal development in young cochlear implant recipients. *Journal of Speech, Language, and Hearing Research* 50:393–407.

Frank, A.M. 2009. *Strukturelle og temporale trekk i norske spedbarns lydutvikling.* PhD thesis NTNU, Trondheim.

Oller, D.K. 2000. *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates Inc., Publishers

Wie, O. B. 2010. Language development in children after receiving bilateral cochlear implants between 5 and 18 months. *International Journal of Pediatric Otorhinolaryngology* 74:1258–1266.

Possible explanation of Chinese misidentified tones

Guohua Hu
Department of Languages and Literatures, Gothenburg University, Gothenburg, Sweden

Abstract

Even though F_0 plays an essential role in tone perception other facts (e.g. different consonant types, vowel quality, see Hombert, 1978) also effect tone perception. This article tries to find some evidence of influence between initial consonants and misidentified tone patterns in terms of the acquisition of Chinese tones. The main results show that voiced sounds, e.g. [l], effect that (T)one 2 (high) is misinterpreted as T3 (low), and that aspirated stops, e.g. [p^h], cause that T3 (low) is misidentified as T2 (high).

Introduction

Stops are universal phenomena. Different languages use various stop systems. Chinese, on one side, has only a voiceless stop system; there the features [±aspirated] play an essential role and voiced sounds include approximants e.g. [l], [w] and the nasals (Lin, 2007:45–65).

On the other side, [±voiced] are distinguishing features in Swedish, see also (Hu, 2012). Apart from the segments, suprasegments in languages give another view. This article only focuses on how Swedes have perceived Chinese tones so a presentation of Swedish suprasegments is irrelevant here.

Chinese is a tone language, which has four lexical tones (T1, T2, T3, and T4, see Figure 1) and one neutral tone (or toneless), e.g. 吗 ma *question mark* (eh?). A classical example often given in literatures is 妈 mā *mother*, 麻 má *hemp*, 马 mǎ *horse*, and 骂 mà *scold*. The words distinguish their meanings by different tones.

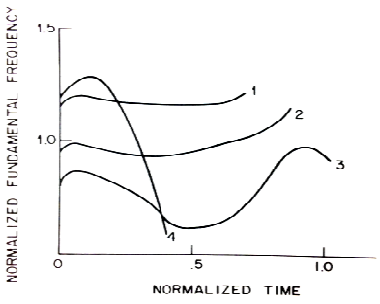


Figure 1: F_0 contours with four Chinese tones (from: Chuang et al., 1972)

Figure 1 shows the F_0 (fundamental frequency) contours of the four Chinese tones. In short, T1 is high and relatively level over most of the duration of the tone. T2 begins with a relative lower pitch compared to T1, and the onset of the rise occurs in the middle portion and ends almost as high as T1. T3 contour displays the lowest region of the F_0 range, although extending at least to the midpoint of the range by the offset. F_0 in the beginning of both T2 and T3 are quiet close to each other. T4 starts high and robustly falls to the bottom during its duration. Phonologically, T1 is high level, T2 is high rising, T3 is dip low and T4 is low falling.

Earlier studies (e.g. Klatt, 1973; Zee, 1985) have paid their attention to the pitch of the vowel (F_0). Many scholars (e.g. Shen & Lin, 1991; Chuang et al., 1972; Kiriloff, 1969) have conducted perception tests outgoing from different native speakers since different languages vary in their pitch patterns and functions. Their main results have shown that T2 is often misidentified as T3 and vice versa, but T3 is not so frequently misidentified as a T2.

The explanation for a misconception of this tone pair is that "neither the falling and rising contour nor the position of the dip point in tone-3 alone can be the perceptual cues to discriminate tone-3 from tone-2" (Chuang et al., 1972:299), see also *Figure 1*.

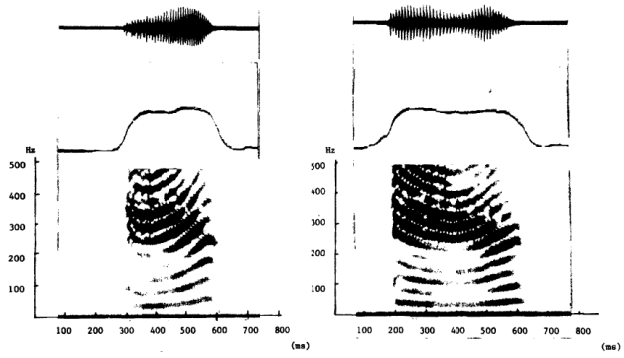


Figure 2: Turning points between T2 and T3 (from: Shen & Lin, 1991).

Shen and Lin (1991) have reported that the time of the turning points is the perceptual cue that differentiates T2 from T3. Moreover, a shift in vowel duration has been shown to affect the perception of T2 and T3, see *Figure 2*.

Significant is that they have chosen test words only with voiced initial consonants e.g. lóu *a storied building* and lǒu *basket*, máo *hair* and mǎo *riveting,* and ní *mud* and nǐ *you*, etc. They have, however, not explained why they used test words only with a voiced initial, so the question is if this is enough for stating that the T2 is confused T3. We lack an explanation for the misperception between T3 and T2.

In a study of tone perception of adult Swedes Hu and Lindh (2010) have shown the similar result, i.e. that T2 and T3 and vice versa are most frequently confused. However, they haven't illustrated *why* this tone conception occurs.

The hypothesis of this article is that the initial consonants affect the tone perception since the feedback from the informants, after the experiment, has revealed that they experienced that the consonants preceding the vowels disturbed the tone perception.

Furthermore, Hombert (1978) has claimed that F_0 perturbation of initial stops does disturb the tone perception. However, he has only analyzed the acoustic features between English and French stops. There was no evidence how English speakers perceived French stops and vice versa.

This article pays its attention to the tone perception based on Hombert. Firstly the perception experiment will be presented. Secondly, the results are illustrated. Next, the initial stops and tone confusions patterns are discussed.

The current study

Twenty-five different unhandled disyllabic words were selected from a textbook for the beginners of Chinese at Gothenburg University. No syllable has a nasal initial. The test words are disyllabic ($\sigma_1\sigma_2$) including 15 (of the possible 19) tone combinations. The test word 劳动 láodòng [^{35}lau^{51}tɔŋ] 'work', to take an example, consists of two syllables. The initial consonant of σ_1 is a lateral approximant (voiced), the initial consonant of σ_2 one is an unaspirated stop. The tone combination of this word is T2+T4.

One male and one female native speaker of Chinese pronounced the words in isolation. Each speaker repeated the words twice with a pause of 1 second in between and there is 2 seconds pause before the following new word. The audio was presented in high quality headphones in the student language lab of the university.

Listening tests have been conducted every year between 2007 and 2013. Each year there have been 25–40 participants (bilingual Chinese immigrants are excluded from the results). The actual group consisted of 18 students who were admitted in the autumn semester of 2007. The subjects had the possibility to listen to the words as many times as they needed for noting correct transcription and tones. The listening test was also an exam, which pushed the participants to perform well.

The students almost correctly transcribed the sounds by using Romanization (Pinyin system). The wrong spellings are so few that they are ignored in the present analysis that focuses only on tone identifications.

Results

The 18 participants had to identify 25 disyllabic words. The total responses were 900 (= 18 × 25 × 2). 278 tones were misidentified (30.88%). The matrix of tone responses is shown in *Table 1*, where **bold** indicates the correct answer.

Table 1: Matrix of tone responses (%)

	Response																			
	Type I				*Type II*					*Type III*										
	σ_1				Σ_2					σ_1				σ_2						
Stimulus (T)	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*0*	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*0*		
1	**75**	13	3	9	**82**	7	0	7	4	**75**	11	8	6	**82**	9	0	9	0		
2	16	**60**	11	13	2	**57**	33	5	3	5	**60**	24	11	0	**58**	33	9	0		
3	10	33	**52**	5	1	23	**69**	5	2	13	13	**54**	20	7	19	**69**	5	0		
4	3	17	1	**79**	3	10	3	**78**	6	5	11	5	**79**	7	9	3	**78**	3		
0					0	0	0	47	**53**					0	0	0	47	**53**		

Depending upon where the misidentified tones of the syllables in the words occur they were divided as follows (note that T_0 never can be distributed in σ_1). The highest percentages of misidentifications are:

- Type I, σ_1: T3→T2 (33%)
- Type II, σ_2: T2→T3 (33%)
- Type III: the tones of both σ_1 (T2→T3, 24%) and σ_2 (T2→T3, 33%) are wrongly perceived

Table 2 below shows the influence of the initial consonants on the incorrect tone responses.

Table 1: Correlation between misidentified tones and the initial consonants of three types (%)

	Sounds	Features	Type I	Type II	Type III	
			σ_1	σ_2	σ_1	σ_2
Voiceless	Stop	Unaspirated	12	17	8	29
		Aspirated	15	13	16	12
	Fricative		31	13	35	22
	Affricate	Unaspirated	6	6	6	14
		Aspirated	7	11	4	2
Voiced	Central Approximant		4	33	4	20
	Lateral Approximant		24	7	27	0
	∑		100	100	100	100

- Type I: The fricatives can be suspected to play a role for tone confusions (31%). Thereafter follow the stops (unaspirated 12% and aspirated 15%). The stops have a stronger connection with misidentified tones than the affricates. Also the lateral seems to have an impact on the tone perception (24%).
- Type II: One third of the central approximant affricates are seen together with tone confusions. The stops have a similar tendency as in Type I. However, fricatives decrease to 13% (31% in Type I). Aspirated affricates tie up stronger (11%) with tone misperception than the unaspirated (6%).
- Type III: If both syllables have a fricative as their initial consonant it looks plausible that they influence the tone perception (35% and 22%). The initial stops of the σ_2 are possibly involved with the wrong tone identifications (29%). Unaspirated affricates (14%) have a substantial relation to misidentified tones. Concerning voiced initials, the results show that the central approximants have a strong association with the wrong tone responses of σ_2 (20%), whereas the lateral (27%) seems to have its impact on σ_1.

A plotting schedule, showing the patterns for tone confusions related to the initial consonants was developed. The space of this article is too small for reproducing it here, but it can be requested from the author. Since the data was very little, we pick only the results according to the analyses above. A suspicion was found that only some initial consonants influence the tone confusion patterns T2→T3 and T3→T2.

- Type I, σ_1: In 20% of the cases the aspirated initial stops seem to have a connection with the tone confusion T3→T2. The initial lateral approximant has a close link to T2→T3 (12%).
- Type II, σ_2: The aspirated initial affricates seem to be involved when the tone confusion T2→T3 occurs in 38% of the cases. Unaspirated stops might have a relationship to the wrong tone pattern T3→T2 (15%), it is however not as frequent as in the case of the aspirated affricates. In 25% a central approximant as initial ties up with this same tone confusion. The same kind of initial even shows a connection with T3→T2 (15%)

- Type III: When a lateral approximant is the initial of σ_1 there often occurs a T2→T3 confusion, 26%. Concerning a possible pattern for the T3→T2 confusion the initial consonants are so scattered that no strong connections can be suspected.

Conclusion

It may be concluded that the initial consonants do affect tone perception. Aspirated initials seem to be linked to cause that T3 is misinterpreted as a T2. Unaspirated stops and approximants show a connection of the wrong perceived tone pattern T2 as a T3. It might be added that this pattern has repeated itself in the tests of 2008–2013 (although the numbers in this article refer to the test 2007).

The department of Chinese has started a perception coaching system. It includes all Chinese disyllabic phonotactic and tonotaxical patterns in order to observe closely the relationship between initial consonants and misidentified tone patterns. Meanwhile, the acoustic data of F_0 disturbance by both Swedish and Chinese consonants preceding vowels should be collected.

References

Chuang, C. K., S. Hiki, T. Sone & T. & Nimura 1972. The Acoustical features and perceptual cues of the four tones of standard colloquial Chinese *Proceedings of the 7th International Congress of Acoustics,* 3297−300.

Hombert, J.-M. 1978. Consonant Types, Vowel Quality and Tone. In: V. A. Fromkin (ed.), *Tone − A Linguistic Survey*, 77–111.

Lin, Y.-H. 2007. The Sounds of Chinese. Cambridge: Cambridge University Press.

Hu, G. 2012. Chinese and Swedish Stops in Contrast. Paper presented at the *Proceedings of Fonetik 2012 Gothenburg Department of Philosophy, Linguistics and Theory of Science*, 77–80.

Hu, G. & J. Lindh. 2010. Perceptual mistakes of Chinese tones in 2-syllable words by Swedish listeners (Abstract). Paper presented at the *Proceedings of the Fourth European Conference on Tone and Intonation* (TIE4). Stockholm, 121.

Kiriloff, C. 1969. On the auditory perception of tones in Mandarin. *Phonetica* 20:63–67.

Klatt, D. 1973. Discrimination of fundamental frequency contours in synthetic speech duplicaitons for models of pitch perception *Journal of the Acoustical Society of America* 53, 8–16.

Shen, X. S. & M. Lin. 1991. A Perceptual Study of mandarin Tones 2 and 3. *Language and Speech* 34:145–156.

Zee, E. (1985). Sound change in syllable final nasal consonants in Chinese *Journal of Chinese Linguistics* 13(1):291–330.

Acoustic data on variation in the Swedish postalveolar sibilant across boundaries

Emanuel Karlsson
Department of Linguistics and Philology, Uppsala University

Abstract

The claim that postalveolarization across word boundaries in Swedish is a non-obligatory process is examined for the postalveolar sibilant in an acoustic study of Standard Swedish. An internal group exclusively showing a postalveolar and a boundary group with a range from a laminal dentoalveolar to a post-alveolar are distinguished. For the boundary group, data on fricative resonant frequency and on duration are analysed further. The presence of a distinct rhotic appears to be not exclusively correlated to fricative quality, and a rhotic may appear also with a postalveolar. It is suggested that a dentoalveolar following a rhotic may be typically apical rather than laminal unless specifically to mark the juncture. The data do not seem to support the claim that the choice of outcome is influenced by speech tempo.

Introduction

The Standard Swedish apical posterior coronals that contrast with plain dentoalveolars are perhaps best described as postalveolar (but often termed 'retroflex' or 'supradental'). Their diachronic source is sequences of rhotic plus plain dentoalveolar (consequently, they do not occur in word-initial position in isolated words). Postalveolars also appear across word and morpheme boundaries where a final lexical /r/ of one element meets with an initial dento-alveolar of the next. At word boundaries, post-alveolars interchange with sequences of rhotic plus dentoalveolar (cf. e.g. Garlén, 1988).

Among Swedish phonologists in general, postalveolars in all positions are viewed phonologically as the product of synchronic derivational rules of assimilation and deletion on underlying phonematic sequences of /r/ plus dentoalveolar (e.g. Elert, 1957; Eliasson, 1986; Riad, 2010). Norwegian phonologists, on the other hand, tend to treat the same phenomenon differently, with the corresponding internal postalveolars as phonematic for Norwegian (e.g. Kristoffersen, 2000). This is also the approach used in UPSID (Maddieson, 1984, following Vanvik, 1972).

The original motivation behind the sequence analysis seems to have been one of inventory economy and – either implicitly or explicitly – of the now widely rejected principle of *biuniqueness*, which rests on the theoretical presumption that sounds should map to uniquely predictable underliers (i.e., identical phones in a given context represent the same phoneme) and manifests itself as a desire for derivations to work both ways: postalveolars in every position, seeing as they pass for identical, must then all have the same underliers.

Representing all postalveolars as phono-logical sequences of rhotic plus dentoalveolar is complicated by the independent existence, root-internally, of phonetic sequences of rhotic plus dentoalveolar, e.g. *absurd* [apˈsʊɹd]. The sug-gested distribution that sequences occur for voiced dentoalveolars after short vowels may be historically relevant but is not synchronic, as seen in e.g. *imorn* [ɪˈmɔɳː], *Ursula* [ˈʊɹsʊla].

This is resolved (e.g. in Eliasson, 1986) by exploiting an abstract analysis of Swedish quantity, wherein for underlying forms of stressed short vowels followed by two con-sonant phonemes, an otherwise redundant geminate may be added at will to the notation: the postalveolarization rule, then, is taken not to apply with such a geminate /r/. Apart from this being the only case where such arbitrary gemination would be distinctive, it appears to be contradicted by cases like *perception* [pʰæɹsepˈxuːn] and *piercad* [ˈpʰeːɹsad].

Word-internal postalveolars evidently do not exhibit the same degree of variation as those that occur across word boundaries. The variable appearance at boundaries of a simple post-alveolar on the one hand and a sequence of rhotic and dentoalveolar on the other has been ascribed to the non-obligatory nature of the postalveorization process across a word boundary, referring to the application of the above rule. Some factors controlling the choice of sequence vs. postalveolar have been suggested (Witting, 1959; Gårding, 1967; Malmberg, 1968; Eliasson, 1986; Kuronen, 2003), including grammatical-lexical context, phonetic context, speech tempo, and style.

The study

This paper presents an empirical study of the acoustic quality of Swedish postalveolars with reference to morphosyntactic position. The study focuses on the sibilant postalveolar because of its acoustic and perceptual qualities (it is, presumably, the most perceptually salient postalveolar and it is also the one most amenable to acoustic analysis), as well as its high distributional frequency.

The acoustic quality of eligible sibilants is gauged by measuring the lowest resonance frequency of sibilant fricatives. While this is a simplification of spectral properties, it is taken to be indicative of perceived quality and indirectly of tongue position, with apical postalveolars perceived as low-pitched and laminal dentoalveolars as high-pitched.

An analysis of each sibilant token based on the correlation between morphosyntactic context and its lowest resonance frequency was performed. The analysis also considered the duration of the sibilant as well as the preceding rhotic (when present).

Method

The material consists of scripted and unscripted Standard Swedish from broadcasts of radio news programmes *Lunchekot* and *Studio Ett*, as well as unscripted material from the *Swedish Map Task Corpus* (Helgason, 2006). The radio data comprises 10 male and 11 female speakers, with 1 male and 2 female speakers in the SMTC data.

Eligible sibilants were identified on the basis of the occurrence of sequences of <r> and <s> in the orthographic transcription, though certain common cases with no trace of rhotic (especially the copula *är*) were excepted. The eligible cases were then classified into 13 morphosyntactic categories based on morphological and syntactic criteria (but as is shown below, these 13 categories can be collapsed into just two basic categories).

Sections of fricative noise corresponding to eligible sibilants, and the duration of any discrete appertaining rhotic, were labelled in *WaveSurfer* (Sjölander & Beskow, 2000). The frequency of the lowest resonance of the sibilant was measured using *Praat* (Boersma & Weenink, 2001).

Results

Figure 1 shows the frequency spread of the lowest resonance frequency for the tokens in each morphosyntactic category.

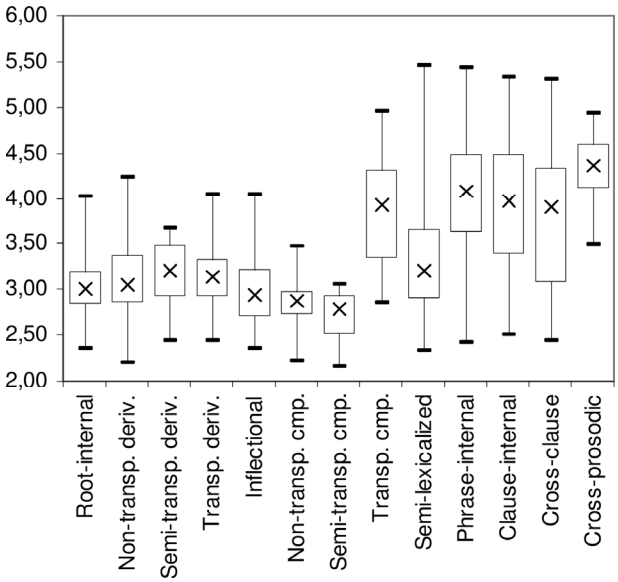


Figure 1. Spread of measured resonance frequencies within each morphosyntactic category.

Two fairly distinct groups of data emerge. Tokens in morpheme-internal position and across most morpheme boundaries have a lower frequency range than those across word boundaries. That is, the former (internal group) is constistently produced as [ʃ], while the latter (boundary group) ranges from [s] to [ʃ].

Token resonance frequency for the two groups is plotted against the total duration of sibilant plus any preceding rhotic in *Figure 2*.

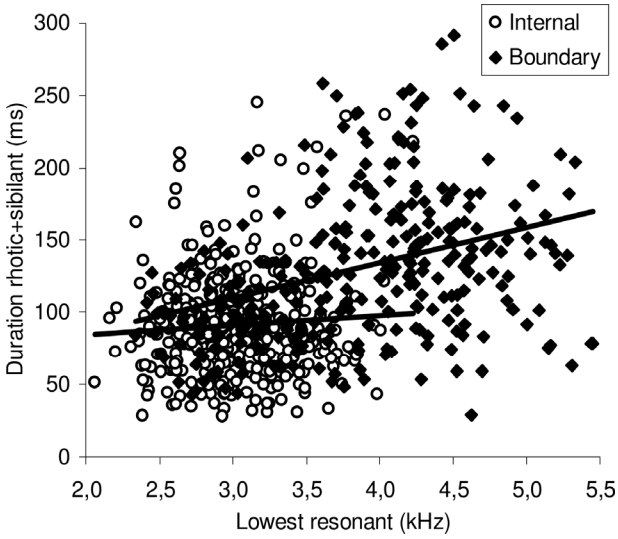


Figure 2. Fricative quality for each token, measured as the lowest resonance frequency (in kHz), plotted against the total duration of rhotic (when present) plus sibilant in internal and boundary groups.

Analysis

The durational data, supported by spectrographic evidence, indicate for the boundary group that a rhotic can also occur with a postalveolar, i.e. [ɹʃ]. The rhotic may also be completely absent before a laminal dentoalveolar sibilant. This is common with plural and present suffixes in *-r* and with frequent function words (e.g. *för*, *var*, *hur*).

Transparent (i.e. non-lexicalized) compounds pertain to the boundary group and inflections to the internal group, as seen in *Figure 1*. The low mean for semi-lexicalized phrases may suggest some orthographically multi-word units (e.g. *år sen*) in fact pertain to the internal group.

Data on sibilant duration does not show any convincing evidence of connection between duration and resonant frequency, which would have been suggestive of outcome choice influenced by speech tempo (i.e., that postalveolars are more frequent in faster speech).

Figure 3 shows the distribution of boundary group resonant frequencies for female speakers. A bimodal distribution can be observed, the lower frequency peak indicating postalveolars and the higher peak indicating dentoalveolars.

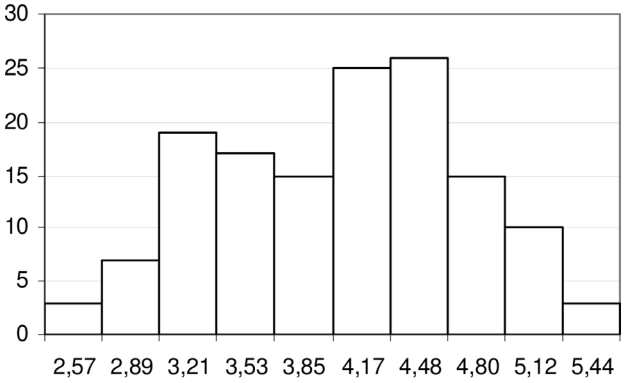


Figure 3. Distribution of sibilant frequencies (kHz) in the boundary group for female speakers.

Discussion

Considering frequency as a function of morphosyntactic condition, two major groups of data emerge, referred to here as the internal group and the boundary group. For the internal group, a postalveolar is obligatory and contrasts sequence of rhotic plus dentoalveolar. The boundary group has a more variable distribution with regard to lowest resonance frequency, as well as, semi-independently, the relative presence (or absence) of a rhotic preceding the sibilant, where the rhotic also varies both in duration and intensity.

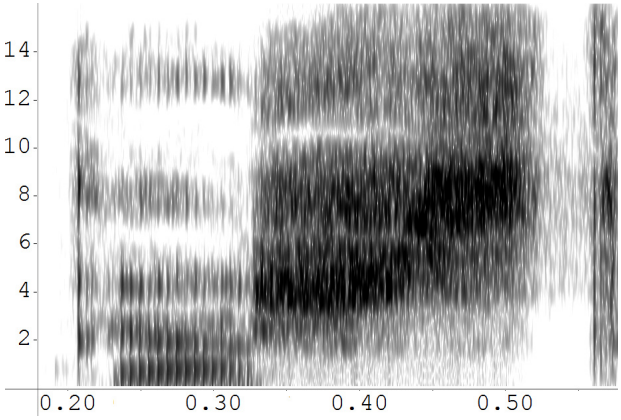


Figure 4. Spectrogram of [ɡaʃ̟st] *in the utterance* (da)gars st(ämma) *showing juncture marking.*

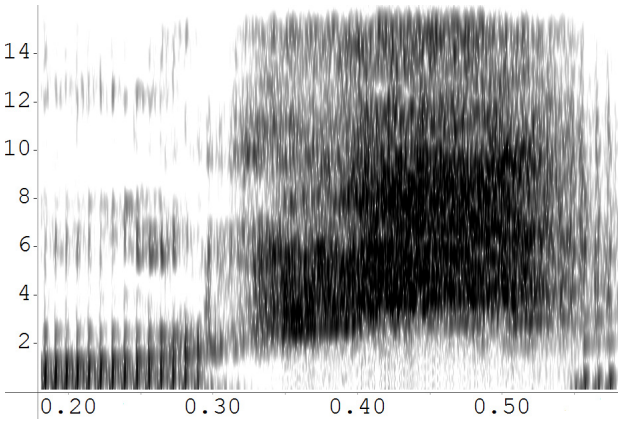


Figure 5. Spectrogram of [œɹʃ̟si] *in the utterance* (f)ör si(na) *showing probable repair.*

It may be hypothesized that, among dentoalveolars in the boundary group, an apical realization [s̪] is more common, as the sibilant is usually coarticulated with the rhotic, while a laminal [ʃ̟] is used as a cue for marking the juncture. *Figure 3* may hint that dentoalveolars more commonly exhibit lower frequencies, and thus are apical. This use of laminal is similar to one way of marking juncture as seen in *Figure 4*, where an internal postalveolar transitions into a laminal sibilant, [ʃ̟], across a boundary, as opposed to collapsing sibilants into a single postalveolar sound.

Several variations of this use of transition exist. A sibilant following a final, internal postalveolar, as opposed to a simple rhotic, is more likely to be postalveolar, suggesting that when it is not, this is a marker. *Figure 5* shows lexical /r#s/ as [ɹʃ̟], possibly as a kind of boundary repair, with a similar transition from postalveolar to apical dentoalveolar within the sibilant, while lexically it is past the boundary.

It appears that pragmatic features (as in juncture marking, speech tempo) may make significant use of and interact with the variation in a phonologically meaningful way. This may include influence on coarticulation in apical vs. laminal boundary dentoalveolars. The boundary postalveolar, however, patterns largely with the internal postalveolar, not indicating juncture.

The pragmatic aspect of the boundary group can be seen as a correlate of the H&H continuum (cf. Lindblom, 1990), reflected in the choice of postalveolar vs. dentoalveolar, where the latter might function as a juncture marker in its own right, seeing as hyperspeech possibly favours separation of gestures. (*Figure 3* shows dentoalveolars to be more common, but this could be attributed to style.) The choice of outcome is likely based on a complex of factors, where speech tempo could play a role in that faster and thus less enunciated speech implies hypospeech (i.e., postalveolar).

From the point of view of Articulatory Phonology (cf. Browman & Goldstein, 1992), the relationship between [ɹs] and [ʃ] can be described as different phasings of the same gestures, illustrated in *Figure 6*. Linked to this is how a postalveolar gesture may or may not span over a stretch of concurrent gestures.

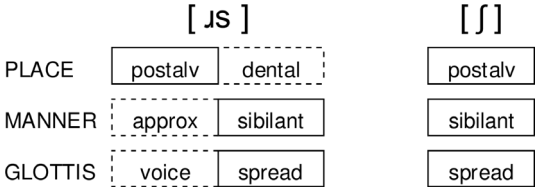


Figure 6. Schematic gestural scores of [ɹs] *and* [ʃ].

Variation can be expressed as different constraints on the range of allowed outcomes, which may be seen in terms of phasing and degree of activation (of the same component gestures), modification of both of which the boundary variation is contingent on (though it is unclear how they relate).

Boundary behaviour (as apparent occasional absence of the historical process) can be viewed as an overlay effect on an otherwise continuous progression of speech, arising out of analogical pressure from the different connected contextual forms. A realization [ɹʃ] could likewise stem from analogical influence of the lexical /r/. Affixes, on the other hand, remain unisolable, meaning in gestural terms that they are simple specifications of gestures, with the implication that the relation of forms can be described non-linearly.

Change in variational constraints could be understood to reflect postalveolar diachrony within an H&H model. If obligatory internal postalveolar was, historically, preceded by a stage of variation (as possibly still seen in the lateral), the development can be conceived in terms of drift of the variation spectrum.

References

Boersma, P. & D. Weenink. 2001. PRAAT, a system for doing phonetics by computer. *Glot International* 5:341–345.

Browman, C. & L. Goldstein. 1992. Articulatory Phonology: An Overview. *Phonetica* 49:155–180.

Elert, C.-C. 1957. Bidrag till en fonematisk beskrivning av svenska. *Arkiv för nordisk filologi* 72:35–60.

Eliasson, S. 1986. Sandhi in peninsular Scandinavian. In: A. Henning (ed.), *Sandhi phenomena in the languages of Europe*, Berlin: Walter de Gruyter, 271–300.

Garlén, C. 1988. *Svenskans fonologi*. Lund: Studentlitteratur.

Gårding, E. 1967. *Internal juncture in Swedish*. Lund: Gleerup.

Helgason, P. 2006. SMTC: A Swedish Map Task Corpus. *Lund Working Papers in Linguistics* 52:57–60.

Kristoffersen, G. 2000. *The phonology of Norwegian.* Oxford University Press.

Kuronen, M. 2003. Finns det supradentala konsonanter även i finlandssvenskan? *Svenskans beskrivning* 26:172–177.

Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. *Speech Production and Speech Modelling* 55:403–439.

Maddieson, I. 1984. *Patterns of Sounds*. Cambridge University Press.

Malmberg, B. 1968. *Svensk fonetik*. Lund: Gleerup.

Riad, T. 2010. Retroflektering. In: K. Jóhannesson et al. (eds.), *Bo 65: Festskrift till Bo Ralph*, Göteborg: Meijerbergs institut för svensk etymologisk forskning, 214–227.

Sjölander, K. & J. Beskow. 2001. WaveSurfer – an open source speech tool. *Proceedings of ICSLP 2000*, 464–467.

Vanvik, A. 1972. A phonetic-phonemic analysis of Standard Eastern Norwegian. *Norwegian Journal of Linguistics* 26:119–164.

Witting, C. 1959. *Physical and functional aspects of speech sounds: with special application to standard Swedish*. Uppsala: Almqvist & Wiksells.

Visual speaker gender affects vowel identification in Danish

Charlotte Larsen & John Tøndering
Department of Scandinavian Studies and Linguistics, University of Copenhagen, Denmark

Abstract

The experiment examined the effect of visual speaker gender on the vowel perception of 20 native Danish-speaking subjects. Auditory stimuli consisting of a continuum between /muːlə/ 'muzzle' and /moːlə/ 'pier' generated using TANDEM-STRAIGHT matched with video clips of a female and a male speaker were used to determine whether visual speaker gender affected Danish listeners similarly to American English-speaking listeners tested in a similar way.

Introduction

The purpose of the experiment reported here is to determine whether visual information about speaker gender affects Danish listeners' vowel perception. Participants were presented with audiovisual stimuli consisting of combinations of a woman's and a man's face with two auditory continua with vowel qualities between /muːlə/ 'muzzle' and /moːlə/ 'pier', one with a woman's voice, the other with a man's. Based on similar experiments conducted with American English speakers and listeners, the participants were expected to identify more steps on both auditory continua as /muːlə/ when they were paired with the female face than when they were paired with the male.

The existence of such an effect of visual gender on the vowel perception of Danish listeners would not only provide information on listeners' expectations regarding women's and men's speech, it would also contribute to theories of speech perception, and in particular to theories of speaker normalization.

Speaker normalization

It is widely accepted that most men's voices sound different from most women's. This is partly due to differences in average size and shape of the vocal tract – e.g. women's vocal tracts tend to be shorter than men's, producing vowels with higher average formant frequencies (Ladefoged & Broadbent, 1957). Differences may also result from factors not directly linked to physical differences, as evidenced by the fact that listeners are able to tell girls' voices apart from boys' before the children are old enough to have developed the physiological differences which might otherwise account for voice differences (Perry, Ohde & Ashmead, 2001). This suggests that social or cultural factors may also be involved.

For a variety of reasons, then, sounds which listeners have no trouble categorizing as instances of the same phoneme are acoustically quite variable depending on the gender of the speaker. Theories of speech perception need to be able to account for this through an explanation of the phenomenon known as speaker normalization, the process by which listeners fit input from individual speakers to phoneme categories available in their language.

Theories of speaker normalization have traditionally focused on physical differences between speakers, for example the theory, a version of which is put forth by Potter and Steinberg (1950), that sets of vowels produced by different speakers have about the same *relative* distribution in acoustical or auditory space. Another theory holds that listeners construct a mental model of a speaker's vocal tract, allowing them to correct for the effect of its size and shape and extract a set of absolute formant frequencies common to all speakers of a particular dialect independent of what Joos (1948) terms 'PERSONAL ERROR' (1948: 6).

Neither theory accounts for variation caused by non-physiological factors, however, and the contribution of visual cues to speech perception is ignored entirely.

Visual integration in speech perception

Perhaps the most famous example of how visual information affects speech perception is the McGurk-effect (McGurk & MacDonald, 1976) which demonstrated how listeners could be made to perceive a third sound by presenting them with audiovisual stimuli with visual articulatory information pointing to one sound, auditory information to another.

Later experiments have shown a similar effect of articulatory information in vowel perception, e.g. Traunmüller and Öhrström (2006).

However, there is evidence that information about place or manner of articulation is not the only type of visual information relevant to vowel perception. In a series of experiments, Johnson, Strand and D'Imperio (1999) found that American English listeners were likely to perceive more steps on a continuum of auditory stimuli comprising a continuum between /hood/ and /hud/ as /hood/ when they were paired with a video clip of a woman speaking than when the speaker was visually male. An effect was found not just of visual gender but also of the degree of gender stereotypicality of different voices and faces as judged by another group of participants. This last finding in particular would be difficult to explain under a theory of speaker normalization which only concerns itself with average physiological gender differences.

Based on these findings, which suggested that more than one kind of visual information was integrated during speech perception, the authors advocated a theory of speaker normalization which includes 'abstract, subjective talker representations' (ibid; 1999:380).

The experiment described in this paper is based on one of the experiments presented in Johnson, Strand and D'Imperio (1999), aiming to discover whether a similar effect of visual gender can be shown for Danish speakers and listeners.

Method

The stimuli were produced using sound and video recordings (head and shoulders) of a 31 year old woman and a 34 year old man pronouncing the words /mu:lə/ and /mo:lə/. Video was recorded in QuickTime format in the resolution 1920×1080 using a JVC GY-MH100 camera, while sound was recorded in wav format, 16 bit, 44,100 Hz stereo using an Olympus LS10 digital recorder.

Several repetitions of each word were produced, and video and sound clips chosen for further manipulation did not come from the same instance in the original recording, removing the risk of some finished audiovisual stimuli seeming better synchronized than others.

In order to avoid a learning effect relating a particular intonation to a particular vowel quality, the four sound clips were manipulated in Praat (v. 5.3.03) to keep F0 constant throughout each clip and identical for the two clips produced by the same speaker, at 220 Hz for the female voice, 133 Hz for the male voice, these frequencies being the averages of the average pitch values for each pair of clips.

The two pairs of words were manipulated using TANDEM-STRAIGHT, a speech manipulation program which allows auditory morphing based on source-filter analysis of recorded speech (Kawahara et al., 2009). For each speaker, a continuum was generated with nine steps between [mo:lə] and [mu:lə] which will be referred to as auditory stimulus 1–9, 1 being 100% [mo:lə], 9 being 0% [mo:lə], that is, 100% [mu:lə]. As the purpose of the experiment was primarily to reveal the difference in perceptual phoneme boundary of one set of stimuli compared to others, and not to provide an absolute value of, e.g., frequency, no perceptually motivated scale was used to determine the degree of morphing; the nine steps were simply morphed with equal percentual intervals so that stimulus 5 equals 50% morphing between [mo:lə] and [mu:lə].

Likewise, for each stimulus all parameters which TANDEM-STRAIGHT manipulates were set to the same degree of morphing, except 'Time' which was set to 50% morphing for all stimuli in order to avoid differences in synchronization between sound and video. This approach yielded a continuum of auditory stimuli which are morphed in more dimensions than, for example, a continuum of synthetically generated vowels inserted in the desired context, but it also means that it is impossible to determine exactly which acoustical features were relevant for participants' perception of vowel quality in the finished stimuli.

Using Final Cut Pro 7, each of the 18 auditory stimuli (two voices × nine steps on the vowel quality continuum) was paired with each of the four visual stimuli (two faces × two visual pronunciations, /mo:lə/ and /mu:lə/), creating a total of 72 different audiovisual stimuli. In order to avoid a sequence effect, stimuli were administered using one of four randomized lists.

Twenty-three linguistics students from University of Copenhagen participated in the experiment. Of these, 20 had Danish as their native language while three had Faroese or Faroese and Danish. Only the replies of participants with Danish as their (only) native language were included in the study. Of these 20, 12 were female, 8 were male. Average birth year was 1988, median birth year 1990, and participants' regional backgrounds were mixed, with 75% having been raised on Zealand.

Participants were tested individually in a quiet room using a laptop computer and a set of headphones. As it was vital for participants to keep their eyes on the screen while the stimuli were played, they were instructed to answer verbally, and in order to avoid any effect of participants themselves pronouncing the stimulus words between stimuli, the replies were given in the form of the numbers 'one' and 'two' rather than the words themselves.

Results

Overall, the results of the experiment showed the expected effect of visual speaker gender on vowel perception. 56.1% of stimuli with the visually female speaker were perceived by participants as /mu:lə/ while the same was true for only 50.4% of stimuli with the visually male speaker. A chi squared test showed this difference to be significant ($p < 0.01$).

There were, however, substantial differences in the way the effect manifested itself in different sets of stimuli – or failed to show up at all.

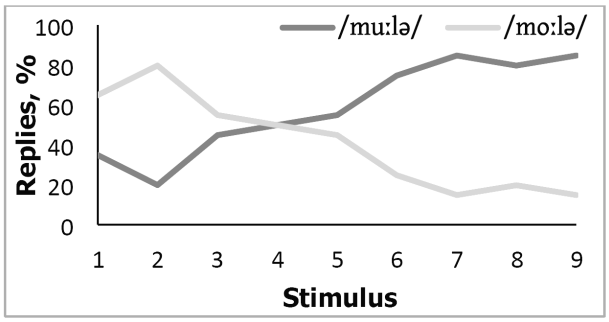


Figure 1. Replies for the set of stimuli with visual /mo:lə/-pronunciation+female voice+female face.

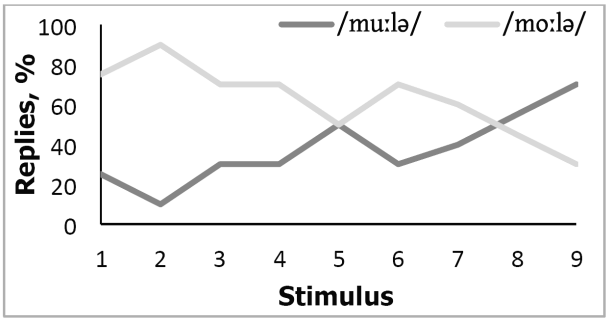


Figure 2. Replies for the set of stimuli with visual /mo:lə/-pronunciation+female voice+male face.

Figures 1–2 are examples of the visual representation of the results, showing the answers for two sets of stimuli which differ only in the variable visual gender. The point where answers are split evenly between /mu:lə/ and /mo:lə/ was chosen as a convenient numerical measure for further analysis, hereafter named *the perceptual crossover point*. It should be mentioned that some sets had two potential crossover points, as seen in *Figure 2*. Here, 50% of participants perceived stimulus 5 as /mu:lə/, however, as all stimuli 1–7 with the exception of stimulus 5 were perceived as /mo:lə/ by more than 50% of participants, 7.67 is considered the actual perceptual crossover point.

Table 1. Perceptual crossover points for all combinations of visual and auditory gender and visual pronunciation.

	Visual /mo:lə/		Visual /mu:lə/	
	Female voice	Male voice	Female voice	Male voice
Female face	4	5.63	4.14	3.6
Male face	7.67	7.25	2.2	4.64

As *Table 1* shows, the expected effect – a perceptual crossover point closer to 1 (= more /mu:lə/-answers) for stimuli sets with the female face than with the male – is seen for three out of four combinations of auditory gender and visual pronunciation. However, for the set with a male voice and visual /mu:lə/-pronunciation, the number of /mu:lə/-replies across all stimuli is actually close to being the same for both visual genders, 56.1% for the visually male speaker, 55% for the visually female one – a small difference in the opposite direction of the one predicted by the hypothesis, despite the perceptual crossover point for the set with the visually female speaker being closer to 1 as predicted.

A possible explanation for this discrepancy is found in the fact that there is considerably less agreement about the classification of stimuli for the set showing a female face with a male voice than for the set with matched visual and auditory gender. Stimulus 2–6 in the gender-mismatched set were each identified by less than 70% of participants as being either /mu:lə/ or /mo:lə/, and no stimulus in the set was classified the same by 90% of participants. Generally, there was less agreement about the sets with mismatch between visual and auditory gender than about the ones with matched genders, possibly because participants were aware of and distracted by the discrepancy.

The variables auditory gender and visual pronunciation were also found to have the expected effect on vowel perception, that is, listeners identified significantly more stimuli as /muːlə/ when they heard the female voice or saw the speaker of either gender pronouncing this word.

Discussion and conclusion

The experiment demonstrated the integration during vowel perception of two distinct kinds of visual information: articulatory information, the effect of which was expected based on the findings of e.g. Traunmüller and Öhrström (2006), and visual information about speaker gender. The effect of visual speaker gender was similar to the one found by Johnson, Strand and D'Imperio (1999), so the results appear to support their view of speaker normalization as based on several different kinds of information and partly dependant on listeners' representations of speakers, not only with respect to the size of their vocal tract.

While the findings certainly support this view, when taken alone, they are not strictly incompatible with a theory of normalization based on individual physical differences – e.g. listeners may simply have noted that the male speaker was larger than the female. An effect of the perceived visual gender stereotypicality of speakers independent of speaker size would disprove this alternative explanation, and this will be the focus of further research.

Regarding the methodology of the experiment, it should be mentioned that, surprisingly, out of all audiovisual stimuli, only one was identified by all twenty participants as the same phoneme. To our ears, the end points of each manipulated auditory continuum were all clearly identifiable as the word they were 'meant' to represent when heard in isolation, suggesting that the ambiguity arose from the combination of auditory and visual stimuli, but as this was not verified by a separate test, we cannot rule out the possibility that the process used for auditory morphing in itself introduced an unintended perceptual ambiguity.

Furthermore, as only four video clips were used, each representing a combination of gender and pronunciation, variation between clips, such as overarticulation on one clip, may have seriously impacted results, masking the effect of visual gender. Analysis of results broken down by visual stimulus suggests this may well have been the case for the two sets of stimuli which did not show the expected effect, but further research would be necessary to determine whether this was in fact the case, as well as to determine the exact relation between the method of auditory manipulation and the perceptual ambiguity discussed above.

Finally, as the participants in this study were not selected to be representative, and there is a strong possibility that participant variables such as age, gender and regional background affect the outcome, the findings cannot be said to apply to Danish listeners in general. The mere fact that an effect was shown for this particular group does however demonstrate that the effect of visual gender on vowel perception is not unique to the American English-speaking populations examined by Johnson, Strand and D'Imperio (1999) and others, and underscores the need for theories of speaker normalization to take into account not just physical differences between speakers of different genders but also, for example, listener expectations of how women and men are 'supposed' to speak.

References

Johnson, K., E. A. Strand & M. D'Imperio. 1999. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27:359−384.

Joos, M. A. 1948. Acoustic phonetics. *Language* Supplement 24(2):1–136.

Kawahara, H., T. Toru, M. Takahashi, M. Morise & H. Banno. 2009. Development of exploratory research tools based on TANDEM-STRAIGHT. *Proc. APSIPA*, Sapporo, 111–120.

Ladefoged, P. & D. D. Broadbent. 1957. Information Conveyed by Vowels. *Journal of the Acoustical Society of America* 29(1):98–104.

McGurk, H. & J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264:746–748.

Perry, T. L., R. N. Ohde & D. H. Ashmead. 2001. The acoustic bases for gender identification from children's voices. *Journal of the Acoustical Society of America* 109(6):2988–2998.

Potter, R. K & J. C. Steinberg. 1950. Toward the Specification of Speech. *Journal of the Acoustical Society of America* 22(6):807–820.

Traunmüller, H. & N. Öhrström. 2007. Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics* 35:244–258.

Perceptual evaluations of children with language impairment and deviant Voice Onset Time

Inger Lundeborg, Theodor Ricklefs & Lovisa Tunedal
Department of Clinical and Experimental Medicine, Division of Speech and Language Pathology, Linköping University, Linköping, Sweden

Abstract

This study investigated how naïve listeners assess words with initial plosives and deviant voice onset time produced by children with language impairment. Thirty-four adults (19 ♂ and 15 ♀) aged 22–60 years assessed 102 words with deviant VOT produced by children (4;2–11;6 y) recorded at SLP-clinics and language pre-school and school units. The participants identified 79% of the words correctly. Words with deviant short VOT had a slightly higher rate of error responses in comparison to words produced with a deviant long VOT. Most participants perceived the words with deviant long VOT as produced with voiceless plosives and words produced with deviant short VOT as having a voiced plosive. The results indicate that other factors than VOT must be taken into consideration for perceptual discrimination of the plosives.

Introduction

Children with speech and language deficits form a large but heterogeneous group. In many cases the causation is obvious. However, there is a large subgroup with articulatory problems without identifiable aetiology.

Within the field of speech and language disorders there is an on-going debate whether the problems of these children are caused by insufficient linguistic knowledge, processing deficits or limitations of motor speech control (Marinis, 2008; Alcock, 2006). Regarding the aspect of motor speech control, different speech sounds require different degree of motor coordination and control. Plosives are among the sounds that place high demands on motoric skills with requirement of a close co-ordination between the larynx and the lips, tongue and jaws (Auzou et al., 2000).

Auditory perceptual evaluations form the basis for the routine assessment of speech and language in clinical speech and language pathologist practice. However, by using acoustic analysis further insights of error patterns can be provided (Ballard & Robin, 2002).

Voice onset time is regarded to be a reliable acoustic cue for the distinction between voiceless and voiced plosives (Helgason & Ringen, 2008), and is considered to reflect the co-ordination between articulatory gestures and phonation (Hoit-Dalgaard, Murray & Kopp, 1983). It is defined as the time between the release of the oral closure and the onset of voicing and is measured in milliseconds. For the voiceless plosives typical Swedish adult values range between 49 and and 78 milliseconds and the corresponding range for the voiced plosives is –91 and –61 milliseconds (Lundeborg et al., 2012).

The perceptual boundary for correctly identifying a plosive as voiced or unvoiced is between +20 and +40 ms (Zlatin & Koenigsknecht, 1976). In two studies of typically developing Swedish children, a clear developmental trend was found both regarding length of VOT for the voiceless plosives and the occurrence of prevoicing in their voiced counterpats.

Adult-like values were established somewhere between 9 and 10 years of age (Larsson, & Wiman, 2010; Lundeborg et al., 2012). In children with phonological impairment atypical VOT-patterns are found and the variability is greater compared to normative data (Bond & Wilson, 1980; Lundeborg et al., submitted).

The aim of the present study is to investigate how productions with atypical VOT-values by children with speech and language impairment are perceived by adult naïve listeners.

Material and methods

A total of 34 mono-lingual Swedish speaking adults (19 ♂ and 15 ♀, 22–60 years, median age 31.5 years) with normal hearing and no specific knowledge of speech and language development participated in the study. The

participants listened to a sound file consisting of recordings of 102 plosive words with deviant VOT produced by 38 children with speech and language impairment.

The material was collected in a previous study (Lundeborg et al., submitted). VOT-values above or below one standard deviation from the mean values for the typically developed children in Larsson and Wiman (2010) and Lundeborg et al. (submitted) were defined as deviant. The participant's task was to select which one out of two words within a minimal pair they heard and also indicate if they were secure or insecure when choosing. A random selection of ten productions was duplicated into the sound file to check for intra-rater agreement.

Statistical analysis

Data were expressed with descriptive statistics for the occurrence of wrong answers. Comparisons between the genders and between words with VOT deviation above and below 1 SD from norm values were made by the use of MannWhitney U-test and a $p < 0.05$ was considered statistically significant.

Results

The participants identified 79% of the words correctly. Words with deviant short VOT-values (below 1SD from norm value) were harder to identify than words with deviant long VOT-values ($p < 0.001$). No gender difference was found. The greatest number of indications for unsecure choices was for the words with deviant long VOT. The intra-rater agreement was .89.

Discussion

Despite listening to productions with deviant VOT values between 1 and 3 SD form norm values the proportion of correct identifications was considerably high (79%). This could be interpreted as that the listeners used other cues in the word identification process. One such factor could be the fortis-lenis distinction, namely that the voiceless plosives are produced with a higher burst intensity than their voiced counterparts. Another relevant factor could be that the voiceless plosives in Swedish in a stressed syllable have a prominent aspiration (Lindblom, 2008). Speech rate is also reported to have influence on perception of syllables with initial plosives (Kessinger & Blumstein, 1998). Further research is needed regarding which cues listeners use when identifying words.

Conclusions

Despite deviances in VOT in the production of plosive-initial words the perceptual assessments of the productions were fairly adequate. Further research is needed of which cues listeners use in the identification process when assessing words with initial plosives.

Notes

The study was carried out in accordance with the etical principles for medical research of the Helsinki declaration as revised in 2008 (World Medical Association, 2008).

References

Alcock, K. 2006. The development of oral motor control and language. *Down Syndrome Research and Practice* 11(1):1–8

Auzou, P., C. Özsancak, R. J. Morris, M. Jan, F. Eustache & D. Hannequin. 2000. Voice onset time in aphasia, apraxia of speech and dysarthria: a review. *Clinical Linguistics and Phonetics* 14(2):131–150.

Ballard, K. J. & D. A. Robin. 2002. Assessment of AOS for treament planning. *Seminars in Speech and Language* 23(4):281–292.

Bond, Z. S. & H. F. Wilson.. 1980. Acquisition of the Voicing Contrast by Language-Delayed and Normal-Speaking Children *Journal of Speech and Hearing Research* 23:152–161.

Helgason, P. & C. Ringen. 2008. Voicing and aspiration in Swedish stops. *Journal of Phonetics* 36:607–628.

Hoit-Dalgaard, J., T. Murray & H. G. Kopp. 1983. Voice onset time production and perception in apraxic subjects. *Brain and Language,* 20(2):329–339.

Kessinger, R. H. & S. E. Blumstein. 1998. Effects of speaking rate on voice-onset time and vowel production. *Journal of Phonetics* 26; 117–128.

Larsson, M. & S. Wiman. 2010. Voice onset time hos svenska förskolebarn – Ett utvecklings-perspektiv (Voice onset time in Swedish preschool children - a developmental perspective). Kandidatuppsats i logopedi (bachelor thesis), Linköping University http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-55537

Lindblom, B. 2008. *Röst- och talfunktion*. In: L. Hartelius, U. Nettelbladt & B. Hammarberg (eds.), *Logopedi*. Lund: Studentlitteratur, 95–102.

Lundeborg, I., M. Larsson, S. Wiman & A. McAllister. 2012. Voice Onset Time in Swedish children and adults. *Logopedics Phoniatrics Vocology* 37(3):117–122.

Lundeborg, I., M. Zeipel-Stierna, E. Nordin & A. McAllister. Submitted. Voice Onset Time in Swedish children with phonological impairment. Manuscript submitted to *Logopedics Phoniatrics Vocology*.

Marinis, T. 2008. Syntactic Processing in developmental and aquired language disorders. In: M. J. Ball, M. R. Perkins, N. Muller & S. Howard (eds.), *The handbook of clinical linguistics*. Oxford: Blackwell Publishing.

World Medical Association Declaration of Helsinki, 2008. Available at: http://www.wma.net/en/30publications/10policies/b3/

Zlatin, M. A., & R. A. Koenigsknecht. 1976. Development of the voicing contrast: A comparision of voice onset time in stop perception and production. *Journal of Speech and Hearing Research* 19:93–111.

The effect of vowel height on Voice Onset Time in stop consonants in CV sequences in spontaneous Danish

Johannes Mortensen & John Tøndering
Department of Scandinavian Studies and Linguistics, University of Copenhagen, Denmark

Abstract

Voice onset time has been reported to vary with the height of vowels following the stop consonant. This paper investigates the effects of vowel height on VOT in Danish CV sequences with stop consonants in Danish spontaneous speech. A significant effect of vowel height on VOT was found in the unaspirated stops [b̥ d̥ g̊], but not in the aspirated stops [pʰ tˢ kʰ]. This is contrary to previous findings.

Introduction

The purpose of this paper is to investigate the effect of vowel height on voice onset time (VOT). Earlier studies have found that VOT before high vowels is significantly longer than VOT before low vowels (Barry & Moyle, 2011; Bijankhan & Nourbakhsh, 2009; Esposito, 2002; Fischer-Jørgensen, 1980; Higgins, Netsell & Schulte, 1998). These previous studies are all based on readings of carefully designed material (read aloud speech), mainly consisting of nonsense words. Krull (1991) compared VOT in Swedish laboratory speech with spontaneous speech. She found no significant differences between VOT in the two speaking styles.

In the present study we have examined CV sequences with stop consonants and stressed vowels in a data set of Danish spontaneous speech. As a measure of vowel height the physiologically defined degree of vowel opening is used. The basic research question to be answered is whether the same relationship between VOT and vowel height exists in spontaneous speech as in read aloud speech.

Note on the Danish stop consonants

The Danish stop consonants can be divided into two series: /ptk/ which are voiceless aspirated stops (except that /t/ is affricated and aspirated) and /bdg/ which are voiceless unaspirated stops. The aspiration distinction is present only in syllable initial position before vowels and sonorant consonants. In medial position the stops are often pronounced as weakly voiced [b d g] (Fischer-Jørgensen & Hutters, 1981). This study, however, is only concerned with syllable initial stops followed by stressed vowels. In this position the stops are pronounced [pʰ tˢ kʰ] and [b̥ d̥ g̊]. In the following the simplified transcriptions [ptk] and [bdg] will be used.

Material

The speech material was taken from the *Danish Phonetically Annotated Spontaneous Speech Corpus* (DanPASS, Grønnum, 2013). The recordings where made at The University of Copenhagen in 1996 and 2004. The speakers were given specific map tasks to talk about in both dialogues and monologues. Although the recordings where made in a studio, the corpus represents an approximation to speech in a natural setting.

The total duration of the recordings is more than 9 hours. The material contains recordings of 27 speakers, 10 women and 17 men. The number of tokens of relevant CV-sequences – i.e. a stop consonant followed by a stressed vowel – exceeds 7,000, but only a little more than 3.000 tokens were included in the present investigation. All tokens in the corpus with [ptk] were included (1.967), whereas tokens with [bdg] were taken from 5 randomly selected speakers (1.074). This procedure was chosen because a very clear pattern was found for tokens with [bdg], whereas this was not the case for [ptk].

Method

Segmentation

Measuring VOT is, to a large extent, dependent on the principles of delimitation. In determining the starting and ending point of VOT, the recommendations in Fischer-Jørgensen and Hutters (1981) were followed. Accordingly, VOT starts at the first release of the closure, which can be located in the waveform, and ends with the start of the higher formants of the following vowel, which can be located in the spectrogram. There is an uncertainty of some milliseconds in both points.

Measurements were only made in stressed syllables, and in case of doubt about the segmentation – e.g. if the release of the closure was so smooth that it could not be located in the waveform – no measurement was made.

Vowel classification

Apart from using spontaneous speech, another new methodological approach in this study is the use of physiological vowel description. In the traditional vowel classification the distance from tongue to palate defines vowel height, but in the physiological description vowel height is defined as the degree of constriction at the narrowest point in the vocal tract (Grønnum, 2005). We have chosen to use the physiological vowel description because the most prevailing explanation for the effect of vowel height on VOT is related to the flow of air through the vocal tract. This airflow does not depend (directly) on tongue height, but more precisely on the degree of constriction, or conversely on the degree of opening: One of the conditions for the vocal folds to start vibrating – i.e. voice onset – is that the air pressure above the glottis is lower than the air pressure below the glottis. During the articulation of a stop consonant, however, the supraglottal pressure will rise because the flow of air is restrained at the point of constriction so that the air builds up behind this point – the place of articulation of the stop consonant. After the release of the closure the air will flow rapidly out of the vocal tract and the pressure just above glottis will start to fall again. It will take a little time for the air pressure to drop sufficiently for the vocal cords to start vibrating. How long it takes will depend on the size of the passage through the mouth. If the passage is wide the pressure will drop quickly, but if the passage is narrow the pressure will drop more slowly. The consequence of this is that the narrow passage in high vowels delays the airflow through the mouth and thereby also delays voice onset and increases VOT (Fischer-Jørgensen, 1980). The physiological classification of the Danish vowels follows Grønnum (2005:105), except from merging her levels 4 and 5, resulting in four levels on the scale of degree of opening. The classification of the stressed Danish vowels can be seen from *Table 1*. Level 1 is the narrow end of the scale and level 4 is the open end (notice that the terms high and low makes less sense here, although high would correspond to level 1 and low would correspond to level 4). The merging of level 4 and 5 is done partly because the vowels in these levels are rather infrequent resulting in fewer tokens in the material, partly because some of them can be suspiciously difficult to distinguish.

Table 1. Physiological classification of the degree of opening of Danish vowels. The phonetic transcription is according to the modified version of the IPA for Danish (Grønnum, 2005).

Degree of opening			
1	2	3	4
[i]	[e]	[ɛ]	[æ]
[y]	[ɔ]	[ø]	[œ]
[u]	[ʌ]	[ɑ]	[œ̝]
[o]			[a]
[ɒ]			[ɶ]

A drawback of using this physiological classification of the Danish vowels is that it is based on a combination of x-rays of vowel articulation of one person and introspection (Grønnum, 2005). A proper physiological investigation on this aspect of the Danish vowels does not exist. However, since this study has no intention of detecting minor fluctuations in VOT, but only intends to reach conclusions regarding the overall tendency in the relation between vowel height and VOT, Grønnum's more or less introspective classification should be an acceptable point of departure.

Statistical approaches

The data was analysed using plots and statistical modeling. Mixed effects multiple regression models were used with speaker and word as random effects and the degree of vowel opening as fixed effect. Models with additional variables – such as gender of the speaker – were also tried out, but this did not yield particularly interesting results and will not be reported here. Each stop consonant was analysed separately and after fitting the models, residual diagnostics were carried out to validate model assumptions.

It should be noted, that in order to be able to use regression modeling, the degree of vowel opening was treated as a continuous variable although it strictly speaking is only ordinal. It is reasonable to think that Grønnum (2005) has chosen the levels of vowel height so that the distance between them is fairly constant, but as noted earlier the classification is partly introspective and also the two highest levels are taken as one in this study.

Results

Visual inspection

Figure 1 and *Table 2* show the mean VOT for the different stops according to vowel height. In [bdg] there is a clear tendency for a fall in VOT when the degree of vowel opening increases. The only exception is a rise in [b] from level 3 to 4. The difference in the average VOT of the narrowest vowels and the most open vowels is a bit less than 10 ms.

Table 2. Mean VOT in ms.

Stop consonant		Degree of vowel opening			
	No.	1	2	3	4
[b]	156	19.2	16.3	11.7	14.0
[d]	608	29.2	24.6	20.5	18.0
[g]	310	34.0	29.4	28.2	25.1
[p]	341	77.0	64.0	71.9	58.1
[t]	600	86.6	87.2	81.6	88.7
[k]	1.026	81.0	61.6	60.5	76.1

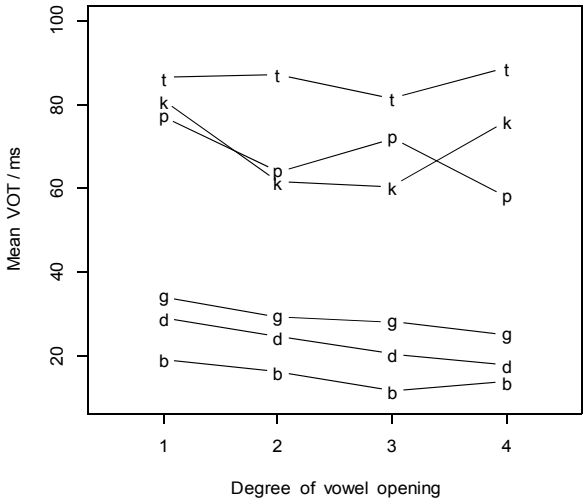


Figure 1. Mean VOT as a function of vowel opening degree.

In [ptk] the picture is less clear. In [k] there is a steep fall from level 1 to 2, then almost no difference from 2 to 3, and then a steep rise to level 4. In [p] the pattern is falling–rising–falling. In [t] the variation seems to be rather small although there is a slight fall from level 2 to 3 and a somewhat larger rise from level 3 to 4. Overall there does not seem to be any consistent effect of vowel height on the VOT of [ptk].

It should be noted that the variation in VOT between individual tokens of CV sequences is larger than the variation across different vowel heights, so that for instance a [d] in level 4 can easily have a longer VOT than a [d] in level 1. This is not apparent in *Figure 1* because it only shows the mean values. Still it is clear that at least in [bdg] vowel height accounts for some of the variation in VOT.

Statistical analysis

Table 3 shows coefficients, standard deviations and p-values for the mixed effects models with degree of vowel opening as the fixed effect. In [b], [d] and [g] the effect of vowel height is significant with p-values below 0.001. In [p], [t] and [k] the effect of vowel height is not significant.

Table 3. Result of the statistical analysis with degree of vowel opening as fixed effect.

Stop consonant	Coefficient	Standard deviation	*p*-value
[b]	-0.173	0.040	< 0.001
[d]	-0.166	0.019	< 0.001
[g]	-0.104	0.018	< 0.001
[p]	-1.766	2.345	0.452
[t]	-0.014	0.016	0.376
[k]	-0.031	0.020	0.114

All the coefficients are negative indicating a fall in VOT when the degree of opening increases. This is also the case in [ptk] although the coefficients here are very small and the effect is insignificant. The reason for the seemingly high values in [p] is technical: the VOT values of the other stop consonants were log-transformed because their distributions were skewed, which was not the case for [p].

Supplementary analysis

To check whether the lack of effect of vowel height on VOT in aspirated stops could be due to the use of the physiologically based classification of vowel height, we also examined the association between VOT and vowel height according to the traditional description. In this analysis the Danish vowels were classified in three degrees as high ([i y u e]), mid ([ɛ ø o œ ɔ]) or low ([a ɶ ʌ ɑ æ ɐ̯ ɒ]). The mean VOT divided by vowel height is shown in *Table 4*. As seen, the overall pattern is the same as when using physiological vowel classification.

Table 4. Mean VOT in ms divided by vowel height.

Stop consonant	Vowel height		
	High	Mid	Low
[b]	19.0	18.0	12.7
[d]	25.9	21.0	18.4
[g]	36.0	28.8	27.4
[p]	70.1	65.0	67.8
[t]	86.6	86.4	83.8
[k]	80.1	62.7	66.3

Discussion and conclusion

The effect of vowel height on VOT was investigated and for [bdg] a clear relation was found. VOT was shorter when the vowels were more open. This result is in agreement with previous findings (Fischer-Jørgensen, 1980, and others). For [ptk], however, no significant effect of vowel height on VOT was found. This is contrary to previous findings and calls for discussion.

The most obvious explanation might be methodological. The use of physiological vowel description could be the reason for the deviating result. However, the supplementary analysis based on the traditional vowel description yielded nearly the same results – ruling out the explanation of difference in vowel description.

The use of spontaneous speech might be another explanation, but it is hard to explain why this should have an effect in [ptk], but not in [bdg].

The deviating results could also be a consequence of differences in the criteria used for the delimitation of VOT. As described in Fischer-Jørgensen and Hutters (1981) it is possible to consider the vowel to start at different points, and the choice of point might have serious consequences for the length of VOT, especially before low vowels. Fischer-Jørgensen and Hutters (1981) argues in favour of choosing the start of the higher formants as the vowel onset. This suggestion is followed in the present study, but most studies take the start of voicing as the endpoint of VOT, which does not seem unreasonable when dealing with the concept voice onset time (Barry & Moyle, 2011; Bijankhan & Nourbakhsh, 2009).

There might, however, also be another explanation for the lack of effect of vowel height on VOT in [ptk] found in this study. As noted earlier, the most prevailing explanation in the literature for the effect of vowel height on VOT is related to the flow of air through the vocal tract. But the question is, whether this explanation holds when VOT is relatively long. In other words, maybe the VOT in [ptk] is so extensive that the sufficient oral pressure drop is reached long before voice onset – regardless of the height of the following vowel. If this is the case, vowel height might not have much effect on VOT in [ptk]. One way to investigate this could be to examine if the effect of vowel height on VOT in [ptk] is present for some speakers and not for others, and whether this possible pattern would be related to speakers's overall level of voice onset time. This will be the subject for further studies.

References

Barry, J. & M. Moyle. 2011. Covariation among vowel height effects on acoustic measures. *Journal of the Acoustical Society of America*, 130(5): 365−371.

Bijankhan, M. & M. Nourbakhsh. 2009. Voice onset time in Persian initial and intervocalic stop production. *Journal of the International Phonetic Association* 39:335−364.

Esposito, A. 2002. On vowel height and consonantal voicing effects: Data from Italian. *Phonetica* 59:197−231.

Fischer-Jørgensen, E. 1980. Temporal relations in Danish tautosyllabic CV sequences with stop consonants. *ARIPUC* 14:207−261.

Fischer-Jørgensen, E. & B. Hütters. 1981. Aspirated stop consonants before low vowels. A problem of delimitation. *ARIPUC* 15:77–102.

Grønnum, N. 2013. Danish Phonetically Annotated Spontaneous Speech.

Grønnum, N. 2005. *Fonetik og Fonologi. Almen og Dansk.* Copenhagen: Akademisk Forlag.

Higgins, M. B., R. Netsell & L. Schulte. 1998. Vowel-related differences in laryngeal articulatory and phonatory function. *Journal of Speech, Language and Hearing Research* 41(4):712−724.

Krull, D. 1991. VOT in spontaneous speech and in citation form words. Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (*PERILUS*). No. XII:101−107.

Focus type effects on focal accents and boundary tones

Sara Myrberg
Department of Swedish Language and Multilingualism, Stockholm University, Stockholm, Sweden

Abstract

The paper presents a production experiment on the prosodic realization of focus types in Stockholm Swedish. It is shown that focal accents are realized with higher f0 and longer duration when signalling narrow or corrective focus, than when signalling all-new focus. In addition, focal accents are more often followed by a L% when signalling narrow or corrective focus, and f0 falls due to L% are larger.

Introduction

In the literature, different types of focus have sometimes been distinguished and placed on a scale. It has been hypothesized that the higher up on the scale a focus is, the more likely is it that the focus will receive a phonological correlate, and/or the stronger will its correlate be. For example, Féry (to appear), proposes that *all-new information* focus is a weak focus type, *narrow information* focus is stronger, and *corrective* focus is yet stronger (but note that these are not the only focus types on her scale).

Many questions remain to be answered in relation to such focus scales, however. It is unknown how many focus types languages distinguish, and which the primary correlates of focus are in a typological perspective. *Prominence* in terms of e.g. increased f0 and duration play a very important role in the Germanic languages. In many other languages, *phrasing* has been shown to play a crucial role in focus expression (e.g. Féry, to appear). In the Germanic languages, the role of phrasing in the expression of focus is less well studied than the role of prominence.

In Swedish, the effect of focus types on the realization of the *focal accent* has not been well studied, in spite of the fact that the focal accent has several properties which makes it unique among (at least) Germanic languages (e.g. Myrberg, 2013). Also, in (Stockholm) Swedish, a categorical distinction can be made between focal accents directly followed by a *boundary tone fall*, here analyzed as L%, and focal accents followed by a *high plateau*. Plateaus result from interpolation with a following H, and appear in the absence of a L% (Bruce, 1987; Myrberg, 2013). Because Swedish intonation has these uncommon features, increased knowledge of the expression of focus types may have interesting typological implications.

This paper presents a production study addressing the question how the three focus types *all-new*, *narrow* and *corrective* focus affect *i*) the maximal f0 height in focal accents, *ii*) how often there is a *tonal fall* L% directly following a focal accent, as well as the size of such falls, and *iii*) word duration in these three focus types.

Experiment design

Six female Stockholm Swedish speakers read question–answer pairs as in (1).

(1) *Context questions*:

Q-a. Varför ringer äggklockan ute i köket?
Why is the timer ringing in the kitchen?

Q-b. Vad har kokat färdigt?
What is done boiling?

Q-c. Har äggen kokat färdigt?
Are the eggs done boiling?

Target sentences:

A-a. [1]**Hummern** har kokat färdigt.
The lobster is done boiling.

A-b. [2]**Havren** har kokat färdigt.
The oat is done boiling.

Each target sentence was read with three different context questions, triggering all-new focus (Q-a), narrow focus (Q-b), and corrective focus (Q-c). There were five different target sentences, each of which appeared once with an accent 1 subject (A-a), and once with an accent 2 subject (A-b). Every question–answer pair was repeated three times by each speaker. This procedure resulted in a corpus of 540 sentences (5 were lost due to technical problems).

Syntactically, the target sentences consisted of a one-word subject (disyllabic, initial stress), followed by an intransitive predicate. The predicates belong to the group of intransitives which allow an all-new reading when the strongest prominence of the sentence (a focal accent in Swedish) is located on the subject (e.g. Gussenhoven, 1983; Selkirk, 1984). In contrast to the sentences used here, most sentences can receive an all-new interpretation only when the strongest prominence appears sentence-finally.

It was important that the target word would receive the strongest prominence in all three focus conditions, and that it would appear non-finally in the sentence. Sentence-finally, there is always a L%, and this L% can be expected to reach the lowest f0 of the sentence, simply because it is final. However, in non-final position, the insertion and realization of a L% are independent of finality-effects. Thus, in non-final position, we can observe focus effects on L% insertion and realization.

The intransitives used here made it possible to compare focal accents that express all-new, narrow, and corrective focus, while keeping constant the position of the focal accent in the sentence.

Annotation procedure

The sentences were annotated by the author using Praat (Boersma & Weenink, 2013). In each accent 1 subject, a L*H focal accent contour was annotated, and in each accent 2 subject a H*LH focal accent contour was annotated. This annotation is based on standard assumptions of the tonal structure in Stockholm Swedish (e.g. Bruce, 1977, 1998; Heldner, 2001; Myrberg, 2013). Henceforth, the rightmost H-tones in the focal accent tonal sequences L***H** and H*L**H** will be referred to as the *focal H tone*. The focal H tones are structurally equal in words with accent 1 and accent 2, and correlate with information structural focus in Stockholm Swedish (Bruce, 1977).

In addition to the focal accent contours, the lowest f0 point following the focal H tone inside the subject was annotated. This low point corresponds to a potential L% following the focal accent. Furthermore, the author made a (subjective) judgment with respect to whether or not a L% followed the focal accent on the subject, and with respect to whether or not the predicate contained a focal accent in addition to the one on the subject.

Results and discussion

Four basic tonal patterns could be distinguished in the target sentences, as shown in *Table 1* (for sentences with an accent 1 target word) and *Table 2* (for sentences with an accent 2 target word). The patterns differ *i*) in terms of a focal accent in the verb phrase, and *ii*) in terms of a boundary tone (L%) immediately following the focal accent in the subject. The focal accent on the verb is present in a and b, and the boundary tone in the subject is present in a and c.

Table 1. Distribution of phrasing patterns of sentences in (1). Accent 1 target word. 100%=267

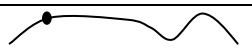
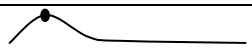
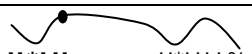
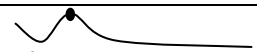
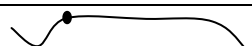
[1]**Hummern**... ...[2]färdigt	all-new (1) Q-a	narrow (1) Q-b	correct. (1) Q-c
a.  **L*H L%** H*LH L%	17.6% (47)	-	-
b.  **L*H** H*LH L%	1.5% (4)	-	-
c.  **L*H L%**	13.9% (37)	33.7% (90)	33.3% (89)
d.  **L*H** L%	-	-	-

Table 2. Distribution of phrasing patterns of sentences in (1). Accent 2 target word. 100%=268

[2]**Havren**... ...[2]färdigt	all-new (1) Q-a	narrow (1) Q-b	correct. (1) Q-c
a.  **H*LH L%** H*LH L%	6.3% (17)	0.3% (1)	-
b.  **H*LH** H*LH L%	10.1% (27)	-	-
c.  **H*LH L%**	3.7% (10)	16.8% (45)	17.5% (47)
d.  **H*LH** L%	12.7% (34)	16.4% (44)	16.0% (43)

The right sides of *Tables 1 and 2* show the number of occurrences in the three focus types for each pattern. As for the focal accent on the verb (a and b), it should be noted that this pattern appears only in the all-new condition.

As for the occurrence of boundary tones in the subject, there is an effect of focus type for the accent 2 words, but not for the accent 1 words. L% following the subject is as good as obligatory for accent 1 words in all focus types (hence patterns b and d are virtually unattested). However, in the accent 2 words, speakers insert a L% more restrictively. This restrictiveness allows the observation that L% insertion is more common with narrow and corrective focus than with all-new focus. This observation is made by comparing the number of occurrences of patterns a and c (with L%), to the number of occurrences of patterns b and d (without L%), in the three focus types. A Chi-square test shows a significant effect of focus type in relation to the presence of a L% (χ^2-squared = 10.5484, df = 2, p = 0.005122).

Much remains to be understood when it comes to the distribution of L% boundary tones directly following focal accents, but the fact that accent 2 (which contains 3 tonal targets) have such tones much less often than accent 1 (which contains 2 tonal targets) suggest that in

addition to focus type, space inside the target word is a major factor in determining whether a focal accent is followed by a L%.

It can be hypothetically concluded that when there is ample space, as in the accent 1 words, speakers insert a L% following focal accents, independent of context. However, even when there is some shortage of space, insertion of a L% is sometimes forced, as in the accent 2 words. It is in these cases, where L% is not always present, that we can observe the higher pressure for a L% to be inserted with narrow and corrective focus than with all-new focus.

The following sections describe in turn how focus type affects *i*) the f0 height of the focal H tone in the subject, marked by a dot in Tables 1 and 2, *ii)* the realization of the L% fall following the focal accent in the subject, as in patterns a and c, and *iii*) the duration of focally accented words.

Height of the focal H tone in the subject

The boxplots in *Figures 1 and 2* illustrate how the f0 of the focal H varies between the focus types.

A Wilcoxon test reveals that the focal H tone is significantly lower in the all-new condition than in the narrow condition. This is true for accent 1 ($W = 2840$, $p = 0.001617$) as well as accent 2 ($W = 2641$, $p = 0.000125$). The all-new condition is also significantly lower than the corrective condition, in accent 1 ($W = 3009$, $p = 0.01075$) and in accent 2 ($W = 2396$, $p < 0.0001$). Between the corrective and the narrow condition, however, there is no significant difference, for accent 1 ($W = 4357$, $p = 0.3106$) or 2 ($W = 3841$, $p = 0.5508$).

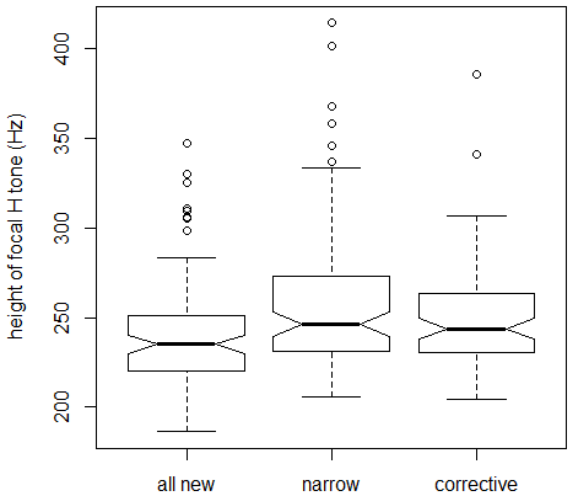


Figure 1. Height of focal H tone. Accent 1.

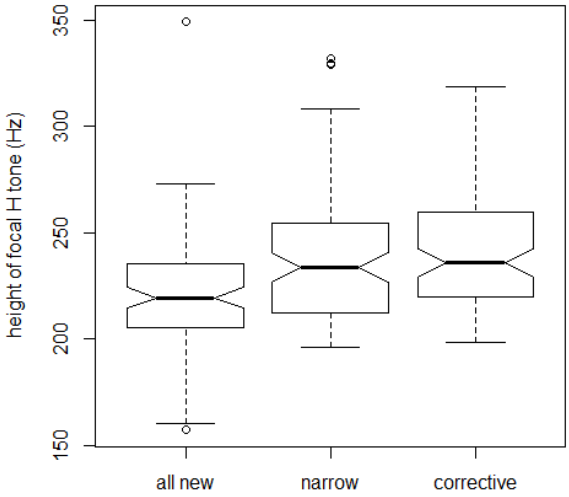


Figure 2. Height of focal H tone. Accent 2.

Size of the L% boundary fall

The size of the falls from the focal H tone to the L% in patterns a and c from *Tables 1 and 2* were measured, to see whether these falls were affected by focus type.

The fall size reported here amounts to the difference between the focal H tone (as reported in the previous section), and the lowest f0 point following the focal tone *within* the target word. Only sentences realized with a L% (i.e. patterns a and c from *Tables 1 and 2*) were included in the reported measures (it can be noted, however, that similar results were obtained when including sentences realized with patterns b and d).

The size of the fall in the three focus conditions is illustrated in *Figure 3* (for accent 1 subjects) and *Figure 4* (for accent 2 subjects).

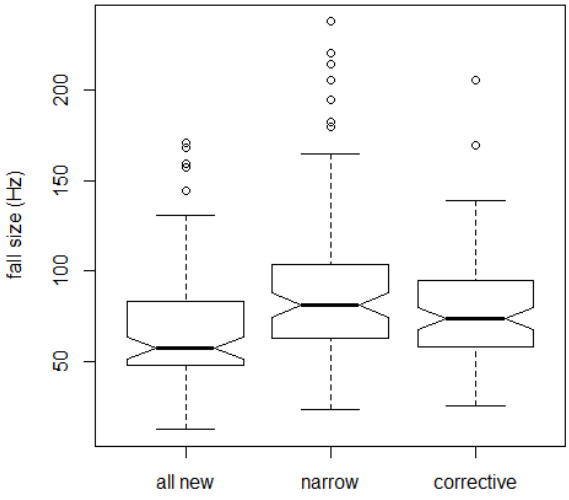


Figure 3. Fall size from the focal H to L%, in sentences with patterns a and c from Table 1. Accent 1.

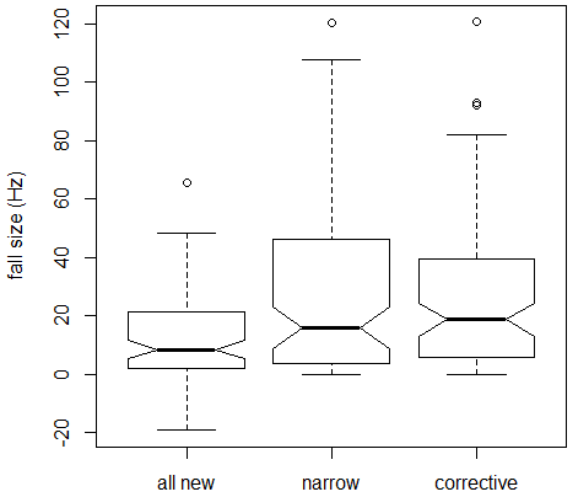


Figure 4. Fall size from the focal H to L%, in sentences with patterns a and c from Table 2. Accent 2.

A Wilcoxon test shows that the fall is smaller in the all-new condition than in the narrow condition, for accent 1 ($W = 2377$, $p < 0.0001$) as well as accent 2 ($W = 2849$, $p = 0.001234$). The all-new condition is also significantly different from the corrective condition, both for accent 1 ($W = 2689$, $p = 0.0009858$) and accent 2 ($W = 2648$, $p = 0.0001358$). However, the narrow and the corrective conditions are not significantly different (accent 1: $W = 4501$, $p = 0.1158$; accent 2: $W = 3931$, $p = 0.7346$).

Word duration

The duration of the target words, too, reflects the different focus conditions. The duration of the target words is illustrated in *Figure 5*.

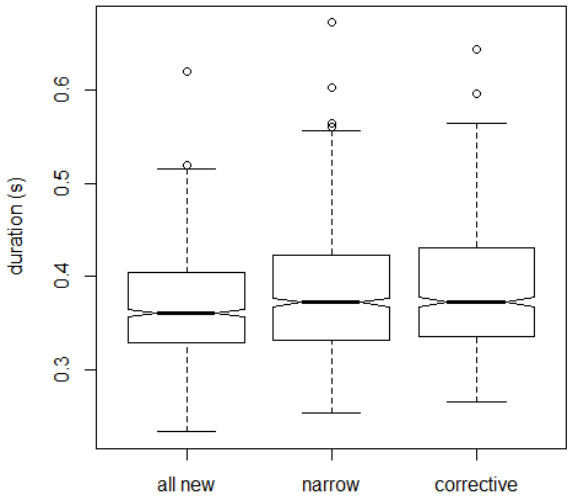


Figure 5. Duration of the target words in the three focus conditions. Accent 1 and 2.

The Wilcoxon test shows that words in the all-new condition are significantly shorter than words in the narrow condition ($W = 288730$, $p = 0.0005403$), as well as words in the corrective condition ($W = 285258$, $p = 0.0002355$). However, words in the narrow and the corrective conditions are not significantly different ($W = 323350$, $p = 0.7427$).

Conclusion

The results of the experiment reveal that all-new focus is distinguished from narrow focus and corrective focus in Stockholm Swedish. However, they also indicate that no distinction is made between narrow and corrective focus.

All-new focus is distinguished from narrow and corrective focus in terms of the f0 height of the focal H, in terms of the size of a L% tone fall inside a focally accented word, and in terms of word duration. The first two effects are interpreted here as effects on the phonetic realization of the tonal targets involved in expressing focus.

The results also show a higher pressure to insert a L% boundary tone after narrow and corrective focus than after all-new focus. This is interpreted as indicating that, alongside the obligatory focus prominence, there is a pressure to align focus with a prosodic edge in Swedish, along the lines of focus alignment as proposed by Féry (to appear).

Acknowledgements

For financial support, I gratefully acknowledge Sven och Dagmar Saléns Stiftelse and Anna Ahlström och Ellen Terserus Stiftelse.

References

Boersma, P. & D. Weenink. 2013. Praat: doing phonetics by computer [Computer program].

Bruce, G. 1977. *Swedish word accents in sentence perspective*. Lund: Liber Läromedel.

Bruce, G. 1987. How Floating is Focal Accent? In: K. Gregersen & H. Basbøll (eds.) *Nordic Prosody* 4. Odense: Odense University Press. 41–49.

Bruce, G. 1998. *Allmän och svensk prosodi. Praktisk Lingvistik* 16. Lund: Department of Linguistics, Lund University.

Féry, C. To appear. Focus as prosodic alignment. *Natural Language & Linguistic Theory* 31(4). (2013)

Gussenhoven, C. 1983. Focus, mode and the nucleus. *Journal of Linguistics* 19:377–417.

Heldner, M. 2001. *Focal accent – F0 movement and beyond. PHONUM* 8. Umeå University.

Myrberg, S. 2013. Sisterhood in prosodic branching. *Phonology* 30(1):73–124.

Selkirk, E. 1984. *Phonology and Syntax*. Cambridge, Massachusetts: The MIT Press.

Disfluency in child-directed speech

Kristina Nilsson Björkenstam [1], Mats Wirén [1] & Robert Eklund [2]
[1] Department of Linguistics, Stockholm University, Stockholm, Sweden
[2] Department of Culture and Communication, Linköping University, Linköping, Sweden

Abstract

We report results from a longitudinal study of the rate and location of disfluencies in child-directed speech, using data for children between 0;6 and 2;9 years. We compare these results to adult-directed speech by the same speakers.

Introduction

From a language acquisition perspective, disfluency (for example, "uh" and "um") is interesting because it could arguably make learning harder. Put differently, it looks like yet another manifestation of the poverty of the stimulus. Seen from this perspective, it is natural that child-directed utterances are not only short and slow, but also highly fluent compared to adult-directed speech (ADS). Even though the adult disfluency rate increases with the age of the child, child-directed speech (CDS) is consistently less disfluent than ADS (Broen, 1972). However, it has recently been shown that disfluencies contain information that helps the child to interpret the input from a certain age: disfluencies tend to occur before words that are unfamiliar, infrequent or new in the discourse, and thereby provide a cue about a speaker's intended referent or communicative intention (Kidd, White & Aslin, 2011). To corroborate this finding, we must begin by investigating the disfluencies that children hear at different ages. To this end, we report results from a longitudinal study of the rate and location in utterances of disfluencies in child-directed speech, using data for children between 0;6 and 2;9 years.

Fluency and disfluency in child-directed speech

Spontaneous speech in adult–adult conversations typically includes disfluencies such as filled pauses, segment prolongations, hesitations, repetitions, and truncated words at a rate of about 6% of all words uttered (Eklund & Wirén, 2010; Fox Tree, 1995).

When talking to young children, adults modify their speech, e.g. by using fewer words per utterance, slower speech rate, more repetitions, and decreased syntactic complexity compared to ADS (Broen, 1972). Typically, CDS is described as *fluent* speech (Clark, 2009:36). Over time, as caretakers use longer, more complicated utterances at a faster speech rate, the disfluency rate increases accordingly; Kidd, White and Aslin (2011) report that filled pauses occur at a rate of 1/1000 words in speech directed at 2-year olds in the CHILDES database, and that this rate increases with the age of the child. This can be compared to a reported filled pause incidence of 1.9% to 4.4% in scientific works covering the period 1959 to 2007 (Eklund, 2010:25).

The most prevalent type of disfluency is the filled pause (FP), e.g. *um, öh*. Eklund and Wirén (2010) list five hypotheses regarding the function(s) of FPs in speech:

1) Floor-holding hypothesis
2) Help-me-out hypothesis
3) Self-monitoring/error-detection hypothesis
4) Many-options hypothesis
5) Attention-getting signal

Eklund and Wirén (2010) point out that these hypotheses are not mutually exclusive and that FPs may serve more than one function, but that there is strong support for the many-options hypothesis. In a CDS scenario, the first two hypotheses are less likely than the latter three since the adult is typically very attentive to vocalizations by the child.

Corpus data

The data consist of audio and video recordings of free play sessions in a recording studio at the Phonetics laboratory at Stockholm University. The free play sessions are in most cases followed by a session when the parent and the experiment leader chat informally while working through The Swedish Early Communicative Development Inventory (SECDI, a version of the MacArthur Communicative Development Inventory) with the child in the room.

The data consist of 31 recordings of four children (age 6–33 months), three girls and one boy, interacting with their Swedish-speaking mothers or fathers (mean recordings/child 7, range 11–5).

All utterances by both parent and child in these audio and video recordings have been transcribed using ELAN. The utterances by the parents have been orthographically transcribed, with additional labels for features like laughter, onomatopoeia, and disfluency according to the MINGLE annotation guidelines (Nilsson Björkenstam, 2012). Utterances interpreted as exclamations, appeals, or orders are marked with an exclamation mark, and questions with a question mark. Utterances interpreted as adult-directed are labeled as such, while the default is child-directed speech. A subset of this data, named MINGLE-2, has also been annotated with eye gaze, hand gestures, and object-related actions (Nilsson Björkenstam & Wirén, 2012).

MINGLE-4 consists of a total of about 59600 words, with about 24100 words ADS, and 35500 words CDS. Due to the set-up of the experiment these recordings originate from, there is little (or in some cases no) ADS in sessions recorded with older children (>16 months). The CDS word average per session is 1145 (range 565–2305), while the ADS average is 778 (range 0–4203).

Disfluency annotation

In MINGLE-4, the following disfluency categories are annotated: truncated words and phrases, prolongations, hesitations, and filled pauses. Below, examples from both CDS and ADS are presented.

Truncated words (marked by &word):

1) CDS: *ska du **göm&** gömma Kucka i väskan?* (“are you going to **&hi** hide Kucka in the bag?”)
2) ADS: *ja hon brukar det i alla fall när jag ger henne **&bo** tandborsten* (“yes she does at least when I give her the **&br** toothbrush”)

Truncated phrases (marked as &(phrase)):

3) CDS: ***&(här kommer nä)** här kommer nämligen Kucka* (lit. “**&(here comes ac)** here comes actually Kucka”)
4) ADS: ***&(titta kan)** titta förstår hon* (lit. “**&(look knows)** look understands she”)

Prolongations (marked with :):

5) CDS: *kan det vara en ee **ha:j**?* (“can that be a ee **sha:rk**?”)

Hesitations (marked with _):

6) ADS: *ee hon **förstå_r** kom hit* (”ee she **understa_nds** come here”)

Filled pauses (e.g. ee, eh, uu, uh, öö, öh)

7) CDS: *kani& **ee** Kucka måste ha den där* (”the rabbi& **ee** Kucka needs that”)

Note that the primary annotation task was orthographic transcription, not disfluency annotation, and thus our results may underestimate the true disfluency rate.

Categorization of filled pauses

We distinguish between filled pauses in initial, internal, and final position within an utterance, clause, and/or phrase.

Initial: the FP occurs in the beginning of an utterance, e.g.:

8) a. ADS: *ee jaha ee det gör hon ju rätt ofta faktiskt* (“ee yeah ee she does that quite often actually”)
   b. CDS: *ee är du hungrig?* (“ee are you hungry?”)

Internal: the FP is located within a clause, or within a phrase, e.g. a verb phrase (“sees my keys” in 9a), a proper name (“Kucka”, “Ulla” in 10a, b), or in the beginning of (“roosters” in 11a) or within a noun phrase (“her different nicknames” in 11b):

9) a. ADS: *i hissen då får hon ee se mina nycklar och så* (“In the lift then she ee sees my keys”)
   b. CDS: *&(ska vi) ska vi ee hitta namn till allihopa?* (“&(shall we) shall we ee make up names for all of them?”)
10) a. ADS: *men kollar ni alltså på det hon gör nu när ee Ulla och jag pratar* (“but do you look at what she is doing now when ee Ulla and I are talkning”)
   b. CDS: *här kan du få ee Kucka* (“here you can have ee Kucka”)
11) a. ADS: *vi hade ee tuppar också* (“we hade ee roosters as well”)
   b. ADS: *ja hon förstår ju sitt eget namn och hon förstår sina olika ee smeknamn* (“yes she understands her own name and she understands her different ee nicknames”)

Final: the FP marks the end of an utterance:

13) ADS: *igår så tittade hon och hennes pappa på en tavla ee* (“yesterday she and her father looked at a painting ee”)

Data extraction

For this study, we divide all utterances into two categories, adult-directed (AD) or child-directed (CD). We further categorize utterances based on the age of the child, and the gender of the caretaker.

Based on the disfluency annotation described above, disfluent utterances were extracted using the ELAN search tool. The categorization of filled pauses into initial, internal, or final position was performed manually.

Results

Disfluency in child-directed speech

Table 1 shows the disfluency frequency and rate per 100 words in ADS and CDS utterances in MINGLE-4. As shown, there is a difference between ADS (2.60 disfluencies/100 words) and CDS (0.88 disfluencies/100 words). This difference is statistically significant given a Log-Likelihood test (Log-Likelihood value 262.03, $p < 0.0001$).

Table 1. Disfluency frequency, word frequency, and disfluency rate per 100 words in Adult-Directed and Child-Directed speech in MINGLE-4.

	Disfl	Words	Disfl/100 w
AD	627	24109	**2.60**
CD	314	35485	**0.88**

In *Table 2*, the ADS and CDS utterances are divided in two categories based on the age of the child: infants (7–12 months) and one-year olds (13–24 months). *Table 2* shows that the disfluency rate per 100 words for ADS is the same regardless of the age of the child present during recording (Log-Likelihood value 0.04), but that there is a significant increase of disfluency in CDS as the children develop (Log-Likelihood value 18.10, $p < 0.0001$).

Table 2. Disfluency frequency, word frequency, and disfluency rate per 100 words in Adult-Directed and Child-Directed speech categorized by child age.

	Infants (6–12 mnts)			Toddlers (13–24 mnts)		
	Disfl	Words	Disfl/ 100 w	Disfl	Words	Disfl/ 100 w
AD	232	8991	**2.58**	395	15042	**2.63**
CD	64	11057	**0.58**	219	21271	**1.03**

Table 3 shows the disfluency frequency and rate in ADS and CDS utterances to one-year olds categorized by the gender of the caretaker. As *Table 3* shows, there is a difference between male and female speakers in disfluency frequency in ADS (Log-Likelihood value 4.90, $p < 0.05$) but, interestingly, there is no significant difference in CDS (Log-Likelihood value 2.2) between male and female speakers.

Filled pauses in Child-Directed Speech

We find that in our data, the majority of FPs in ADS (70%) occurs in initial position, whereas in CDS, FPs are evenly distributed between initial and utterance-internal position and there are no FPs in final position.

Table 3. Disfluency frequency, word frequency, and disfluency rate per 100 words in Adult-Directed and Child-Directed speech to children age 13–24 months, categorized by the gender of the caretaker.

	Men			Women		
	Disfl	Words	Disfl/ 100 w	Disfl	Words	Disfl/ 100 w
AD-1	211	7130	**2.96**	221	9244	**2.39**
CD-1	134	11880	**1.13**	109	11696	**0.93**

There are only 19 occurrences of FPs in our CDS data, but among these we find patterns of usage for FPs in both initial and internal position, as shown in *Table 4*.

Table 4. Frequency for Filled Pauses in Adult-Directed and Child-Directed speech in MINGLE-4, categorized by the position of the FP.

	Initial (%)	Internal (%)	Final (%)	TOTAL (%)
AD	174 (70%)	36 (14%)	39 (16%)	249 (100%)
CD	11 (58%)	8 (42%)	0	19 (100%)

Out of 11 initial FPs, 6 occur as attention-getting signals, and 5 precede utterance fragments. The initial FPs as attention-getting signals are followed by the child's name (e.g. 14), a question (e.g. 15), or an imperative (e.g. 16).

There are only 19 occurrences of FPs in our CDS data, but among these we find patterns of usage for FPs in both initial and internal position. Out of 11 initial FPs, 6 occur as attention-getting signals, and 5 precede utterance fragments. The initial FPs as attention-getting signals are followed by the child's name (e.g. 14), a question (e.g. 15), or an imperative (e.g. 16).

14) *ee hörru Cornelia* ("ee hey you Cornelia")

15) *oj! ee ska du dricka upp all min mjölk?* "oi! ee are you going to drink all my milk?")

16) *ee öppna munnen!* ("ee open your mouth!")

In the utterance fragments following initial FPs, objects (e.g. 17) or actions (e.g. 18) are named:

17) *ee Kucka*

18) *ee blåsa* ("ee blow")

The internal FPs in our CDS data (8 occurrences) precede unfamiliar or discourse-new objects referred to by names (e.g. 19) or noun phrases (e.g. "a little tanktop" in 20):

19) *kan du mata ee Kucka* ("can you feed ee Kucka")
20) *kan vara ett ee ett litet linne?* ("could be a ee a little tanktop?")

Discussion

There is a significant difference in disfluency rate between ADS and CDS in our data, and further, we find a significant increase in the rate of disfluency coupled with increasing child age when comparing CDS directed at infants (age 6 to 12 months) to CDS directed at one-year olds. These results for Swedish CDS are consistent with previous research on English CDS (Broen, 1972; Kidd, White & Aslin, 2011).

Previous research suggests that FPs commonly precede infrequent or discourse-new words, and may be a result of delay in lexical retrieval (Clark & Fox Tree, 2002). Although there are few occurrences of FPs in CDS, we find clear patterns of usage where FPs in initial position tends to function as attention-getting signals or to precede utterance fragments, while the internal FPs precede discourse-new information. However, since disfluencies such as FPs are infrequent in CDS, further data collection and analysis are needed.

Shriberg (1996) finds that the FP rate in the Switchboard corpus correlates with gender in that men produce significantly higher rates of FPs than women. We find the same pattern in our data, where there is a significant difference in disfluency rate (including FPs) by male and female speakers when talking to the (female) experiment leader, but interestingly there is no difference in disfluency rate between the male and female speakers when talking to children. Previous studies have reported no gender differences in Swedish as regards filled pauses production (e.g. Bell, Eklund & Gustafson, 2000).

Acknowledgements

This research of the first and the second author is part of the project "Modelling the emergence of linguistic structures in early childhood", funded by the Swedish Research Council as 2011-675-86010-31. Thanks to the section for Phonetics, SU, for making this data available for us. Thanks also to our team of transcribers: Anna Ericsson, Joel Ivre, and Johan Sjons.

References

Bell, L., R. Eklund & J. Gustafson. 2000. A Comparison of Disfluency Distribution in a Unimodal and a Multimodal Human–Machine Interface. *Proceedings of ICSLP '00*, Beijing, 16–20 October 2000, 3:626–629.

Broen, P.A. 1972. *The Verbal Environment of the Language-Learning Child*. ASHA Monographs, No. 17. American Speech and Hearing Association: Washington, DC.

Clark, E. 2009. *First Language Acquisition.* 2nd Edition. Cambridge University Press: Cambridge.

Clark, H. & J. E Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111.

Fox Tree, J. E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language* 34:709–738.

Eklund, R. 2010. The Effect of Directed and Open Disambiguation Prompts in Authentic Call Center Data on the Frequency and Distribution of Filled Pauses and Possible Implications for Filled Pause Hypotheses and Data Collection Methodology. *Proceedings of DiSS-LPSS Joint Workshop 2010, The 5th Workshop on Disfluency in Spontaneous Speech and The 2nd International Symposium on Linguistic Patterns in Spontaneous Speech*. University of Tokyo, 25–26 September 2010, Tokyo, Japan, 23–26.

Eklund, R. & M. Wirén. 2010. Effects of open and directed prompts on filled pauses and utterance production. In: *Proceedings of Fonetik 2010*, Lund University, 23–28.

Kidd, C., K. S. White & R. N. Aslin. 2011. Learning the Meaning of "Um": Toddlers' developing use of speech disfluencies as cues to speakers' referential intentions. In I. Arnon & E. V. Clark (eds.), *Experience, Variation, and Generaliza-tion: Learning a First Language (Trends in Language Acquisition Research)*. Amsterdam: John Benjamins, 91–106.

Nilsson Björkenstam, K. 2012. The MINGLE annotation scheme: Multimodal annotation of parent-child interaction in a free play setting (v. 1.0). *Papers from the Institute of Linguistics, University of Stockholm (PILUS)*, ISSN 0348-3223:63.

Nilsson Björkenstam K. & M. Wirén. 2012. Reference to Objects in Longitudinal Parent–Child Interaction. In: *Proceedings of Workshop on Language, Action and Perception (APL)*. Lund, Sweden. [No page numbers.]

Shriberg, E. E. 1996. Disfluencies in SWITCH-BOARD. In: *Proceedings of International Conference on Spoken Language Processing*, Addendum, Philadelphia, PA, 11–14.

Perception of fricative voice distinctions in Greek

Elina Nirgianaki[1], Antonis Botinis[1] & Marios Fourakis[2]
[1]Laboratory of Phonetics and Computational Linguistics, University of Athens, Greece
[2]Department of Communication Sciences and Disorders, University of Wisconsin-Madison, USA

Abstract

This is a study of the perception of the voice distinction in Greek fricatives across different places of articulation. The results indicate that: (1) the acoustic correlate of voicing is not a categorical perception effect of fricative voice distinctions; at least not across all places of articulation, and 2) noise duration has a significant effect on fricative voice perception across all places of articulation, but only in conjunction with glottal vibration.

Introduction

Voice contrasts in Greek fricatives are present at all places of articulation, i.e. labial (labiodental) [f~v], dental [θ~ð], alveolar [s~z], palatal [ç~ʝ] and velar [x~ɣ]. Palatal fricatives have normally no phonemic status but are as a rule context-dependent (Botinis, 2011).

These contrasts affect three main acoustic characteristics. First, voiced fricatives are fully voiced whereas voiceless ones are typically voiceless. Second, voiceless fricatives are considerable longer than voiced ones (see *Table 1*). Third, [voice] differences trigger compensatory durational adjustments on the following vowel: vowel durations are longer after voiced and shorter after voiceless fricatives (Nirgianaki, 2009, 2010, 2013).

In the present study, an experiment was carried out to determine the perceptual correlates of the voicing contrast in Greek fricatives. Two main questions were addressed: (1) does the presence of voicing induce a categorical perceptual effect in distinguishing voiced from voiceless fricatives? And (2), what is the effect of different durational patterns on the perception of the [voice] contrast in these fricatives? Although the acoustics of fricatives (for both place of articulation and [voice]) have been studied extensively (e.g. Baum & Blumstein, 1987; Crystal & House, 1988; Jongman et al., 2000; Maniwa, Jongman & Wade, 2009), their effects on perception have not drawn particular attention. On the other hand, the [voice] contrast in fricatives occupies a central place in the phonetic system of several different language families.

Experimental methodology

Ten (10) female listeners, 20–35 years old, participated to the experiment. All were native speakers of Greek, speaking what is commonly called standard Athenian. None of them had any history of speech or hearing disorders and none had recorded the experimental material. All speakers participated in the experiments as volunteers.

The speech material was recorded by a female native speaker. The six (6) Greek fricative consonants [f], [v], [s], [z], [x] and [ɣ] were recorded in real, two-syllable words of the structure CVCV, stressed on the first syllable. The vowel [a] followed the first fricative consonant. Five repetitions were recorded, from which one was used for the perception experiments. The words were placed in the carrier phrase [ˈipa … ksaˈna] (I said … again). The duration of each fricative and the following vowel were measured with the simultaneous consultation of the waveform and the wideband spectrogram and the mean duration was calculated (*Table 1*).

Table 1. Mean durations (in ms) of each fricative and the following vowel (/a/).

Word	Fricative	Vowel
ˈfata	91	102
ˈvata	69	117
ˈsali	125	139
ˈzali	91	155
ˈxamo	103	116
ˈɣamo	80	140

Then, three extra files were created in Praat (Boersma & Weenink, 2009) for each file that contained a voiced fricative: (a). a file with the voiced fricative prolonged (named ‘Long C’ – each voiced fricative was prolonged by adding to it the mean percentage difference that it had from its voiceless counterpart) (b). a file with the vowel followed the voiced fricative shortened (named ‘Short V’ – the following vowel was shortened by removing the mean percentage difference it had from the one that followed its voiceless counterpart), and

(c). a file with the voiced fricative having no voice bar (named 'No voicing' – for each voiced fricative, the frequencies between 0–500 Hz were filtered out). Four more files were created for each voiced fricative, which resulted from all possible combinations of these three files (i.e. 'Long C + short V', 'Long C + no voicing', 'Short V + no voicing', 'Long C + short V + no voicing').

The stimuli used for the perception experiment were the files that contained a word with a voiced fricative (original, manipulated or a combination of manipulated files). Each such file was heard ten (10) times, yielding a total of eighty (80) stimuli per fricative and 240 stimuli in total.

Three experimental sets were conducted for each participant in a quiet room at the University of Athens. All these experimental sets were two-alternative forced-choice identification tasks, for which clear instructions were provided in Greek. Participants were informed that they would hear the phrase [ˈipa … ksaˈna] "I said … again" containing one of the two words [ˈfata] or [ˈvata] for the first experimental set, [ˈsali] or [ˈzali] for the second experimental set, and [ˈxamo] or [ˈɣamo] for the third experimental set. Participating listeners were asked to identify the word by pressing the appropriate button.

Stimuli were presented to listeners via a program created on Matlab (Mathworks, 2011) on a Hewlett Packard laptop computer over Direct Sound headphones. Following a short familiarization phase of three items, the stimuli were played in random order once each, after pressing a "Play Next" button appearing on the screen.

After each stimulus was played, a prompt appeared on screen containing the two possible response buttons. Participants could press only one of them and after pressing it, the prompt disappeared and they should press the "Play Next" button in order to hear the next item. The two response buttons contained the two Greek words written in Greek orthography, as follows: "φάτα" ([ˈfata]), "βάτα" ([ˈvata]) for the first experimental set, "σάλι" ([ˈsali]), "ζάλη" ([ˈzali]) for the second one, and "χάμω" ([ˈxamo]), "γάμο" ([ˈɣamo]) for the third one. Except for the two buttons corresponding to the two words, a "Replay" button was also included in the prompt, which if pressed, the same phrase was played again.

Listeners could hear each sound three times maximum, by pressing this button two times at most. For each participant, all three experimental sets lasted approximately ¾ hour, including a short break (of around 5 minutes) after each one.

Results

The results of the perception experiment are presented in the following *Table 2* and *Figures 1–2*. Statistical analysis was carried out using the SPSS 19.0 (SPSS Inc., 2010) software package.

It was revealed that voiced fricatives were identified mainly as voiceless, when they were lengthened and their voicing had been eliminated (conditions 'Long C + no voicing' and 'Long C + short V + no voicing'). However, there was a systematic effect of place of articulation on listeners' performance. We used the Generalized Estimating Equations (GEE) analysis, in order to conduct the equivalent of a repeated measures ANOVA. The logistic linking function was used and each of the ten repetitions was treated as a repeated measurement. Place of articulation and condition were within-subjects factors; participant was a between-subjects factor.

Participant had a highly significant main effect (Wald χ^2 (9) = 215360.507, $p < .0001$): the identification of voiceless sounds ranged from 16.3% to 49.6% (*Figure 1*). Condition was also significant (Wald χ^2 (7) = 1982.839, $p < .0001$). The identification of voiceless sounds ranged from 2.7% in the 'Short V' condition to 73% in the 'Long C + no voicing' condition (*Table 2*).

There was a significant main effect of place of articulation (Wald χ^2 (2) = 787.771, $p < .0001$) with higher identification of voiceless sounds in the velar place (velars: 50.2%, alveolars: 31.6%, labiodentals: 14.5%). Individual comparisons against the velars (which showed the highest performance on voiceless identification) showed that in the other two places of articulation voiceless sounds were identified significantly less (labiodentals: Wald χ^2 (1) = 79.683, $p < .0001$; alveolars: Wald χ^2 (1) = 75.290, $p < .0001$).

A place x condition interaction (Wald χ^2 (9) = 209.027, $p < .0001$) was mainly driven by the velars (*Figure 2*), which, contrary to the other two places, were identified as voiceless in all 'no voicing' conditions (i.e. apart from 'Long C + no voicing' and 'Long C + short V + no voicing', also in 'No voicing' and 'Short V + no voicing' conditions).

Table 2. Mean percentage score of voiceless response (%) for each place of articulation and averaged across all three places.

CONDITION	Labials	Alveolars	Velars	Mean
Original file	0	0	9	3
Long C	0	29	20	16.3
Short V	0	1	7	2.7
Long C + short V	0	27	21	16
No voicing	13	17	71	33.7
Short V + no voicing	21	22	78	40.3
Long C + no voicing	42	80	97	73
Long C + short V + no voicing	40	77	99	72

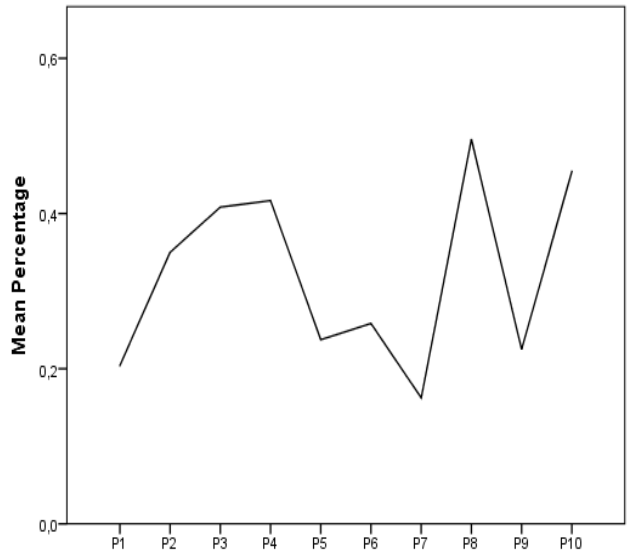


Figure 1. Mean percentage score (%) of voiceless response for each participant.

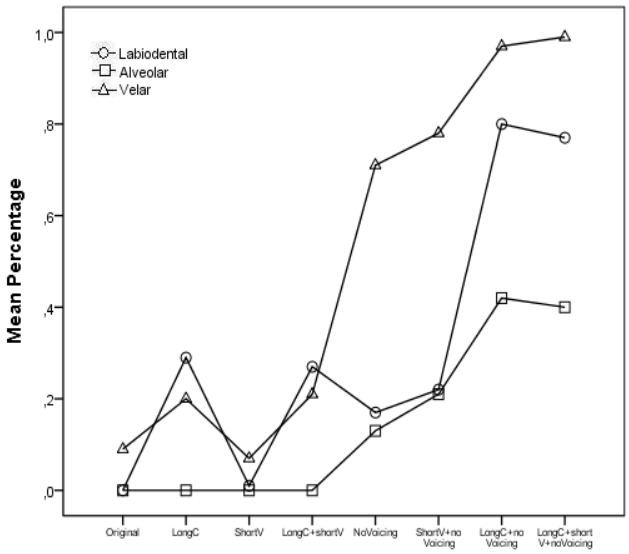


Figure 2. Mean percentage score (%) of voiceless response as a function of place of articulation and condition.

Discussion and conclusions

In accordance with the questions put in the Introduction, the results of this study indicate the following. First, Greek voiced fricatives of all the examined places of articulation were identified mostly as voiceless when they were prolonged to match the duration of their voiceless counterparts and had their glottal vibration removed. Second, voiced velar fricatives were identified as voiceless significantly more than alveolars and labiodentals (velars > alveolars > labiodentals). Third, velar fricatives were also identified as voiceless when they had just their glottal vibration removed.

These results have several implications: 1) the acoustic correlate of voicing is not a categorical perception effect of fricative voice distinctions; at least not across all places of articulation, 2) although the duration of the post-fricative vowel acts compensatory acoustically, it does not account for the perception of the fricative's voicing status, and 3) noise duration has a significant effect on fricative voice perception, but only in conjunction with glottal vibration.

Similarly, the decrease of the frication noise duration of English voiceless fricatives (in syllable-initial position) has been shown to result in an increase of their perception as their voiced counterparts (Cole & Cooper, 1975). Moreover, the fact that there are durational differences between fricatives of the same voicing but different places of articulation (also reported for fricatives with different amplitude by Behrens and Blumstein, 1988) does not seem to play any role, since voiced velars were identified the most as voiceless ones, although shorter than alveolars, and with smaller percentage of lengthening (velars: 28.7%, alveolars: 37.7%, labiodentals: 31.2%).

However, the present study did not examine the effect of F1 (both in terms of onset frequency and change) at the CV boundary for fricatives of different places of articulation, which may have played an important role on listeners' voicing judgements. Results of Stevens et al., (1992) suggest that "listeners base their voicing judgments of intervocalic fricatives on an assessment of the time interval in the fricative during which there is no glottal vibration, and this time interval depends on the extent to which the F1 transitions are truncated at the consonant boundaries".

As stated in the Introduction, [voice] contrasts occur at all places of articulation in Greek. The contrast is a binary specification at the underlying representation level, i.e. ±[voice]. However, in accordance with the results of the present study, the functional binary nature of this contrast is related to several acoustic correlates, in addition to voicing, including, in the first place, duration. Further acoustic correlates include the durations of post-fricative vowels and, presumably, the acoustic structure of the immediate syllabic context in general.

Earlier research indicates that duration is a fairly constant acoustic correlate of voicing distinctions in Greek (Nirgianaki, 2009, 2010, 2013). However, this is evident in simple CV syllabic contexts. In palatalised contexts, which are very common in Greek, duration does not seem to contribute to voicing fricative distinctions. Thus, in words such as [kuˈpça] (oars) and [kuˈbja] (buttons) the duration of the voiceless and voiced fricative after the stop is approximately the same. This surface phonetic representation has been derived from a series of rules that result in a voiceless and voiced fricative production after voiceless and voiced consonant productions respectively (Botinis, 2011). In other words, the fricative voicing distinction in a palatalised context is related to the immediate syllabic context, which carries the respective voicing distinctions.

Thus, the duration patterns corresponding to fricative [voice] contrasts in different contexts is not related to any inherent production specifications but rather to the overall sequence structure. Given that there is no phonemic length distinction in Greek consonants, it is remarkable that durational patterns seem to contribute to a variety of surface phonetic distinctions in the consonant system of Greek. On the other hand, an effective speech communication process has to guarantee that the surface phonetic contrast be effectively conveyed and this results in the association of fricative [voice] distinctions with several acoustic correlates.

Acknowledgements

Thanks to Special Research Account of Athens University for a travel grant to Fonetik 2013 conference.

References

Baum, S. R. & S. E. Blumstein. 1987. Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *Journal of the Acoustical Society of America* 82:1073–1077.

Behrens, S. J. & S. E. Blumstein. 1988. Acoustic characteristics of English voiceless fricatives: A descriptive analysis. *J. Phonet.*16, 295-298.

Boersma, P. & D. Weenink. 2009. Praat: doing phonetics by computer, version 5.1.01. www.praat.org

Botinis, A. 2011. *The Phonetics of Greek* (in Greek, Fonitiki tis Elinikis). Athens: ISEL Editions.

Cole, R.A. & W. E. Cooper. 1975. Perception of voicing in English affricates and fricatives *Journal of the Acoustical Society of America* 58:1280–1287.

Crystal, T. & A. House. 1988. Segmental durations in connected speech signals: Current results. *Journal of the Acoustical Society of America* 83:1553–1573.

Jongman, A., Wayland, R. & S. Wong. 2000. Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America* 108:1252–1263.

Maniwa, K., A. Jongman & T. Wade. 2009. Acoustic characteristics of clearly spoken English fricatives. *Journal of the Acoustical Society of America* 125:3962–3973.

Mathworks, Inc., The. 2000. MATLAB R2011b, The language of technical computing, version 7.13.0.564.

Nirgianaki, E. 2013. *Acoustic and Perceptual Characteristics of Greek Fricative Consonants* (in Greek). PhD Thesis, University of Athens.

Nirgianaki, E., A. Chaida & M. Fourakis. 2010. Acoustic structure of fricative consonants in Greek. *Proceedings ExLing 2010*, Athens, Greece, 125–128.

Nirgianaki, E., A. Chaida & M. Fourakis. 2009. Temporal characteristics of Greek fricatives. *Proc. ICGL*, 25–33.

SPSS Inc. 2010. SPSS for Windows: version 9 2010 SPSS Chicago IL.

Stevens, K. N., S. E. Blumstein, L. Glicksman, M. Burton & K. Kurowski. 1992. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America* 91:2979-3000.

A phonetic pilot study of chirp, chatter, tweet and tweedle in three domestic cats

Susanne Schötz
Centre for Languages and Literature, Lund University, Sweden

Abstract

This study collected 257 vocalisations from three domestic cats when they were watching birds through the window. The sounds were subdivided into the types chatter, chirp, tweet and tweedle, and analysed for duration and F_0. Variation was found within and between these types as well as within and between the three cats in both duration and F_0. A tentative taxonomy of prey-observing cat vocalisations is suggested based on words used for bird sounds.

Introduction

The domestic cat has an extensive and highly variable vocal repertoire, generally divided into three categories: sounds produced with the mouth closed, sounds produced with an opening–closing mouth, and sounds produced with the mouth held tensely open (Moelk, 1944; McKinley, 1982). Many descriptions of domestic cat vocalisations refer to the 16 phonetic patterns suggested by Moelk (1944). However, most of them fail to include the sounds uttered in the vicinity of prey, often referred to as chirp, chatter or other words for bird sounds. The majority of references to such sounds are found in articles about wild cats or on cat-related pages on the Internet, e.g. online dictionaries and articles. Wikipedia (2013) states that "Cats sometimes make chirping or chattering noises when observing prey". Researchers in the Brazilian Amazon have observed wild cats (margays) imitating the calls of tamarin monkeys to lure them within range when hunting. Observations of jaguars and pumas mimicking their prey have also been reported. According to Wildlife Conversation Society (2010) researcher Fabio Rohe "Cats are known for their physical agility, but this vocal manipulation of prey species indicates a psychological cunning that merits further study" (State Journal, 2010; Kelley, 2010).

Chirp and chatter

Chirp and chatter are said to be common vocalisations in some felids and in other small mammals, including the badger, guinea pig and rat. It has been suggested that wild cats are able to copy the calls of their prey, and that this hunting instinct is prevailed in the domestic cat. These sounds are usually elicited when a bird or insect catches the attention of the cat, e.g. by making a sound. The cat becomes riveted to the prey, and will start to chirp, tweet and chatter.

A *chirp* is a voiced short call said to be mimicking a bird or rodent chirp. Stoeger-Horwath and Schwammer (2003) found that juvenile cheetahs produce two distinct cries; chirping and churring, and describe chirping as "a high-intensity call of a bird". Ruiz-Miranda et al. (1998:7) found that chirp were the most common vocalisation in male cheetahs during separations. Acoustic analysis showed high individual variation, with significant differences between individuals in all acoustic measures of fundamental frequency and duration. Feuerstein and Terkel (2008:155) describe chirping as a "sound similar to a high-pitched phone ring, tone often rises near the end", which cats use during play.

Chatter or teeth chattering are very quick stuttering or clicking sounds with the jaws juddering. A Pawsonline (2013) online article argues that chattering is an involuntary, (e.g. at the sight of prey outside the window) enacting of a special type of juddering jaw movement used by the cat to kill its prey while reducing any risk of injury to itself.

Prey-observing cats may also use other types of sounds, which to the author's knowledge have not yet been studied in detail. This study is an attempt to learn more about chirp, chatter and similar sounds. In an earlier study (Schötz, 2012), 18 out of 538 collected cat vocalisations were identified as 'chatter', and an acoustic analysis showed a fairly large variation in F_0 and duration within this type. By recording and analysing a larger number of prey-observing domestic cat vocalisations, the aim is to identify different types and phonetic patterns, and the purpose is to investigate the variation within and between these types.

Words for bird sounds used for cat calls

Dictionaries contain numerous words for various bird sounds, and many of these have been used for also for cat vocalisations. *Table 1*

lists a selection of such words and descriptions, in this case taken from the online dictionary TheFreeDictionary (2013). The words with descriptions that also fitted the different types of prey-related cat vocalisations identified in this study are marked with bold typeface.

Table 1. Words used to describe bird and/or cat sounds (words used in this study for different types of prey-observing cat vocalisations in bold).

Word	Description(s)
chatter	a rapid series of short, inarticulate, speech-like sounds / a rapid rattling or clicking noise by striking together (the teeth)
chirring or churring	a sharp/shrill whirring or trilling/vibrant sound made by some insects and birds, such as the grasshopper and partridge
chirp	a short sharp/high-pitched sound made by small birds and certain insects
cheep	a faint shrill/short weak high-pitched sound like that of a young bird; a chirp
chitter	chiefly US twitter or chirp / twitter or chatter, as a bird
chirrup	a series of chirps / clucking and clicking sounds to urge on a horse
peep	a short, soft, high-pitched sound, like that of a baby bird; cheep / to speak in a hesitant, thin, high-pitched voice
pipe	a birdcall, to chirp or whistle, as a bird does, to utter in a shrill reedy tone
tweet	a weak chirping sound, as of a young or small bird; an imitation or representation of the thin chirping sound made by small or young birds (often reiterated)
twitter	a succession of light chirping or tremulous sounds; chirrup, a light tremulous speech or giggle; titter
tweedle	sing in modulation; play negligently on a musical instrument
warble	sing (a note or song, for example) with trills, runs, quavers or other melodic embellishments
quaver	speak or sing with a trembling voice; (esp. of the voice) to quiver, tremble, or shake

Material and method

Recording procedure

Video and audio recordings of 257 vocalisations were collected from the three domestic cats Donna, Rocky and Turbo (D, R and T; 1 female, 2 males, all 2.5 year old siblings from the same litter) when they were watching birds. The recordings took place in their home between December 2012 and March 2013. A remote-controlled video camera recorder with an electret condenser microphone (either a Sony DCR-PC100E with a Sony ECM-DS70P stereo microphone or a Sony HDR-CX730E with Sony ECM-CG50 shotgun microphone) was positioned on a tripod next to a large window. A variety of bird food was arranged outside (apples and bird seeds on the ground, bird feeders containing peanuts and fat ball nets on strings and poles). When the cats were vocalising at birds through the window, recording could be started from an adjacent room using the remote control without disturbing the cats. Additional recordings were done with an Apple iPhone 3G. *Figure 1* shows the set-up for the video camera with the shotgun microphone.

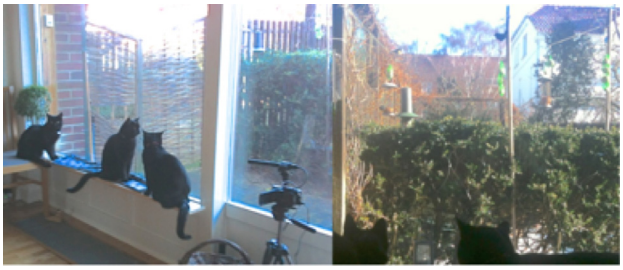


Figure 1. Recording set-up with bird food outside the window and video camera with shotgun microphone.

Preprocessing, segmentation and analysis

Audio files (wav, 44.1 kHz, 16 bit, mono) were extracted with *Extract Movie Soundtrack*. The waveforms were normalised for amplitude and the vocalisations segmented and labelled in *Praat* (Boersma & Weenink, 2013) using the labels chatter, chirp, tweet, and tweedle. Sequences of chirps and tweets were also labelled chirrups and twitter. Measures of duration and F_0 were obtained with a *Praat* script and manually checked.

Results

In this corpus, 97 vocalisation tokens were single sounds, and 49 were phrases of two or more sounds. The most frequent vocalisation type was chirp with 169 tokens. 30 tweedle, 22 chatter and 22 tweet sounds were also recorded. R was the most vocal cat with a total of 119 vocalisations, followed by T: 103 and D: 35 sounds. *Table 2* shows the number of vocalisations for each cat. The results of the acoustic analysis of the four vocalisations types are described below. Median values were very close to mean values, and therefore only mean values are presented here.

Table 2. Number of vocalisations of the three cats in the study by type (CHA = chatter, CHI = chirp, TWE = tweet, TWD = tweedle).

Cat	*CHA*	*CHI*	*TWE*	*TWD*	*Total*
D	3	19	6	6	*35*
R	7	70	19	22	*119*
T	12	80	9	2	*103*
Total	*22*	*169*	*34*	*30*	*257*

Chirp (CHI) and chirrup

A chirp can be described phonetically as a glottal stop [ʔ] followed by a short, often harsh or raspy vowel, e.g. [ə], [e] or [ɛ], produced with an open mouth. Chirps were either single [ʔə] or reiterated [ʔɛʔɛʔɛ...]. Each chirp was fairly short with a mean duration of 0.15 sec. F_0 varied from 229 to 1199 Hz, with a mean F_0 of 661 Hz. Numeric values for chirps are shown in *Table 3*. Sequences of chirps were labelled *chirrups*, and the individual sounds were analysed together with the single chirps.

Table 3. Mean durations, as well as minimum, maximum and mean F_0 of chirp vocalisations (CHI).

Cat	*meanDur*	*minF_0*	*maxF_0*	*meanF_0 (sd)*
D	0.15 s	623 Hz	944 Hz	797 (92) Hz
R	0.14 s	229 Hz	1071 Hz	698 (157) Hz
T	0.17 s	253 Hz	1199 Hz	589 (173) Hz
All	*0.15 s*	*229 Hz*	*1199 Hz*	*661 (194) Hz*

Chatter (CHA)

Chatter sounds were produced with a tensely open mouth, often in sequences. They are phonetically similar to unaspirated voiceless palatal or velar plosives [k̟̄]. The mean duration of individual chatter sounds was 0.03 seconds. As the chatter recorded in this study was mostly voiceless, no measures of F_0 were obtained. Figure 2 shows the waveform, broadband spectrogram and F_0 contour of an example phrase consisting of two sequences of chatter followed by chirps.

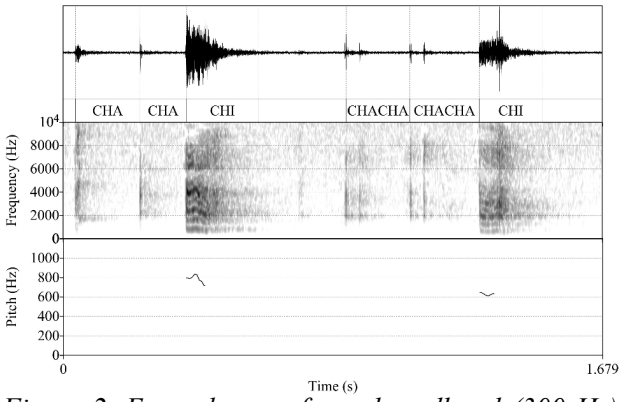


Figure 2. Example waveform, broadband (300 Hz) spectrogram and F_0 contour of two sequences of chatter (CHA) followed by chirps (CHI).

Tweet (TWE) and twitter

Tweets were produced as soft weak chirps, often without any clear initial [ʔ] and with varying vowel qualities, e.g. [wi] or [ɦɛu]. Sequences of tweets were labelled twitter, and the individual sounds analysed together with the single tweets, as they were similar in duration and phonetic quality. *Table 4* shows the numeric values for tweet, which had a mean duration of 0.20 sec. F_0 ranged from 306 to 939 Hz, and the mean F_0 was 635 Hz.

Table 4. Mean durations, as well as minimum, maximum and mean F_0 of tweet (TWE).

Cat	*meanDur*	*minF_0*	*maxF_0*	*meanF_0 (sd)*
D	0.18 s	533 Hz	939 Hz	840 (128) Hz
R	0.19 s	306 Hz	937 Hz	596 (104) Hz
T	0.22 s	514 Hz	648 Hz	586 (26) Hz
All	*0.20 s*	*306 Hz*	*939 Hz*	*635 (132) Hz*

Tweedle (TWD)

A tweedle sounded like a prolonged chirp or tweet, often with some voice modulation, like tremor or quaver, e.g. [ʔəɛəɯə]. The mean duration of tweedle was 0.51 sec, mean F_0 was 578 Hz, ranging from 147 to 936 Hz. *Table 5* shows the duration and F_0 results for tweedle. An example waveform, spectrogram and F_0 contour of a phrase consisting of one tweedle and three tweets is shown in *Figure 3*.

Table 5. Mean durations, as well as minimum, maximum and mean F_0 of tweedle (TWD).

Cat	*meanDur*	*minF_0*	*maxF_0*	*meanF_0 (sd)*
D	0.63 s	528 Hz	936 Hz	820 (61) Hz
R	0.50 s	147 Hz	785 Hz	530 (180) Hz
T	0.29 s	409 Hz	552 Hz	496 (29) Hz
All	*0.51 s*	*147 Hz*	*936 Hz*	*578 (194) Hz*

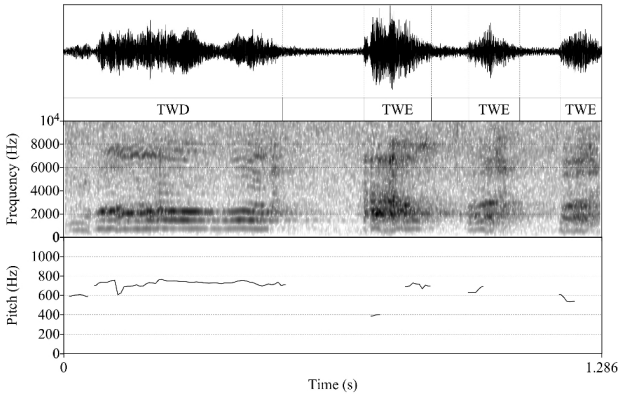


Figure 3. Example waveform, broadband (300 Hz) spectrogram and F_0 contour of a tweedle (TWD) followed by a sequence of three tweets (TWE).

Discussion and future work

The three cats of this study mainly chirped in the vicinity of prey (birds outside the window). Some chirps rose in tone toward the end (cf. Feuerstein & Terkel, 2008), but level or falling F_0 contours were equally common, suggesting that cats are able to vary the intonation of chirp sounds as much as other vocalisation types (see e.g. Schötz, 2012).

Table 6. Tentative taxonomy of prey-observing cat vocalisation types based on words for bird sounds.

feature / **sound**	*voice*	*pitch*	*loudness*	*length*	*rate*	*modulation/reiteration*	*other descriptions or comments*
chatter	unvoiced	-	-	short	rapid	rapid series or rattling	sometimes used for voiced sounds too
chirp	voiced	high	sharp	short	-	-	-
chirrup	voiced	high	sharp	short	-	series	clicking sounds used to urge on a horse
tweet	voiced	high	weak	short	-	often reiterated	-
twitter	voiced	light	light	long	-	succession of tremulous sounds	titter, nervous giggle
tweedle	voiced	-	-	long	-	modulation	-

Several other distinct phonetic patterns were identified within the same behavioural context of this study, which motivated a subdivision into further types. Just like chirp and chatter, words generally used for bird sounds could be used for the additional types as well. This study used the types chatter, tweet (weak chirp), and tweedle (long modulated tweet) for single sounds, and chirrup and twitter for sequences of chirps and tweets. Phrases consisting of two or more types were also not uncommon. Based on the analysis of the recordings of this study, a tentative taxonomy of the sound types and their corresponding features is suggested, as shown in *Table 6*.

Within each type, there was considerable inter- and intra-cat variation in F_0 (except for the voiceless type chatter), and also some variation in duration. This is in line with Ruiz-Miranda et al. (1998), who found significant differences between chirps in cheetahs, with Schötz (2012), where large variation in three other vocalisation types was observed, and also with Moelk (1944:185), who found that the vocal repertoire of the domestic cat is characterised by "an indefinitely wide variation of sound and of patterning".

The results of this pilot study should be regarded as tentative, due to the often limited number of tokens analysed of each type. Future work includes a larger study of similar and also other types of vocalisations collected from a larger number of cats.

Acknowledgements

The author gratefully acknowledges support from the Linnaeus environment Thinking in Time: Cognition, Communication and Learning, financed by the Swedish Research Council, grant no. 349-2007-8695. A warm thanks also goes to the cats participating in this study.

References

Boersma, P., Weenink, D. (2013) Doing phonetics by computer [Computer program]. Version 5.3.23. Retrieved from http://www.praat.org/

Feuerstein, N., J. Terkel. 2008. Interrelationships of Dogs (*Canis familiaris*) and Cats (*Felis catus L.*) Living under the Same Roof. *Applied Animal Behaviour Science* 113, 150–165.

Kelley, J. A. September, 2010. *Scientists Discover New Meaning for Cat "Chattering"*. Retrieved from http://www.catster.com/the-scoop/scientists-discover-new-meaning-for-cat-chattering

McKinley, P. E. 1982. Cluster analysis of the domestic cat's vocal repertoire. Unpublished doctoral dissertation. University of Maryland, College Park.

Moelk, M. 1944. Vocalizing in the House-Cat; A Phonetic and Functional Study. *The American Journal of Psychology* 57(2):184–205.

Pawsonline. 2013. Feline Teeth-chattering. Retrieved from http://www.pawsonline.info/teeth_chattering.htm

Ruiz-Miranda, C. R., S. A. Wells, R. Golden & J. Seidensticker. 1998. Vocalizations and other behavioral responses of male cheetahs (Acinonyx jubatus) during experimental separation and reunion trials. *Zoo Biology* 17:1–16.

Schötz, S. 2012. A phonetic pilot study of vocalisations in three cats. In *Proceedings of Fonetik 2012*, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, 45–48.

State Journal. October, 2010. *Is your kitten a real copy cat?* Retrieved from http://www.statejournal.com/story/13301372/is-your-kitten-a-real-copy-cat?

Stoeger-Horwath, A. S. & H. M. Schwammer. 2003. Vocalizations of Juvenile Cheetahs During Feeding at Schönbrunn Zoo. *International Zoo News* 50(8):468–474.

TheFreeDictionary. 2013. Online dictionary. http://www.thefreedictionary.com

Wildlife Conversation Society. 2010. *Copycat of the Amazon*. Retrieved 7 February 2013 from http://www.wcs.org/news-and-features-main/mimicking-margays-make-monkey-meals.aspx

Wikipedia. 2013. Cat communication. Retrieved 6 February 2013 from http://en.wikipedia.org/wiki/Cat_communication

Functional data analysis of tongue articulation in Gothenburg and Malmöhus Swedish /iː, yː, ʉː/

Susanne Schötz[1], Johan Frid[1], Lars Gustafsson[1] & Anders Löfqvist[2]
[1] Humanities Lab, Centre for Languages & Literature, Lund University, Sweden
[2] Dept. of Logopedics, Phoniatrics and Audiology, Lund University, Sweden

Abstract

Articulatory data collected from nine speakers each of Gothenburg and Malmöhus Swedish were used in a Functional Data Analysis (FDA) to study tongue articulation dynamics, more specifically the height and frontness of the tongue body and tip in the palatal vowels /iː, yː, ʉː/. Standard z-score transformations were used for speaker normalisation. Results showed that the tongue articulation for /iː/ and /yː/ is generally similar, and significantly different from /ʉː/ in both Malmöhus and Gothenburg Swedish. We also found a subdivision of Gothenburg Swedish into two subtypes, where type 1 resembled Malmöhus Swedish more. Significant differences in tongue body height were found between all varieties for all of the vowels, except for /yː/ between Gothenburg type 1 and Malmöhus Swedish.

Introduction

The Swedish vowel system is fairly rich, and Swedish vowels have some particularly unusual and distinctive features. One such feature is that there are three contrastive long front, close vowels /iː, yː, ʉː/, characterised by a relatively small acoustic and perceptual distance. The magnitude of the lip opening is regarded as the major distinctive feature: unrounded /iː/, outrounded /yː/, and inrounded /ʉː/ (Fant, 1959; Ladefoged & Maddieson, 1996). Specifically the contrast between /yː/ and /ʉː/ is considered highly unusual among the world's languages. The tongue articulation is assumed to be basically identical, but the documentation of this is incomplete, especially for the articulatory dynamics (Ladefoged & Maddieson, 1996:295–296).

In many varieties of Swedish, /iː, yː, ʉː/ are also characterised by a slight diphthongisation or consonantal offglide at the end. For /iː/ and /yː/, this is typically made with the tongue dorsum as a [j] sound, while for /ʉː/ the gesture is achieved by the lips approaching each other as a [β] sound (McAllister et al., 1974; Hadding et al., 1976). The different diphthongisations at the end of these vowels contribute to maintaining the distinctions between them. The articulatory dynamics of vowels in Swedish, specifically of palatal vowels, has not been subjected to any systematic phonetic production study. Spectral changes have been claimed to be more important for vowel perception than static cues, see e.g. (Nearey, 1989; Strange, 1989).

Another rare feature is the nowadays fairly wide-spread realisation of /iː/ and /yː/ in Swedish with Viby-colouring, i.e. with a "damped" quality /i̞ː/ and /ʉ̞ː/ (Ladefoged & Maddiesson, 1996; Bruce, 2010). There is disagreement in the Swedish phonetics literature if the major constriction for the damped /iː/ and /yː/ is further front compared to their regular counterparts, and basically alveolar, or instead further back and rather central (Björsten et al, 1999; Engstrand et al., 2000). However, as adequate articulatory data seem to be lacking, these views are at best intelligent speculations.

The purpose of this study was to use Functional Data Analysis to examine tongue articulation of Swedish vowels. We focus on the vowels /iː, yː, ʉː/ in two regional varieties of Swedish; Gothenburg Swedish (GS) and Malmöhus Swedish (MS), spoken in and near Gothenburgh and Malmö, respectively. The aim was to find out if the tongue positions are similar for these vowels as previously assumed, and if there are any regional differences. An additional aim was to learn more about the articulatory dynamics of palatal vowels in Swedish. We expected the tongue positions in the dimensions open–close and front–back to be similar in for /iː/, /yː/ and /ʉː/ in both GS and MS. Furthermore, we expected to find regional differences in the articulation of /iː/ and /yː/, as Viby-colouring is more common in GS than in MS (Bruce, 2010).

Material and method

Nine speakers of GS (5 females, 4 males, 20–47 years) and 9 speakers of MS (4 females and 5 males, 23–62 years) were recorded by means of electromagnetic articulography (Carstens AG 500). Twelve sensors were placed on the

lips, jaw and tongue, and also on the nose ridge and behind the ear to correct for head movements. *Figure 1* shows the sensor positions and one subject with sensors attached. In this study, our focus was on the tongue tip and body (sensors 1 and 2). The speech material consisted of 20 repetitions from each speaker of /iː, yː, ʉː/ in carrier sentences *De va inte hVt utan hVt ja sa* (It was not hVt, but hVt I said), where the target words were stressed. The sentences were displayed on a computer screen one at a time in random order, and the speakers were instructed to read them in their own dialect at a comfortable speech rate.

Figure 1: The twelve sensor positions recorded, and one speaker with the sensors attached.

Error detection and speaker normalisation

Noise and measurement errors in articulatory data may come from a) quick movements by the speaker, b) sensors moving too close to each other, c) sensors breaking or falling off, and d) calculation errors. In order to detect and exclude such errors, we used a two-step process. All /iː, yː, ʉː/ vowels were segmented manually in Praat (Boersma & Weenink, 2013) and used as acoustic landmarks to trim the data set. Plots for sensors traces 1–3 were used to visually identify and exclude vowels with errors. The remaining errors and outliers were removed with the package 'robustbase' (Rousseeuw et al., 2012) in the R statistical environment (R Development Core Team, 2013), using a method that calculates location (mu) and scale (tau) from articulatory data using robust methods (Maronna & Zamar, 2002). In our case, all the position data in all repetitions of each vowel /iː, yː, ʉː/, each of the sensors (1–3), and each spatial dimension (x, y, z), for each speaker were used to calculate the mean value of all the individual repetitions of each vowel. If the mean value of a repetition was above or below mu +/- tau, it was marked as an outlier and excluded.

In order to compensate for differences in oral anatomy between speakers, data was normalized using *z*-score transformation.

FDA smoothing and aligning

Functional Data Analysis (FDA) is a technique for timewarping and aligning a set of signals to examine differences between them. FDA techniques and applications to speech analysis were first introduced by Ramsay et al. (1996), and further developed by Lucero et al. (1997), Lucero and Löfqvist (2005) and Gubian et al. (2011). In FDA, a function or function system is fitted to the data, and the fitting coefficients are examined instead of the original data. A commonly used function form are B-spline functions (Ramsey et al. 2009), which are flexible building blocks for fitting curves to approximate a large number of different shapes. In essence, spline functions are placed at overlapping, equidistant intervals throughout a sensor trace. By selecting weights for each spline, the overall shape becomes similar to the actual sensor trace. The degree of similarity may be controlled so that it does not overfit. It is possible to select: a) the number of spline functions ('knots'), b) the order (how well higher-order derivatives are preserved) and c) the amount of roughness ('lambda'). In this study, FDA was used to smooth the sensor traces, and to standardise the time to facilitate comparisons between repetitions. All FDA processing was done using the R package 'fda' and the following parameters for creating the B-spline basis: knots=20, order=6, lambda=$1e^{-2}$.

Analysis of tongue height and frontness

Sensors 1 and 2 were selected to represent the tongue tip and body (see *Figure 1*). We plotted the FDA processed contours for the tongue dynamics in height and frontness for the tongue body and tongue tip, and compared the positions and dynamics within each regional group as well as across the two regional varieties. Statistical analysis was done with functional *t*-tests, an extension of the classical *t*-test where the *t*-statistic is a function of time, using the function *tperm.fd* in the 'fda' package. Functional *t*-tests are described in detail in Ramsey et al. (2009).

Results

Tongue body height

Within each variety, the contours for /ʉː/ are often clearly separated from /iː/ and /yː/, which in turn often overlap, and significant differences in tongue body height (*Figure 2*, column 1) were found between /ʉː/ and /iː, yː/ (pairwise functional *t*-tests, $p < 0.05$).

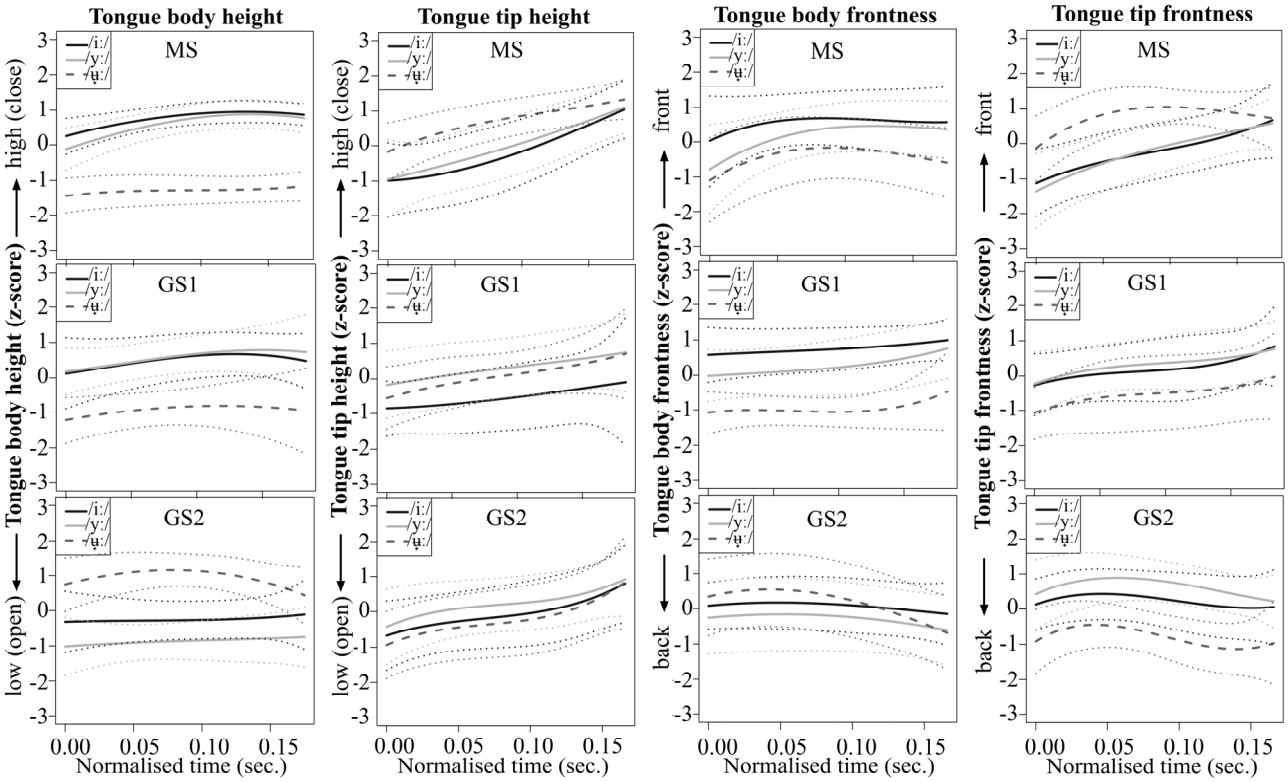


Figure 2. Tongue articulation (z-scores of tongue body and tongue tip height as well as frontness) as a function of normalised time for the vowels /iː, yː, ʉː/ in Malmöhus Swedish (MS) and two types of Gothenburg Swedish (GS1 and GS2); mean values for each variety (dotted lines: standard deviation).

The GS speakers generally displayed more variation than the MS speakers in all tongue positions and all vowels. Among the GS speakers we found a subdivision between four speakers (type GS1) who articulated the vowels with similar tongue positions as the MS speakers, and five speakers (type GS2) who generally had a different tongue positions than slightly higher tongue body for /iː/ than for /yː/. The dynamics, represented by the mean contour shapes, are fairly level for /iː/ and /yː/ in GS1 and GS2, but there seems to be some individual variation. In MS, the contours for /iː/ and /yː/ are slightly rising, suggesting a mild closing diphthongisation. /ʉː/ is relatively level in MS and GS1, but is somewhat arch-shaped in GS2.

Tongue tip height

The tongue tip height contours for /iː/ and /yː/ in MS are similar and somewhat lower than for /ʉː/ (*Figure 2*, column 2). In GS1, /iː/ seems to be produced with a lower tongue tip than /yː/ and /ʉː/, while GS2 has similar contours for all three vowels. The dynamics for all the vowels in all three varieties is represented by slightly rising contours, suggesting closing diphthongisations, although some individual variation was observed.

Tongue body frontness

While the tongue body in MS and GS1 is more front for /iː/ and /yː/ than for /ʉː/ (*Figure 2*, column 3), the opposite pattern is shown in GS2, except towards the final part of the vowels, where there is more variation. In addition, the tongue body seems to be slightly more front for /iː/ than for /yː/ in all varieties. All MS vowel contours rise initially, indicating a forward motion. /iː/ and /yː/ are fairly level in GS1 and GS2, while the tongue body seems to move slightly forward (GS1) or backward (GS2) in the final part of /ʉː/.

Tongue tip frontness

In MS the tongue tip is further back in /iː/ and /yː/ compared to /ʉː/, while the opposite pattern is found for GS1 and GS2 (*Figure 2*, column 4). The contours for /iː/ and /yː/ are similar and overlapping in MS and GS1, while /yː/ tends to be a bit more front than /iː/ in GS2. In MS the /iː/ and /yː/ contours are rising, indicating height-harmonic diphthongisations towards more peripheral vowels, while the GS1 contours are moving slightly forward. The GS2 contours are slightly arch-shaped.

Discussion and future work

The results of this study indicate that the tongue articulation for /ʉː/ is significantly different from /iː/ and /yː/ in both MS and GS. Our hypothesis of similar tongue articulation for the

three vowels was thus rejected. In addition, we found more intra-regional variation in GS than in MS, which led to the subdivision into the two types GS1 and GS2. A closer look showed that the GS1 speakers were more often from the outskirts of the Gothenburg area than the GS2 speakers. Furthermore, most GS2 speakers had clear Viby-coloured /iː/ and /yː/, which was not the case for all GS1 speakers. No MS speakers used Viby-colouring. The Viby-colouring may offer one explanation for the differences in tongue articulation. However, a few GS1 speakers did use some kind of Viby-colouring, and we need to investigate further how the speakers articulated both general and Viby-coloured vowels. We will also compare this data to acoustic data, e.g. formant frequencies.

Considerable regional variation was found in this study, not only for each vowel in the front–back and open–close dimensions, but also in the vowel dynamics (diphthongisation). Our hypothesis of different articulation strategies in different regional varieties was thus supported.

In this study we analysed only two discrete points and two dimensions of the tongue: tongue tip and body height and frontness, and used a standard *z*-score transformation for speaker normalisation. Although we did not look at lip rounding, traditionally regarded as the main difference between /iː/, /yː/ and /ʉː/, our results clearly show differences between these vowels in tongue body height as well. In future studies, tongue body height will be compared to other tongue articulation dimensions as well as to lip rounding. Moreover, we will include other palatal vowels, and compare tongue articulation in MS and GS to that of Stockholm Swedish. We will also investigate more sophisticated speaker normalisation methods.

Acknowledgements

This study was carried out within the project Exotic Vowels in Swedish: an articulographic study of palatal vowels. The authors also gratefully acknowledge support from the Linnaeus environment Thinking in Time: Cognition, Communication and Learning, financed by the Swedish Research Council, grant no. 349-2007-8695.

References

Björsten, S., G. Bruce, C-C Elert, O. Engstrand, A. Eriksson, E. Strangert & P. Wretling. 1999. Svensk dialektologi och fonetik – tjänster och gentjänster. *Svenska landsmål och svenskt folkliv*, 7–23.

Boersma, P. & D. Weenink. 2013. *Doing phonetics by computer*. Retrieved from http://www.praat.org/

Bruce, G. 2010. *Vår fonetiska geografi*. Lund: Studentlitteratur.

Fant, G. 1959. Acoustic description and classification of phonetic units. *Ericsson Technics* 1 (Reprinted 1983 in G. Fant, *Speech Sounds and Features*, 32–83. Cambridge, MA: MIT Press)

Engstrand, O., S. Björsten, B. Lindblom, G. Bruce & A. Eriksson. 2000. Hur udda är Viby-i? Experimentella och typologiska observationer. *Folkmålsstudier* 39:83–95. Helsingfors.

Gubian, G., F. Cangemi & L. Boves. 2011. Joint analysis of F0 and speech rate with functional data analysis. *ICASSP*, 4972–4975, Prague.

Hadding, K., H. Hirose & K. S. Harris. 1976. Facial muscle activity in the production of Swedish vowels: An electromyographic study. *Journal of Phonetics* 4:233–245.

Ladefoged, P. & I. Maddieson. 1996. *The Sounds of the World's Languages*. Oxford: Blackwells.

Lucero, J., K. Munhall, V. Gracco & J. Ramsay. 1997. On the registration of time and the patterning of speech movements, *Journal of Speech, Language and Hearing Research* 40:1111–1117.

Lucero, J. & A. Löfqvist. 2005. Measures of articulatory variability in VCV sequence, *Acoust. Res. Lett. Online* 6:80–84.

Maronna,R. A. & R. H. Zamar. 2002. Robust estimates of location and dispersion of high-dimensional datasets, *Technometrics* 44(4):307–317.

McAllister, R., J. Lubker & B. Lindblom. 1974. An EMG study of some characteristics of the Swedish rounded vowels. *Journal of Phonetics* 2:267–278.

Nearey, T. 1989. Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America* 85: 2088–2113.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*, URL: http://www.R-project.org.

Ramsay, J., G. Hooker & S. Graves. 2009. *Functional Data Analysis with R and MATLAB*. Springer.

Ramsay, J. O., K. G. Munhall, V. L. Gracco & D. J. Ostry. 1996. Functional data analysis of lip motion, *Journal of the Acoustical Society of America* 99:3718–3727.

Rousseeuw, P., C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, M. Maechler. 2012. *robustbase: Basic Robust Statistics*. R package version 0.9-7. URL: http://CRAN.R-project.org/package=robustbase.

Strange, W. 1989. Evolving theories of vowel perception. *Journal of the Acoustical Society of America* 85:2081–2087.

Interviewing Swedes about phonetic transcriptions

Michaël Stenberg
Centre for Languages and Literature (Phonetics), Lund University, Lund, Sweden

Abstract

The present paper reports on an ongoing series of in-depth interviews with Swedes about phonetic transcriptions. These interviews form an important part of a doctoral thesis on phonetic transcriptions (e.g. IPA) in encyclopaedias in Swedish language. For context and comparison reasons, the thesis also treats other publications featuring phonetic transcriptions, such as textbooks, dictionaries and wine catalogues. People who professionally speak in public are interviewed in detail about their behaviour when seeking information on pronunciation of difficult items, typically proper names or words of foreign origin. They are also asked to make assessments of a number of existing transcription systems with respect to rendering of stress, segments and prosodic features. Among the interviewees are lecturers, actors and broadcasters. Preliminary results show that users of encyclopaedias rarely consult the pronunciation key when trying to interpret transcriptions. As to the issue of whether words from all languages should be transcribed with equal accuracy, in the first instance the interviewees are apt to be in favour of that proposal, but often modify their standpoint when pondering the matter. Furthermore, for encyclopaedia entries regarding human beings, a vast majority want phonetic transcriptions to be supplied not only for surnames—as is the conventional usage in Sweden—but also for first names; these have come more and more into use, and not necessarily for disambiguation reasons.

Introduction

You may write an entire book, be it a novel or a doctoral dissertation, without bothering about the pronunciation of any word in it. But someone reading a novel aloud, or presenting a paper at a conference will have to make a choice—consciously or unconsciously—of how to pronounce every detail in it. Likewise, a playwright or a poet, already when conceiving a work, will have to consider what the final result will sound like when performed.

In Sweden, encyclopaedias form an important source of information on pronunciation, particularly of names and words of foreign origin. Adequate pronunciation of these is considered a mark of prestige. There is a long-standing tradition of encyclopaedias to provide users with phonetic transcriptions, at least of entry headwords. In addition, from the late 19th century (Lyttkens & Wulff, 1889) until the 1990s (Hedelin, 1997), no dedicated pronunciation dictionaries for Swedish were published. Hitherto, there has been a lack of knowledge about users' behaviour when seeking pronunciation information. Likewise, little is known about their attitudes and preferences regarding such information, be it in the shape of phonetic transcriptions or clickable audio pronunciations. A pilot study, a quantitative survey, presented by me at *Fonetik 2009* (Stockholm University), yielded shallow results. I decided that applying a qualitative method would probably prove more fruitful. The present paper reports on an ongoing series of in-depth interviews with professionals likely to request reliable information about pronunciation in order to fulfil their tasks in the best way possible. The interviews make up an important part of my dissertation work entitled *Phonetic transcriptions in encyclopaedias in Swedish.*

Background

Swedish encyclopaedias have been using phonetic transcriptions for more than 130 years. Still, little is known about the reception of this pronunciation advice among users. Also, editorial boards seem to start from scratch when preparing for a new work, each of them thus inventing a transcription system of their own, without much regard to previous or surrounding ones. The resulting manifold systems may be confusing for the reference library user, even if IPA-based transcriptions have become increasingly used, and dominate since the 1970s.

Method

One of the main objectives of my dissertation work is to investigate how users of

encyclopaedias act when in need of pronunciation advice, to find out their wishes about it and to study their attitudes to various existing transcription systems. Carrying out a survey based on a random sample of 1000 individuals and then treating the data statistically would make no sense. Therefore, I decided to conduct a series of in-depth interviews of about 45 minutes each with around 20 persons who in their profession are carefully listened to. Consequently, they ought to care about their pronunciation and seek reliable information when in doubt. My interview method is based on the guidelines in the Norwegian–Danish handbook *Den kvalitativa forskningsintervjun* (Swedish translation) (Kvale & Brinkmann, 2009).

Subjects and issues

The interviewees are recruited from the following groups of professionals:

(a) broadcast media journalists;
(b) university lecturers;
(c) commercial event lecturers;
(d) librarians in touch with customers;
(e) film and theatre directors and actors, including radio theatre;
(f) musical drama directors and singers;
(g) narrators of radio novels or audiobooks.

Initially, the interviewees are asked to tell something about their background: their knowledge of foreign languages, studies in phonetics, if any, etc. Then an informal conversation follows, where topics are discussed in an associative order. The idealized questions I want to get answered are of three types:

Issues on attitudes

- How important do you think it is to pronounce according to current norms?
- To what degree do you trust the data on pronunciation you have found?
- What is your opinion on the phonetic transcriptions you have met?
- What do you think of clickable sound recordings of phonetic transcriptions?
- Could such recordings be substituted for phonetic transcriptions?
- In what way could phonetic transcriptions be improved?
- How should stress advice be marked in normal orthography (e.g. in entry headwords)? (Printed examples are shown.)
- What kind of phonetic transcription (marking of stress included) do you prefer for some given words? (Printed examples are shown.)

Issues on use of pronunciation advice

- What are your reasons for seeking pronunciation advice?
- Where do you turn to do so?
- In what concrete situation (where and when) do you do so?
- Exactly how do you proceed when searching for information about pronunciation?
- How do you memorize the pronunciation you have found?
- Do you use the keys, i.e. the explanations of the signs used?

Issues on principles

- Should information about pronunciation in encyclopaedias be *descriptive, normative* or *guiding?*
- How should the meaning of signs be explained (specimen words from various languages, verbal articulation descriptions, images illustrating articulation)?
- For human beings, is it sufficient to supply pronunciation of surnames?
- On what variety of Swedish should the pronunciation information be based? (How to render <rs>, <rd> etc.?)
- Ought pronunciation of all languages to be treated with the same accuracy? If not, what criteria should decide?
- Should more than one pronunciation variety (depending on geographical origin) be used for certain languages (English, Spanish, Portuguese, etc.)?
- Should features like Swedish and Norwegian *accent 1* and *2,* or Danish *stød* be rendered?
- Should tones in for example Chinese be rendered?

Procedure

The interviews are recorded on a Roland Edirol R-09HR solid state audio recorder with two built-in omnidirectional electret microphones. The recording mode applied is MP3, 128 kbps at 44.1 kHz sampling frequency (allowing a recording time of 490 min on an SD card of 512 MB capacity). When finished, the recordings are transferred to a PC via High-speed USB 2.0 for storage and playback in Windows Media Player.

In order to facilitate for interviewees to compare and evaluate various transcription systems, I use a display book with transparent pockets, in which I have put white sheets of paper in DIN A4 size, so that they can be shown one by one at a reading distance of about 50 cm.

On one sheet, five different systems for marking stress in normal orthography are shown, either as entry headwords or alternative varieties of those, such as scientific names of species. Each of the systems is represented by three tokens of varying length. Typeface is Times New Roman boldface in 24 point.

The remaining sheets contain examples of several kinds of phonetic transcriptions to be evaluated by the interviewees, in a process similar to that of determining a prescription for eyeglasses. Gentium Plus normal in 36 point is used as default typeface.

The examples feature transcription systems from the following works:

- Bonniers Lexikon (Swedish encyclopaedia)
- Bra Böckers Lexikon (Swedish encyclopaedia)
- Den Store Danske Encyklopædi (Danish encyclopaedia)
- Duden Aussprachewörterbuch (German pronunciation dictionary)
- Nationalencyklopedin (Swedish encyclopaedia)
- Nordisk Familjebok, 2nd edn. (Swedish encyclopaedia)
- Nordisk Familjebok, 4th edn. (Swedish encyclopaedia)
- Respons (Swedish encyclopaedia)
- Svensk uppslagsbok, 2nd edn. (Swedish encyclopaedia) and Svenska Akademiens ordlista, 13th edn. (Dictionary of the Swedish Academy)
- Wells, Longman Pronunciation Dictionary, 3rd edn.
- Wikipedia German-language edition.

Processing of interviews

Shortly after each interview, I write down a résumé to be completed when listening to the recording. I also make notes about the attitude of the interviewee (interested, careful, energetic, etc.). Later, when listening to the recording in detail, I transcribe important passages (providing answers to my idealized questions or containing remarkable or unexpected comments). The highlights of the text thus created constitute the base of my further writing.

Results

Preliminary results show that encyclopaedia users seldom consult the pronunciation key before trying to interpret the phonetic transcriptions. When asked if transcriptions of words from all languages of the world should be equally accurate, their spontaneous answer tend to be yes, but when pondering the matter and becoming aware of disregarded problems, they often change their opinion. For encyclopaedia entries regarding human beings, there is an almost unanimous agreement that phonetic transcriptions should be supplied not only for surnames—as is the conventional usage in Sweden—but also for first names; these have come more and more into use, and not necessarily for disambiguation reasons.

Further research

Much work is left to be done; about half of the planned interviews are pending. When complete, their outcome will hopefully make it possible to give recommendations for designing phonetic transcription systems suitable for different kinds of encyclopaedias and other reference books or corresponding internet sites.

Acknowledgements

Best thanks to all the anonymous interviewees who until now have spent their time and energy on being interviewed.

References and bibliography

Bonniers Lexikon. 1961–67. Stockholm: Nordiska uppslagsböcker.

Bra Böckers Lexikon 1973–97 (1st–4th edns.). Höganäs, Sweden: Bra Böcker.

Den Store Danske Encyklopædi. 1994–2001. Copenhagen: Danmarks Nationalleksikon.

Duden Aussprachewörterbuch. 2000 (4th revised & updated edn.). Mannheim &c.: Dudenverlag.

Garlén, C. 2003. *Svenska språknämndens uttalsordbok: 67 000 ord i svenskan och deras uttal.* Stockholm: Svenska språknämnden: Norstedts ordbok.

Hedelin, P. 1997. *Norstedts svenska uttalsordbok.* Stockholm: Norstedts.

Kvale, S. & S. Brinkmann 2009 (2nd edn.). *Den kvalitativa forskningsintervjun.* Lund: Studentlitteratur.

Lyttkens, I. A. & F. A. Wulff. 1889. *Svensk uttalsordbok.* Lund: C. W. K. Gleerup.

Nationalencyklopedin. 1989–96. Höganäs, Sweden: Bra Böcker.

Nordisk Familjebok. 1875–99 (1st edn.). Stockholm.

Nordisk Familjebok. 1904–26 (2nd edn.). Stockholm.

Nordisk Familjebok. 1951–55 (4th edn.). Malmö, Sweden: Förlagshuset Norden.

Respons. Its four alphabetical volumes. 1997–98. Malmö, Sweden: Bertmarks.

Svenska Akademiens ordlista. 2006 (13th edn.). Stockholm: Norstedts akademiska förlag (distributor). Also available free of charge on: http://www.svenskaakademien.se/svenska_sprakет/svenska_akademiens_ordlista/saol_pa_natet/ordlista

Svensk uppslagsbok 1947–55 (2nd edn.). Malmö, Sweden: Förlagshuset Norden.

Wells, J. C. 2008 (3rd edn.). *Longman Pronunciation Dictionary.* Harlow, UK: Pearson Education.

Wikipedia German-language edition. A choice of narrow IPA transcriptions of Swedish proper names (e.g. Hjalmar Bergman; Ingmar Bergman; Ingrid Bergman; Dag Hammarskjöld; G. H. von Koch; A. von Krusenstjerna; M. Zetterling). http://de.wikipedia.org/wiki/Kategorie:Schwede

Constant tonal alignment in Swedish word accent II

Malin Svensson
Linguistics and Phonetics, Lund University, Lund, Sweden

Abstract

Studies on accentual tonal alignment of intonation languages suggest that L in rising (LH) pre-nuclear accents anchors with a specific point in the segmental string, while the timing of H varies. This study investigates if lexical accents, too, exhibit a constant alignment by testing the South Swedish word Accent II. When under the strain of tempo variability the L-target was found not to be anchored with syllable onset. The results were not fully conclusive regarding H, but no clear evidence was found against anchoring of H, which could mean that H is an important phonological event in Accent II, while L is not.

Introduction

Over a period of 13 years there has been an on-and-off debate within intonational phonology on whether accentual tonal targets (L, H) are constantly aligned with the segmental string. Most studies have focused on pre-nuclear rising accents in intonation languages. A language with lexical accents has not been taken into account in the recent research on constant alignment so far.

Tonal alignment

Tonal alignment might be seen upon as a wider notion for other concepts such as timing, tonal association or segmental anchoring. *The segmental anchoring principle* presupposes that tonal targets are constantly aligned, and thus anchored at specific points in the segmental string. Studies that second the principle have looked at rising pre-nuclear accents in intonation languages (Arviniti et al., 1998; Atterer & Ladd, 2004; Ladd et al. 1999).

Previous studies in the field have displayed an unambiguous case of the tonal target L aligning with syllable onset in pre-nuclear accents, though the precise timing seems to vary across languages. Results include the L-target occurring just before the onset of the accented syllable (Arvaniti et al., 1998, for Greek; Niemann et al., 2011, for Italian), at syllable onset (Caspers & Van Heuven, 1993, for Dutch; Ladd et al., 1999, for English; Atterer & Ladd, 2004, for German), or after syllable onset (Xu, 1998, for Mandarin).

The precise timing seems to vary, but the studies do show anchoring of L with the beginning of the syllable, while the same consistent result does not exist for the H target. Some studies found H anchoring after syllable offset (Atterer & Ladd, 2004; Xu, 1998; Arvaniti et al., 1998), or somewhere late in the syllable (Ladd et al., 1999). Caspers and Van Heuven (1993) found that the end of the rise, the H-target, varied considerably under time pressure and thus discarded that it did anchor with the segment.

Swedish accents

In the prosodic typology of Swedish intonation provided by the Lund Model (Bruce & Gårding, 1978; Bruce, 2007), the two Swedish word accents are assumed to be represented by a fall in the stressed syllable in a prosodic word, where the two accents differ in the timing of the fall. There is a variation between the Swedish dialects. For example in the South Swedish dialect (South) both accents are timed considerably later than in the Central Swedish dialect (Svea): in South Swedish the high level in Accent I is associated with the stressed vowel and in Accent II with offset of the stressed syllable.

The original dialect typology has later been revised by Bruce (2007), who identified, for all dialects, an LHL tonal gesture from which bitonal gestures are extracted; either a fall, H+L, or a rise, L+H. Bruce generalized for Accent II an association of a fall in the dialects with an early timing of the accents (Svea and Göta) and a rise in the dialects with a late timing (South, Gotland, Dala, North). For the South Swedish dialect, a late timed dialect type, Bruce made the specific assumption of a fall, an H+L pattern, for Accent I and a rise, an L+H pattern, for Accent II. The rise in South Swedish Accent II has indeed been shown to be relevant from a perceptual point of view (Ambrazaitis & Bruce, 2006).

As if by chance, there is a rough phonetic match between the timing of the lexical Accent II in South Swedish and the pre-nuclear accents in the already mentioned studies on tonal alignment in intonation languages. The research question formulated here is whether or not

additional phonetic features are similar such as if L and H are anchored with a segment, as is assumed by the segmental anchoring principle. The present study is a production study where the hypothesis of segmental anchoring is tested on the Swedish word Accent II. Speech rate is used as an experimental tool and is based on the idea that speakers will try to retain primary features of phonological properties, while they will let other features be modified under time pressure (Caspers & Van Heuven, 1993). Speech rate has been used successfully by a number of researchers in studies concerning tonal alignment (Caspers & Van Heuven, 1993; Ladd et al., 1999; Xu, 1998). Because the Lund Model (revised by Bruce, 2007) assumes that both L and H in the rising L+H gesture of Accent II are phonologically relevant, anchoring of L and H is expected.

Method

Speakers and recording

The material was initially recorded for a different study in which two age groups were recorded. For this study only the older speakers were tested due to technical issues. There were seven speakers, four males and three females, and the average age of the speakers was 72 years. All speakers were voluntary and spoke the same variety of the South Swedish dialect. A criterion for speaker selection was that they had all lived most of their lives in the same area in the northeastern part of the South Swedish region. Moreover, their parents also had to have lived most of their lives in the area.

All of the recordings were made in people's homes. An IMG Stage boundary microphone (table-microphone) with phantom power was used (ECM-302B) since it is non-invasive and the speakers were expected to be naïve with no prior recording experiences.

The material was read twice by each speaker at three different speech rates: normal, slow and fast. The recording leader set the pace of the speech rate with the leading question and the speaker was asked to answer the question and to follow the speech rate of the recording leader.

Speech materials and data processing

The materials consisted of three test sentences with the same test word. The materials were mixed with 37 further sentences not investigated here. The three test words fit the following criteria: an unbroken tonal curve, a word Accent II, identical segmental surroundings ([9 syllables] bisyllabic target word [2 syllables]) and that neither syllable and vowel onset, nor syllable and vowel offset coincided. The target word *många* [ˈmɔŋːa] 'many' occurred before nuclear accent in all three sentences.

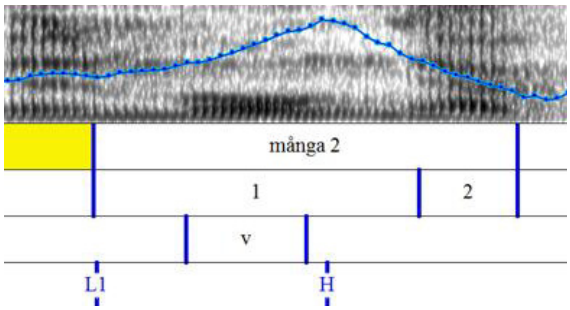


Figure 1. Image of praat window. Target word in sentence 2 Nej, Anna bilar till Polen med många VÄNNER *by speaker M63 in normal speech rate.*

The author performed segmentation and annotation in Praat (Boersma & Weenink, 2010). Since each speaker was recorded twice, the material consisted of 126 items (3 sentences × 2 repetitions × 3 speech rates × 7 speakers). Each target word was segmented into syllables, and in addition, the boundaries of the accented vowel were determined (*Figure 1*). Since the border between the two syllables was difficult to distinguish, offset of accented syllable was calculated to occur in between offset of accented vowel and onset of the following vowel. The tonal curve has been semi-automatically annotated for the tonal targets L and H. Extracted measures were the start and the end of the rise (L and H), syllable onset and syllable offset.

Results

Average syllable duration show a difference in speech rate between the recordings (*Figure 2*). An ANOVA confirmed that speech rate had a significant effect on syllable duration ($F = 66.490$, df 1, $p = .000$), concluding that the rate manipulation was successful.

Since the segments are affected by speech rate, the temporal distance between the tonal targets L and H should also be affected by speech rate, if they are anchored in the segmental string. An ANOVA showed a significant effect of speech rate on the temporal distance between L and H (henceforth, rise time) ($F = 17.129$, df 1, $p = .006$) resulting in shorter rise times for faster speech.

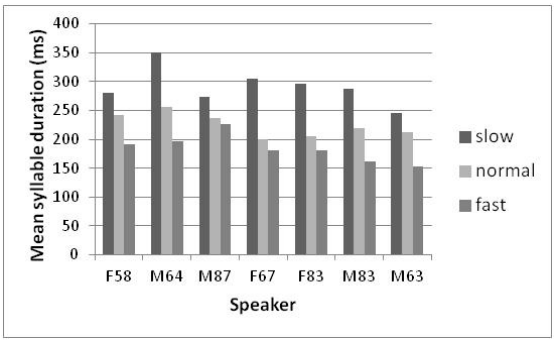


Figure 2. Graph of average syllable duration for each speaker in each speech rate.

If anchoring of the tonal targets with specific points in the segmental string occurs this would necessitate a correlation between segment duration and distance between tonal targets. This was tested by means of a Correlations Pearsons (2-tailed) test. There appears to be a weak to moderate positive linear relationship ($R = 0.433$, $N = 74$), which indicates a correlation. However, there does not seem to be a convincingly strong correlation (*Figure 3*).

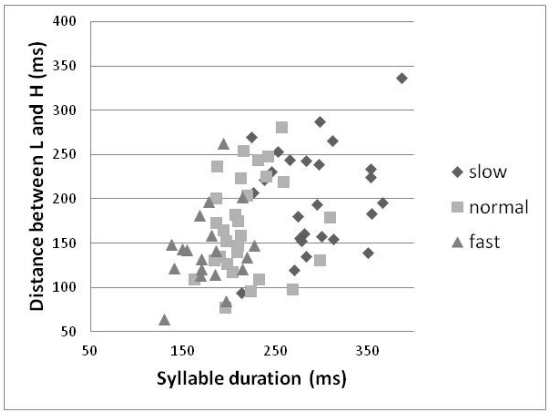


Figure 3. Scatter plot of the relationship between syllable duration and the distance between tonal targets L and H.

To test whether the weak to moderate relationship might indicate that either only one or neither of the targets is anchored, two new measures were calculated: distance between L and syllable onset, and H and syllable offset. An ANOVA was first run with speaker as random sample. To account for missing data, an average value was first calculated for each speaker across the available items for each condition. The ANOVA showed no significant effect of speech rate on the distance between the tonal target L and syllable onset (F =0.460, df 1, p = .523). An ANOVA was also run on the distance between H and syllable offset, showing no significant effect of speech rate on anchoring of H (F = .702, df 1, p = .434). However, *Table 1* displays quite a large standard deviation as well as the anomaly of only one available item in some conditions for three of the speakers (M64, M87 and F83). To avoid a type II error, additional ANOVAs were made with target words as random sample.

*Table 1. Average distance between tonal target and segment border (ms) for each speaker. Negative number indicate that the target is before the border. Standard deviation in parentheses. * Only one item in this condition.*

	L - syllable onset			H - syllable offset		
	slow	normal	fast	slow	normal	Fast
F58	45 (47)	-6 (97)	-9 (1)	-45 (18)	-58 (14)	-79 (32)
M64	-16 (22)	-40 (*)	82 (*)	-103 (49)	-16 (*)	-29 (*)
M87	-2 (43)	-39 (44)	3 (*)	-56 (32)	-29 (48)	-76 (*)
F67	62 (56)	-4 (9)	0 (4)	-35 (21)	-37 (10)	-43 (17)
F83	34 (*)	67 (39)	-30 (28)	-24 (*)	-19 (13)	-15 (18)
M83	57 (26)	15 (38)	-1 (45)	-93 (9)	-54 (20)	-47 (7)
M63	-31 (36)	7 (29)	-7 (8)	-50 (26)	-32 (25)	-33 (31)
All	24 (51)	9 (54)	-6 (36)	-65 (37)	-38 (23)	-40 (27)

The ANOVA with target word as sample showed that speech rate did in fact have a significant effect on the distance between L and syllable onset (F = 14.095, df 1, p = .013). This is evidence against segmental anchoring of the L target. An ANOVA on the distance between H and syllable offset was also calculated which shows a low p-value; however not statistically significant (F = 5.159, df 1, p = .072). Speech rate appears to not affect the possible anchoring of H. Average and standard deviations are shown in *Table 2*.

Table 2. Average distance between tonal target and segment border (ms) for each target word. Negative number indicate that the target is before the border. Standard deviation in parentheses.

	L - syllable onset			H - syllable offset		
	slow	normal	fast	slow	normal	fast
1	4 (22)	-6 (37)	-9 (37)	-39 (32)	-28 (25)	-47 (37)
2	23 (65)	23 (51)	-1 (18)	-76 (25)	-56 (19)	-27 (37)
3	5 (50)	-2 (59)	-27 (34)	-78 (54)	-42 (19)	-28 (17)
4	54 (56)	22 (78)	-9 (11)	-79 (39)	-40 (21)	-26 (22)
5	35 (50)	23 (59)	-8 (9)	-63 (33)	-23 (17)	-58 (16)
6	27 (64)	6 (60)	18 (59)	-52 (30)	-40 (29)	-47 (18)
All	24 (51)	9 (54)	-6 (36)	-65 (37)	-38 (23)	-40 (27)

Discussion

This study does not support the anchoring of L in the L+H rise of South Swedish Accent II. The rise is surely an important feature of the word accent (Ambrazaitis & Bruce, 2006), but if the start of the rise is not anchored it is possible that L is not a phonological event. The end of the rise, the timing of H, might be an important phonological feature. Independent of

syllable duration the timing of H was found approximately 40 ms before syllable offset, which supports the Lund Model and the accent typology that incorporates the South Swedish dialect.

The data displayed a great variability both between and within speakers. Ladd et al. (1999) also observed a similar degree of variability. By excluding certain speakers that seemed to use a different strategy to define pitch accent, they were able to find support for segmental anchoring. It might be that constant alignment is a strategy only for some of the speakers or that the same speaker might use different strategies for aligning the tones to the segment. The proposition by Niebuhr et al. (2011) to include speaker strategy in investigation on tonal alignment is thus a valid suggestion.

The auxiliary hypothesis that primary features of phonetic properties will try to be retained by speakers, while other features will be allowed to be modified by time pressure might be the case for the normal and the fast rate. The slow rate seemed, however, to divert from the others and induces other compensatory prosodic features. This can be seen in the scatter plot of the correlation test (*Figure 3*) with the slow rate being much more scattered across the graph than the normal or the fast rate. The anomalies found in this study on slow speech rate have also been reported in other studies, where difficulties with the slow speech rate seem to have brought forth additional prosodic features to enable the, perhaps, unnaturally slower speech (Ladd et al., 1999). Even though the results of the study confirmed that the manipulation of speech rate was successful, a future use of speech rate as an experimental tool needs to be further investigated.

The coincidence was pointed out that the rise of the pre-nuclear accent in intonation languages phonetically roughly matched the lexical Accent II in South Swedish. The results, however, did not confirm a phonological match. The start of a rising pre-nuclear accent in an intonation language needs to be anchored, but this does not seem to be the case for a South Swedish Accent II rise. Since evidence was found against L anchoring with syllable onset, the results do not support the revised Lund Model of a LH gesture. Further studies on the anchoring of the LHL tonal gesture in Swedish Accent II are suggested.

Acknowledgements

While working with an earlier version of this master thesis I was supervised by professor Gösta Bruce and associate professor Hugo Quené. I am very grateful for having had their excellent supervision in the initial steps of the study. I also want to extend a big thank you to my supervisor Gilbert Ambrazaitis for his invaluable help, understanding and enthusiasm.

References

Ambrazaitis, G. & G. Bruce. 2006. Perception of south Swedish word Accents. *Working papers* 52:5–8. Lund University, Lund, Sweden.

Arvaniti, A., D. R. Ladd & I. Mennen. 1998. Stability of tonal alignment: the case of Greek prenuclear Accents. *Journal of Phonetics* 26:3–25

Atterer, M. & D. R. Ladd. 2004. On the Phonetics and Phonology of "Segmental anchoring" of F0: Evidence from German. *Journal of Phonetics* 32:177–197.

Boersma, P. & D. Weenink. 2010. Praat: doing phonetics by computer [Computer program]. Version 5.2.03, retrieved 24 November 2010 from http://www.praat.org/

Bruce, G. 2007. Components of a prosodic typology of Swedish intonation. In Tomas Riad & Carlos Gussenhoven (eds.), *Tones and Tunes, Volume 1: Typological Studies in Word and Sentence Prosody*. Berlin: Mouton de Gruyter, 113–146.

Bruce, G. & E. Gårding. 1978. A prosodic typology of Swedish dialects. In: Eva Gårding, Gösta Bruce & Robert Bannert (eds.) *Nordic prosody. Papers from a symposium.* Lund University, Lund, Sweden, 219–228.

Caspers, J. & V. J. Van Heuven. 1993. Effects of time pressure on the phonetic realization of the Dutch Accent-lending pitch rise and fall. *Phonetica* 50:161–171.

Ladd, D. R., D. Faulkner, H. Faulkner & A. Schepman. 1999. Constant "Segmental anchoring" of F0 movements under changes in speech rate. *The Journal of the Acoustical Society of America* 106 (3):1543–1554.

Niebuhr, O., M. D'Imperio, B. G. Fivela & F. Cangemi. 2011. Are there "Shapers" and "Aligners"? Individual differences in signaling pitch accent category. In: *Proceedings of ICPhS XVII*, Hong Kong, 120–123.

Niemann, H., D. Mücke, H. Nam, L. Goldstein & M. Grice. 2011. Tones as Gestures: the Case of Italian and German. In: *Proceedings of ICPhS XVII*, Hong Kong, 1486–1489.

Xu, Y. 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica* 55:179–203.

Relative durations of post-vocalic consonants in read-aloud Spanish by native Swedish L2-learners

Bosse Thorén
Department of Language Studies, Umeå University, Sweden

Abstract

Prosody is said to be a persistent feature in foreign accents, particularly in the speech of adult learners. Native Swedish speakers have been reported to lengthen vowels in Spanish as well as post-vocalic consonants in English and German, more than native speakers of the respective languages. The aim of the present study is to examine whether native Swedish learners of Spanish produce increased post-vocalic consonant durations in a reading aloud exercise. The text contains certain words that could be expected to trigger complementary consonant lengthening in native Swedish speakers. The result shows that there is no general tendency for native Swedes to lengthen post-vocalic consonants more than native speakers of Spanish in the present speech material. There are examples of longer consonant durations in the speech of the native Swedish subjects, but it could be a coincidence. A conclusion is that the text used in the study, and probably Spanish as a language, contains few words that are ideal for triggering lengthening of post-vocalic consonants in native Swedish speakers.

Introduction

It has been suggested by Thorén (2008) that Swedish temporal prosody is a useful feature when teaching Swedish as a second language. Bannert (1987) found that allocation of stress is a crucial property with respect to intelligibility of Swedish and Fant and Kruckenberg (1994) concluded that duration is the most reliable acoustic correlate to stress and that stress affects the durations of both vowels and consonants. The Swedish quantity system is known to be realized as /V:C/ or /VC:/ in most regional varieties. The quantity distinction occurs only in stressed syllables, thereby giving stressed syllables longer duration than unstressed ones. The complementary length pattern is persistent in the sense that it affects how native Swedes pronounce English and German, resulting in longer post-vocalic consonant durations than those produced by native speakers of English and German (Thorén, 2007). Both English and German have phonological contrasts involving duration: In the case of German a clear vowel quantity with minimal pairs such as *beten-betten* 'beg-beds'. English distinguishes e.g. *beet-bit* but there is a debate whether this contrast should be defined as a spectral phonemic contrast between a tense and lax vowel, or as a quantity distinction signalled mainly by duration.

Aronsson (2013) found that Swedish learners of Spanish produced significantly longer vowels in stressed syllables compared to a group of native Spanish speakers. The present study examines some of the recordings of Aronsson (2013) in search of possible long post-vocalic consonants in the speech of the native Swedish learners of Spanish. The aim of the study is to find out whether the lengthening of stressed syllables found by Thorén (2007) is present also in the Spanish spoken by native Swedes. If the complementary /V:C-VC:/-pattern manifests itself in various L2s spoken by native Swedes, it would be an additional reason to suggest the complementary lengthening of stressed syllables as a core feature of Swedish pronunciation (cf. Thorén, 2008).

Theoretical and methodological considerations

Two measurements are used here in an attempt to capture two aspects of Swedish speech rhythm: vowel duration divided by consonant duration (V/C) to reflect the complementary relation between a stressed vowel and a following consonant, consonant duration divided by word duration (C/W) to reflect how the duration of one segment can contribute to the duration of the syllable. Measuring syllable duration is however problematic as syllable boundaries cannot always be readily established, and especially when trying to meet syllabification demands from two languages simultaneously. The word (or the two last syllables in words containing more than two syllables), is considered a reasonable compromise.

Had the present author been more familiar with Spanish phonology, the present material may not have been chosen for the study. Out of 22 potential /VC:/-words, only one, *dormilón* 'sleepyhead' seems to meet the criteria for an ideal potential /VC:/-word, since it has a short non-diphthong vowel as nucleus in a clearly stressed syllable, with a single post-vocalic consonant. Nine words in the sample had VCC-structure, i.e. they had 2 consonants following the stressed vowel. Behne and Czigler (1995) found that more duration is allocated to the first consonant than to the second consonant in VCC-words when effects of inherent durations were neutralized, but still less than to a single phonologically long consonant. Measuring VCC-words could thus be expected to yield less significant lengthening when pronounced by native Swedish speakers. In addition to this, the first consonant in 6 out of 9 VCC-words is a nasal, a category that showed less complementary lengthening in Thorén (2007), a condition that lowers the expectations also for the 'ideal' word *dormilón*. The words *Viejo, dijo, ojos* and *noche* have, according to Spanish phonology, a syllable boundary between the stressed (first) vowel and the following consonant. They are however included as they are regarded as possible /VC:/-candidates in the intuitive perception of a native Swede.

Measuring the duration of the stressed vowel divided by the duration of the following consonant (V/C) is assumed to reflect how the L2-speaker treats the local relationship between vowel and consonant, thus realizing a possible perceived quantity category (cf. Elert, 1964). The ratio of the post-vocalic consonant divided by the duration of the word (C/W), is assumed to reflect the potential impact of a Swedish habit to always give more duration to mainly one segment in a stressed syllable.

Method

The recorded material used in this study is part of a corpus described in Aronsson (2013) and is used here with permission. Below follows a brief description. For more details, see Aronsson (2013).

Reading task

The text used in the study is a childrens' story called *El viejo gallo* 'the old rooster'. It contains 363 words and 22 of those were first chosen for measurement.

Participants

The 8 native Swedish learners of Spanish were all in their first semester of university studies, which means that they had received formal instruction of approximately 250 hours in the Swedish school system. The 8 native Spanish speakers in the control group were from both South America and Spain.

Recordings

The speech was recorded in a language lab with varying recording quality. Some of the measurements had to be omitted since segmentation could not be done satisfyingly, sometimes due to poor technical quality and sometimes due to indistinct articulation as well as the general problem of deciding boundaries between vowel and nasal/liquid.

Material

Fourteen different words, out of originally 22 were measured. 9 were VCC, 4 were V.C but were judged as potential /VC:/. 1 word was a clear VC.

Analysis

The speech material was analysed in Praat (Boersma & Weenink, 2013). Durations of words as well as vowel and post-vocalic consonant in stressed syllables were measured. If the vowel was a diphthong, only word and consonant durations were measured. Vowel duration divided by consonant duration was calculated as a measure of quantity realisation and consonant duration divided by word duration was calculated as a measure of stress induced lengthening. Words like v*iejo, ojos* and *dijo* were regarded as possible /VC:/-words although there is a syllable boundary between vowel and following consonant according to conventional Spanish phonology. Nine words had a post-vocalic two-consonant-cluster, e.g. *alto*, that could be expected to trigger a /VC:C/ pattern in a native Swede. In the word *noche* only the occlusion phase was measured, making segmentation more secure.

Polysyllabic words like *momento, effectivamente, rapidamente, dormilón* were measured including only the 2 last syllables in an attempt to neutralize substantial differences in how the speakers initialized these words as well as differing proficiency in reading aloud.

Result

As shown in *Tables 1 and 2*, the duration values do not suggest that the Swedish speakers in general lengthened post-vocalic consonants more than native speakers of Spanish. Low values for V/C and high values for C/W would reflect the expected complementary pattern. The word *dormilón* is the only word in the present sample that has a single potential long post-vocalic consonant with no syllable

Table 1. V/C-ratios and C/W-ratios (mean values) for the measured VCC-words. Grey filling indicates that ratios reflect longer relative consonant durations for the Swedish L2-speakers compared to the Spanish L1-speakers.

	Alto	*campo*	*(mo-) mento*	*tanto*	*mismo*	*(effectiva-) mente*	*(rapida)- mente*	*triste*	*Quenta*
V/C L1	1,04 N =3	1,06 N=8	1,03 N=16	1,24 N=8	0,72 N=7	1,12 N=8	0,98 N=8	1,12 N=5	
V/C L2	1,41 N=3	0,95 N=8	1,04 N=10	1,26 N=8	0,57 N=6	0,96 N=5	0,96 N=5	0,75 N=5	
C/W L1	0,20 N=6	0,22 N=8	0,20 N=16	0,26 N=8	0,31 N=7	0,19 N=8	0,23 N=8	0,24 N=5	0,25 N=7
C/W L2	0,18 N=4	0,21 N=8	0,16 N=12	0,23 N=8	0,28 N=6	0,21 N=5	0,19 N=5	0,28 N=5	0,21 N=7

Table 2. V/C-ratios and C/W-ratios (mean values) for potential /CVː/ words, single consonant after short stressed vowel. Grey filling as in Table 1.

	viejo	*Dijo*	*ojos*	*noche*	*(Dor) milón*
V/C L1		0,71 N=2	0,61 N=6	1,77 N=12	1,39 N=7
V/C L2		0,76 N=4	0,68 N=5	1,30 N=11	0,77 N=7
C/W L1	0,33 N=45	0,44 N=7	0,30 N=6	0,19 N=15	0,23 N=7
C/W L2	0,32 N=41	0,33 N=6	0,26 N=5	0,17 N=12	0,31 N=7

boundary between V and C. This word also shows the highest degree of consonant lengthening expressed as V/C and C/W differences between L1 and L2 speakers. The box plot in *Figure 1* shows how L1 and L2 speakers realized the temporal relation between V and C, with lower values for the Swedish subjects (i.e. the two last syllables of the word). *Figure 2* shows consonant durations divided by word durations; higher for native Swedish speakers, reflecting the syllable-lengthening effect due to the duration of the post-vocalic consonant.

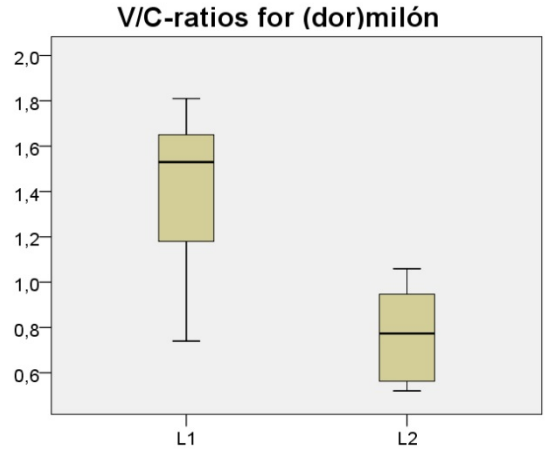


Figure 1. V/C-ratios for the word "dormilón" comparing durations of L1 and L2 speakers. N=7.

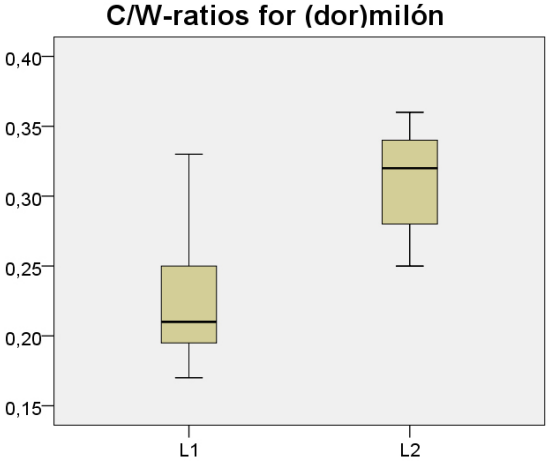


Figure 2. Consonant durations divided by word durations for the word "dormilón", comparing durations for L1 and L2 speakers. N=7.

Discussion

The result does not show any clear differences with respect to expected Swedish temporal patterns. A typical Swedish pattern would have shown overall lower values for C/V-ratios and

higher for C/W for the Swedish L2-speakers compared to Spanish L1-speakers. Both these differences between native Swedish and native Spanish speakers could be caused by extra-long post-vocalic consonants. Lower V/C-values can of course also be the result of extra short vowels, which indeed seems to be the case as mean absolute vowel durations in the L2-versions of *dormilón* are 26% shorter than in the L1-versions. A typical Swedish temporal pattern could also be manifested as extra lengthening of stressed syllables, and in the chosen words this extra duration is expected to be associated with the post-vocalic consonant, manifesting itself as higher C/W-ratio. An extra long consonant will in other words affect both V/C and C/W values. The mean absolute duration of the post-vocalic consonant in the L2 versions of *dormilón* is 25% longer than in the L1 version.

The present author – who has no knowledge of Spanish – could not detect any typical Swedish accent in more than 1 of the native Swedish speakers, although listeners with knowledge of Spanish report clear audible Swedish accent in all native Swedish subjects. This person however did not show more than marginally longer durations in postvocalic consonants than the rest of the L2-speakers.

It could be assumed that a group of native Swedish learners of Spanish who started learning the language as adults, would show more typical Swedish accent, manifested partly as longer post-vocalic consonants, but on the other hand, the speakers in Thorén (2007) had been 9–10 years old when they started to learn English, and they still had a measurable Swedish accent manifested as longer post-vocalic consonants. A possible explanation could be that Spanish phonotactics and prosody generate few structures that would typically trigger the Swedish complementary consonant length.

Acknowledgements

I want to thank Berit Aronsson of Umeå University for giving me access to her recordings that made the present study possible.

References

Aronsson, B. 2013. Patterns of Prominence in Swedish Second Language (L2) Speakers and Native (L1) Speakers of Spanish: Spontaneous Dialogue versus Read Text. *Studies in Hispanic and Lusophone Linguistics* 6(2). [In press.]

Bannert, R. 1987. From prominent syllables to a skeleton of meaning: a model of prosodically guided speech recognition. In: *Proceedings of the XIth ICPhS Tallinn,*73–76.

Boersma, P. & D. Weenink. 2013 *Praat: Doing phonetics by computer* http://www.fon.hum.uva.nl/praat/

Behne, D. & P. Czigler. 1995. Distinctive vowel length and postvocalic consonant clusters in Swedish. *Phonum 3*, Department of Phonetics, Umeå University, 55–63.

Elert C.-C. 1964. *Phonologic Studies of Quantity in Swedish*. Uppsala: Almqvist & Wiksell.

Fant, G. & A. Kruckenberg. 1994. Notes on stress and word accent in Swedish *STL-QPSR* 2–3/1994, 125–144.

Thorén, B. 2007. Swedish Accent – Duration of Postvocalic consonants in native Swedes speaking English and German. *International Congress of Phonetic Sciences* (ICPhS) Saarbrücken, 6–10 August 2007, 1693–1696.

Thorén, B. 2008. *The priority of temporal aspects of L2-Swedish prosody – Studies in perception and production*. Doctoral dissertation. Department of Linguistics, Stockholm University.

Observed pronunciation features in Swedish L2 produced by L1-speakers of Albanian

Mechtild Tronnier[1] *& Elisabeth Zetterholm*[2]
[1] *Centre for Languages and Literature, Lund University, Lund, Sweden*
[2] *Department of Swedish, Linnaeus University, Sweden*

Abstract

The number of immigrants with Albanian L1 has increased in Sweden during the last two decades. This particular group is also present in the SFI-classroom. This contribution aims to present pronunciation features in L2-Swedish produced by Albanian L1-speakers, based on recordings of two Albanian speakers living in Sweden.

Introduction

Since the mid-90s the migration of Albanian L1-speakers across European countries increased due to the Balkan-conflict and political changes in the country of Albania. According to the Swedish Migration Board (www.migrationsverket.se), about 1300 immigrants from Kosovo and Albania have been granted permission of residence in 2012 alone.

Learning the local language is a major issue when establishing a new life in the new country. The number of participants with Albanian as L1 has increased in Swedish as a second language (Tronnier & Zetterholm, 2011).

In this contribution the sound inventory of the Albanian Language is given in a contrastive perspective, i.e. in comparison with Swedish. Furthermore, an investigation of observed pronunciation features in Swedish-L2 based on Albanian-L1 will be presented. Since Swedish L2-learners with Albanian L1 originate from both areas, the Republic of Albania and the Autonomous Province/Republic of Kosovo, two major Albanian dialects, Tosk and Gheg and their sound systems are taken into consideration.

Albanian: typology and geography

Albanian is an Indo–European language which forms its own branch and is currently the official language in the Republic of Albania and one of the two official languages in the Autonomous Province/Republic of Kosovo. It also occurs as a minority language in other parts of Southern Europe.

There are two major Albanian dialects: Gheg and Tosk. Gheg is mainly found in Kosovo and northern Albania and Tosk is found in southern Albania. The Shkumbin river forms a rough dividing line, where a transitional zone between the two dialects can be found.

There has also been an effort to establish a standard variant, which has been subject to emotional and political debate (Moosmueller & Granser, 2003). Since 1972, a variant containing mainly elements from Tosk serves as Standard Albanian not only in the Republic of Albania. Vehbiu (1997) points out that the originally Gheg speaking cultural elite have also adopted the Standard variety.

Some differences between the sound systems of the two dialects are presented below.

The Albanian sound system in comparison with Swedish

Vowels

According to Moosmueller and Granser (2003), the Standard Albanian variety based on Tosk comprises seven vowels: /i y e ə ɑ u o/. There is however quite some allophonic variation for these vowels not only between the different dialects, but also within each dialect. The latter applies specifically for the vowels / ə ɑ /.

In addition, five distinctive nasal vowels can be found in Gheg (Garlén, 1988): /ĩ ỹ ẽ ã ũ/.

Except for the nasal vowels, all Albanian vowels occur in Swedish. On the other hand, coarticulatory vowel nasalisation is found in Swedish. Furthermore, Swedish has a higher number of vowel phonemes, where the vowels /ʉ ø/ are probably most divergent. In Swedish vowel length variation goes together with quality variation – specifically for /a/. Not using the right quality with the correct phonemic length might be a potential foreign accent feature for Albanian L1-speakers

Consonants

The Tosk variant of Albanian has 29 consonantal phonemes, whereas Gheg has three more. Some consonantal phonemes of Albanian

do not occur in Swedish. The phonemes /ʃ ʦ ʧ/ are not exactly overlapping with the Swedish consonants, but are very similar, as there are /ɕ/ and /ʦ/. The latter is not an affricate, but a sequence of phonemes in syllable final position. /ʦɕ/ is used initially in the syllable in some dialects. The Swedish consonantal phoneme /ɧ/ does not occur in Albanian.

As becomes clear from the analysis in the L1-L2-map (Koreman et al., 2011) the voiceless stops are un-aspirated in Albanian, whereas they are clearly and strongly aspirated under several circumstances in Swedish.

Prosody and syllable structure

According to Garlén (1988), stress is variable in placement and distinctive in both Albanian dialects, which is similar to Swedish. Garlén gives the minimal pair ˈbari "the grass" and baˈri "(shep-)herds" as examples. Lloshi (1999) however points out that stress is mainly fixed, and the rhythm is trochaic, which leads to stress on the penultimate syllable in most words.

For Tosk, no quantity distinction is found, a distinction which is found for Gheg. Moosmueller and Granser (2003) refer to long and short vowels in Gheg in their investigation on vowel quality variation in stressed syllables. It is unclear whether vowel quantity distinction only applies to stressed syllables or to unstressed syllables as well. Swedish has a complementary quantity distinction on stressed syllables, which results in rhyme patterns like a) a short vowel combined with a following long consonant or a consonant combination or b) a long vowel with a short or no following consonant.

No word accent distinction occurs in either of the two Albanian dialects, which is characteristic for Swedish. Word accents will however not be discussed here.

The syllable structure in Albanian allows between null and maximally two pre-vocalic consonants. The same is the case for the post-vocalic position (Garlén, 1988). The Swedish syllable structure is similar to the Albanian one, but allows up to three consonants pre-vocalically and in simplex words up to three consonants post-vocalically. In case of morphologically complex suffixing, further consonants may occur post-vocalically in Swedish.

The present study

The current study is based on recordings made of two speakers of Albanian living in the southern part of Sweden. One male and one female speaker were recorded reading Swedish sentences, a short text and when describing a picture story. The sentences were compiled so that words containing all Swedish vowels and consonants and most of the Swedish consonant clusters were present in the material. Furthermore minimal words pairs contrasted by quantity characteristics, stress placement and word accents were built into the sentences. Many of these target words were also present in the short text and supplemented by further words, e.g. compound words.

Both speakers are originally from Kosovo and their dialect is Gheg. They were both in their 20s when they moved to Sweden. The Female speaker however has lived in the Republic of Albania for a longer period of time, where she earned a university degree. She is therefore familiar with the Tosk dialect and also speaks German and Serbo-Croatian. She has been living in Sweden for 27 years and took classes in Swedish language for only 3 months. The male speaker has been living in Sweden for four years at the time of recording and has studied Swedish for 2.5 years. Both speakers have a good enough command of Swedish to manage every day conversations. None of the speakers reports a comfortable command of English.

The recorded material was analyzed in an impressionistic way: when listening to the material, examples of non-L1 accented speech were extracted and transcribed. In many cases the phonetic features on which the foreign accent impression was based could already be pointed out by auditory observation only. Speech wave and spectrogram representations were used to get a closer look and a more detailed auditory impression.

Observed variation in Swedish L2

In General

Some pronunciation variation might have arisen due to the possibility that some words in the text might have been unknown to the readers. This could not easily be controlled for, because the investigation is part of larger study. In any case of a clear hesitation in the flow when reading the material, the observed variation was not considered any further. Other

pronunciation variations may be based on those Swedish letters which do not occur in the Albanian set of letters. These letters are *ä ö å* and are related to the phonemes /ɛ ø o/, but were mistaken for the letters *a o a* and pronounced as a variant of /a o a/. This misinterpretation did not occur for all cases where these particular letters were present.

Vowels

For both speakers the pronunciation of the Swedish vowel /ʉ/ shows the clearest vocalic variation. This phoneme is by one speaker almost always and by the other speaker less frequently pronounced as [u]. For example the words *gul* "yellow" and *jul* "Christmas", pronounced as [gʉ:l] and [jʉ:l] in Swedish, were pronounced as [gu:l] and [ju:l] by both speakers. This might be based on the misapprehension of the letter *u*, which represents a different phoneme in Albanian – which is /u/.

Other vocalic L1-features that occur in Swedish L2 are based on more complex factors and are presented below.

Consonants

The most prominent consonantal feature is the consistent pronunciation variation of /r/. There are many different allophones for /r/ in Swedish, between and within the different dialects. The variants, which the two Albanian speakers use were neither of the allophones in Swedish or were misplaced. In the recordings one speaker almost always uses [ɻ], which is also found in English. Note that the speaker did not report any knowledge of English. If the /r/ occurred in word final position she also used strongly rhoticised vowels instead of a sequence of V+/r/. In that respect the final syllable of the word *hjälper* "helps" is pronounced [pɚ]. The variant the other speaker used is the apical trill [r]. This is one of the Swedish /r/-variants, but scarcely used and hardly ever in unstressed syllables, function words or in syllable final position. Making use of the repetitive trill in all positions results in a very prominent /r/-sound, which deviates from the impression of what the Swedish language should sound like.

Both speakers often add a homorganic stop after the velar nasal, when in intervocalic position: the word *många* "many" is pronounced as [mɔŋga] or [mõga] instead of [mɔŋa]. Vowel nasalization rather than sequencing V+C_{nas} are discussed below.

In Swedish, word initial voiceless stops are usually aspirated, when in a stressed syllable and not part of a consonant cluster: *tall* "pine tree" is pronounced [tʰal:]. The Albanian speakers in most cases either do not produce any aspiration or produce short/weak aspiration and the example above becomes [tal:]. There seems however to be a certain distribution when aspiration is produced and when it lacks. In that sense, aspiration lacks mainly when a frontal stop encounters a back vowel, no matter if the syllable is stressed or not: as in *polis* "police" becomes [pɔˈli:s] and *tomat* "tomato" [tɔˈma:t], *pappa* "dad" [ˈpap:a] and *taket* "the roof" [ˈta:kətʰ].

On the other hand, when a back stop (i.e. a velar stop) encounters a back vowel – as in *kallas* "is called" [kʰal:as] or a frontal stop (an alveolar stop) encounters a front vowel [i y] – as in *tid* "time" [tʰi:d], the aspiration is correctly produced after the stop.

An additional accent feature produced by the two Albanian L1-speakers is final-obstruent devoicing. The Swedish word *golv* "floor" [gɔlv] is produced with a final devoicing, which results in [gɔlv̥]. The word *ägg* "egg" [ɛg:] is produced as [ɛg̊:ʰ], which not only includes final devoicing, but furthers the impression of voicelessness by the aspiration of the stop. Another example shows devoicing medially in a compound word: *guldpapper* "gold glittering paper" becomes [gulḍʰˈpapər] – where stress misplacement also occurs.

In that way, aspiration lacks where expected – which is word initially – and occurs where it doesn't belong, i.e. with a word final devoiced stop, which should be voiced.

Swedish has two fricatives /ɕ ɧ/, which are produced with double articulations in most dialects. The alveolo-palatal fricative [ɕ] – as in *kök* "kitchen" [ɕø:k] – is mostly replaced by the somewhat darker [ʃ], which occurs in Albanian. The fricative [ɧ] as in *sjö* "lake" [ɧø:] is also replaced by [ʃ] by one speaker. The other speaker shifts between [ʃ] and [x]. The latter sound is a dialectal allophone of /ɧ/, but is usually not produced as similarly strong in the corresponding dialect as by the Albanian L1-speaker when speaking Swedish L2.

Prosody

Stress is placed on the wrong syllable in many compound and simplex words by both speakers. It occurs many times on the penultima syllable, a regular place for stress in

Albanian, but it may be misplaced on other syllables, too. Examples where the stress is misplaced on the penultima syllable are: *ˈläraren* "the teacher", *ˈarbetar* "works" and *ˈgrönˌsaker* "vegetable" become **läˈraren* **arˈbetar* and **grönˈsaker*

Stress misplacement onto the final syllable occurs in: *ˈsaluˌhall* "market hall" and *ˈkonstˌbok* "a book about arts". They become **saluˈhall* and **konstˈbok.*

Also in the flow of speech prominence variation produced by the Albanian L1-speakers is not always applied correctly. In a neutral utterance in Swedish, pronouns, clitics and auxiliaries are usually less prominent and in a sequence of adjective + noun, the noun is more prominent: *Hade ˈpå sig* "was wearing", *ˌvackert ˈväder* "nice weather" and *ˌvid ˈjul* "at christmas" become **Hade på ˈsig,* **ˈvackert ˌväder* and **ˈvid ˌjul*

Quantity aspects are in many cases not handled accurately by the two L1-speakers of Albanian. The words of minimal pairs as in: *väggen* "the wall" vs. *vägen* "the road", löss "lice" vs. lös "loose" and *villan* "the house" vs, *vilan* "the recreation" are hardly distinguished in pronunciation, favoring the words containing the long vowels, i.e. *vägen, lös* and *vilan.* This feature also occurs in words, not being part of a minimal pair: *semester* "vacation" [səˈmɛstɛɹ] becomes [səˈmeːstɛr].

The contrast between a stressed and an unstressed syllable is not always as clear and explicit as one would expect for Swedish. This is based on an alteration of length distinction in the rhyme, where the two Albanian L1-speakers seem to produce similar rhyme length for stressed and unstressed syllables.

Complex variations

Sequences of V+C_{nas} are very much influenced by assimilation. Nasalisation of vowels does occur in Swedish, but in many cases the Albanian L1-speakers replaced the whole sequence with a nasal vowel, which sounds unnaturally marked in Swedish: *Man kan* "one can" [man kan] and *glänste* "shone" [glɛnstə] become [mã kã] and [glɛ̃stə]. As the two speakers' dialect is Gheg – where nasal vowels are part of the sound system – they were apparently applying them in Swedish.

Other examples of inappropriate assimilation can be found in the regressive voicing of /s/ in *Israel* (name of the country) – [z] never occurs in Swedish – and progressive devoicing for /v/ in *kvällen* "the evening" and *två* "two" which does not result in a somewhat weaker devoiced [v̥], but in a very prominent [f].

Favoring long vowels over short ones is not equally clear for /a/, where a mismatch with the corresponding vowel quality occurred in words containing the Swedish phoneme /a/, a phoneme which requires the quality of [ɑ] in the case of a long vowel in a stressed syllable and [a] in the case of a corresponding short one. Both speakers also prefer [a] for the long vowel, so that *vas* "vase" correctly pronounced [vɑːs] becomes [vaːs].

Summary and outlook

Although the investigated L1-speakers of Albanian are well understood when speaking Swedish L2, the most prominent pronunciation variation is based on the inaccurate choice of sounds for the phonemes /ʉ/ and /r/, un-aspirated stops, final devoicing, stress misplacement and inaccurate vowel quantity sometimes combined with incorrect vocalic quality. Stress misplacement and the production of inaccurate vowel quantity can lead to misunderstandings, as stress serves as an anchor point to attention and variation in vowel quantity can lead to the wrong word.

An interesting finding was that the stressed and unstressed syllables in L2 were of comparable strengths. An NPVI-analysis could shed further light on that relationship.

References

Garlén, C. 1988. *Svenskans fonologi*. Lund: Studentlitteratur.

Koreman, J., Ø. Bech, O. Husby & P. Wik. 2011. L1-L2*map*: a tool for multi-lingual contrastive analysis. *Proc. 17th Int. Congress of Phonetic Sciences (ICPhS 2011)*, Hong Kong, 599–602.

Lloshi, X. 1999. Albanian. In: U. Hinrichs (ed.), *Handbuch der Südost-Europa-Linguistik,* Wiesbaden: Harrassowitz Verlag, 277–299.

Moosmueller, S. & T. Granser. 2003. The vowels of Standard Albanian. *Proceedings of the 15th Int. Congress of Phonetic Sciences (ICPhS2003)*, Barcelona, 659–662.

Tronnier, M. & E. Zetterholm. 2011. New Foreign Accents in Swedish. *Proc. 17th Int. Congress of Phonetic Sciences (ICPhS 2011)*, Hong Kong, 2018–2021.

Vehbiu, A. 1997. Standard Albanian and the Gheg Renaissance: A Sociolinguistic Perspective. *International Journal of Albanian Studies* 1:1–14.

Author index

< this page intentionally left blank>

ISBN 978-91-7519-582-7
eISBN 978-91-7519-579-7
ISSN 1403-2570