Printed by Universitetsservice AB

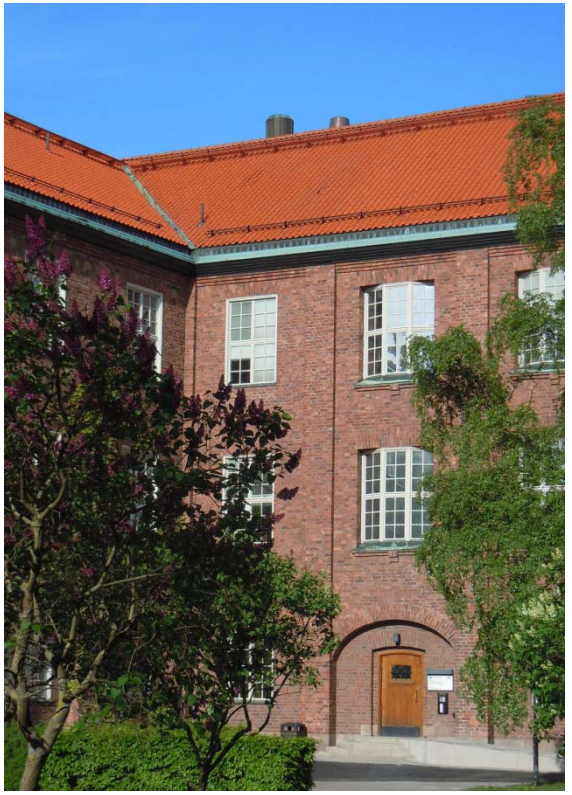
Stockholm 2011

Fonetik 2011

KTH, Speech, Music and Hearing

TMH-QPSR Vol. 51

Fonetik 2011



Speech, Music and Hearing
TMH-QPSR Vol. 51

KTH Computer Science
and Communication

Speech, Music and Hearing

Quarterly Progress and Status Report

TMH-QPSR, Volume 51, 2011

Proceedings from

Fonetik 2011

June 8 – June 10, 2011

Organized by
Department of Speech, Music and Hearing,
and Centre for Speech Technology, CTT,
KTH, Stockholm

Organizing Committee:

Björn Granström
David House
Daniel Neiberg
Sofia Strömbergsson

Front and back cover:
The meeting venue at KTH
Photos by Björn Granström

Published by Department of Speech, Music and Hearing, KTH, Stockholm

TRITA-CSC-TMH 2011:1
ISSN 1104-5787
ISRN KTH/CSC/TMH--11/01--SE

Printed by Universitetsservice AB, Stockholm
2011

Table of Contents

Author index

Preface

This volume of QPSR, the 51th in a long series of KTH publications, contains the 31 contributions to Fonetik 2011, the annual Swedish Phonetics Conference. It has been organised since the mid 1980's by different university departments involved in phonetics. This time the Department of Speech, Music and Hearing at KTH hosted the conference which was held on June 8 – June 10, 2011 at KTH

The conference was attended by close to 75 participants, mainly from Sweden and the other Nordic countries. Fonetik 2011 displays a variety of topics reflecting the wide range of activities in this field.

We thank all the contributors for their co-operative work to make this volume available in time for the conference. The conference activities and the printing of this volume were economically supported by Fonetikstiftelsen (the Swedish Phonetics Foundation) and by CSC (the School of Computer Science and Communication) at KTH, which we gratefully acknowledge.

The contributions in this volume are also published on the web, as are the previous 50 QPSR volumes - http://www.speech.kth.se/qpsr/

The Fonetik 2011 organisers

Björn Granström *David House* *Daniel Neiberg* *Sofia Strömbergsson*

Previous Swedish Phonetics Conferences (from 1986):

1986	Uppsala University
1988	Lund University
1989	KTH Stockholm
1990	Umeå University (Lövånger)
1991	Stockholm University
1992	Chalmers/Göteborg University
1993	Uppsala University
1994	Lund University (Höör)
1995	(XIIIth ICPhS in Stockholm)
1996	KTH Stockholm (Nässlingen)
1997	Umeå University
1998	Stockholm University
1999	Göteborg University
2000	University of Skövde
2001	Lund University (Örenäs)
2002	KTH Stockholm
2003	Umeå University
2004	Stockholm University
2005	Göteborg University
2006	Lund University
2007	KTH Stockholm
2008	Göteborg University
2009	Stockholm University
2010	Lund University

There is a web-page maintained by H. Traunmüller with links to all previous phonetics conferences at:
http://www2.ling.su.se/fon/fonkonfer.html

An acoustic analysis of lion roars. I: Data collection and spectrogram and waveform analyses

Robert Eklund [1,2,3], *Gustav Peters* [4], *Gopal Ananthakrishnan* [5] *& Evans Mabiza* [6]
[1] *Voice Provider, Stockholm, Sweden*
[2] *Department of Cognitive Neuroscience, Karolinska Institute, Stockholm, Sweden*
[3] *Department of Computer Science, Linköping University, Linköping, Sweden*
[4] *Forschungsinstitut Alexander Koenig, Bonn, Germany*
[5] *Centre for Speech Technology, Royal Institute of Technology, Stockholm, Sweden*
[6] *Antelope Park, Gweru, Zimbabwe*

Abstract

This paper describes the collection of lion roar data at two different locations, an outdoor setting at Antelope Park in Zimbabwe and an indoor setting at Parken Zoo in Sweden. Preliminary analyses of spectrographic and waveform data are provided.

Introduction

Felids are one of the most successful carnivore families ever to exist, and within the 35–40 different cat species that exist today several different vocalizations can be found, with different functions, ranging from the well-known purring to the most impressive sound of them all: roaring of lion (*Panthera leo*) fame. This paper focuses on the impressive lion roaring, and highlights methodological problems associated with the collection of animal vocalizations data.

Roaring: a primer

For a human observer the roaring of a lion – even more so that of a whole pride – certainly is one of the most impressive vocalizations in the animal kingdom. In its complete form lion roaring is a species-specific series of calls with a fairly regular structure of the single calls composing it and the series itself, in the latter in terms of the sequence of call types, their change of intensity in the course of the series, the temporal sequencing of the calls and their relative duration and that of the intervals between them. A typical lion roaring can last for more than a minute, usually starting off with a few low-intensity moan-like calls, then progressively increasing in intensity and duration of the calls, and in approaching the intensity climax of the series the calls become shorter again and harsher. After the climax follows a series of short harsh calls, in the beginning uttered at fairly monotonic intensity and brief intervals between the calls, then towards the end of the series gradually decreasing in intensity and with increasing interval duration (called “outro” in this paper).

Given the fact that the colloquial term ‘roar’ is commonly used for various intense animal vocalizations it is not surprising that even in the lion it has been applied to vocalizations which are definitely different from roaring as dealt with here. Early attempts at characterizing it in a more technical manner were published by e.g. Leyhausen (1950), Hemmer (1966) and Schaller (1972). More recent studies of lion roaring include Peters (1978), Peters & Hast (1994), and Pfefferle et al. (2007) .

Weissengruber et al. (2002:208) extended the definition of roaring in a general vertebrate vocalization context suggesting that lion roaring “has two distinct physiological and acoustic components:

1 a low fundamental frequency, made possible by long or heavy vocal folds, which lead to the low pitch of the roar;

2 lowered formant frequencies, made possible by an elongated vocal tract, which provide the impressive baritone timbre of roars.”

(See also Frey & Gebler, 2010).

In this paper, we studied lion roaring ‘proper’ as outlined at the start, in respect of the fine acoustic structure of its component single calls, the structural changes they undergo in the course of the roaring series and possible physiological mechanisms underlying these changes, considering the definition suggested by Weissengruber et al. (2007).

On the function of roaring

The function of lion roars has been discussed extensively in the literature, and several hypotheses have been suggested. Pfefferle et al. (2007:3952) concluded that the "primary function of roars is the advertisement and defense of territory". In support of this hypothesis, it has been shown that lionesses can estimate the number of individuals roaring, and that they are less likely to approach foreign roars when they are outnumbered (McComb, Packer & Pusey, 1994).

Besides territorial defense, an additional function might also be coordination of hunting (Grinnell & McComb, 2001; McComb, Packer & Pusey, 1994; Schaller, 1972).

Method

The following sections describe the data collection, data processing and analysis tools.

Data collection

The data analyzed in this paper were recorded at two different locations, one outdoor and one indoor setting. Recording details are given below.

Antelope Park, Gweru, Zimbabwe

The first set of lion roar recordings was obtained at Antelope Park lion rehabilitation and release into the wild facility at Gweru, Zimbabwe, by the first and last author. Antelope Park presently holds a population of around 100 African (Zimbabwean) lions.

The recordings were made on 23 November 2010, between 0400 and 0600 hours in the morning at the main enclosure centre. This meant that at least 50 lions were within close earshot, and that most of the other 50 lions were also within hearing range, given that lion roars can be heard by humans at a distance of at least 8 kilometers (Sunquist & Sunquist, 2002:294). Estimated distance between the microphone and the lions varied from about four meters to several hundred meters, although the latter roars appeared as fairly weak signals.

The lions that were closest to the microphone were nine males, most of whom were born in 2006. Also close were seven other males with ages between seven and eight years old. Relatively close were another five males who are seven and eight years old, and also a number of females.

As it was more or less pitch-black during the recording it was impossible to know exactly what lion produced exactly what roar, or whether the roars were produced by a male or a female, although the former is more likely. Besides, there were considerable overlap between the roars of several lions (often more than a dozen at a time).

The equipment used was a Canon HG-10 HD camcorder with a clipon DM50 electret stereo condenser shotgun microphone with a 150–15,000 Hz frequency range and a sensitivity of –40 dB. The microphone was directed towards the lions that roared for the moment, and thus its position varied.

Other than slight contamination with morning bird chirping, the soundscape was relatively calm.

The recording location, with setup indications, is shown in *Plate 1*.

Parken Zoo, Eskilstuna, Sweden

Parken Zoo is a wildlife facility about an hour's distance from Stockholm and holds a wide number of exotic animals, including several species of felids. There are presently three Asiatic (Gir) lions there: *Sarla*, a female born in 1997 (estimated 165 kilos); *Ishara*, another female born in 2007 (estimated 165 kilos); and *Kaya*, a male born in 1999 (estimated 180 kilos).

The recordings were made on 7 April 2011, between 0800 and 1000 hours in the morning. The recordings were made indoors to ensure that the lions remained in close proximity to the microphones – in their outdoor enclosure the lions would likely have walked off (far from the microphones). The cameras/microphones were set up by the first author. All three lions were at a distance from the microphones that varied between about one meter to around five meters. Since the recordings were made indoors, there were some echo effects. There was considerable contamination of the soundscape with bird chirps, emanating from a few birds perched somewhere in the enclosure.

The recordings were made with two Canon HG-10 HD camcorders. One camera used the same clipon microphone as is described above, while the other camera used an external professional high-fidelity Audiotechnica AT813 cardoid-pattern, condenser mono microphone, with a frequency range of 30–20,000 Hz and a sensitivity of –44 dB. The two cameras were placed so that, between them, they would cover as much of the enclosure as possible, with the hope of catching roaring sequences on film.

The recording location, with setup indications, is shown in *Plate 2*.

Plates 1 and 2. Recording setups at Antelope Park, Gweru, Zimbabwe (left) and Parken Zoo, Eskilstuna, Sweden (right). Left plate: Orange dots indicate approximate positions of roaring lions while the white dot indicates the most frequent position of the DM50 stereo microphone, ~250 cm above the ground. Right plate: White arrow indicates position of DM50 stereo microphone; orange arrow indicates position of AT813 mono microphone.

Plates 3 and 4. Roaring sequences caught on film at Parken Zoo, Eskilstuna, Sweden. Film captures from the two cameras lifted from the roaring sequences analyzed in this paper. Note that all film/sound files obtained at Antelope Park were recorded in complete darkness (to humans; the lions saw the authors quite well).

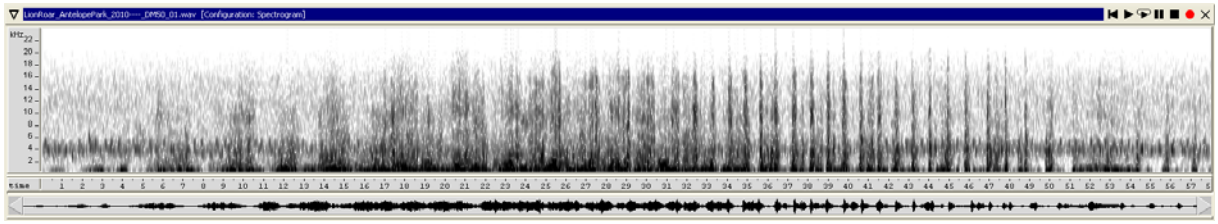


Figure 1. Spectrogram and waveform (excerpt) of multiple lions roaring sequence recorded at Antelope Park, Gweru, Zimbabwe. Canon DM50 clipon stereo microphone. Duration: 58 seconds.

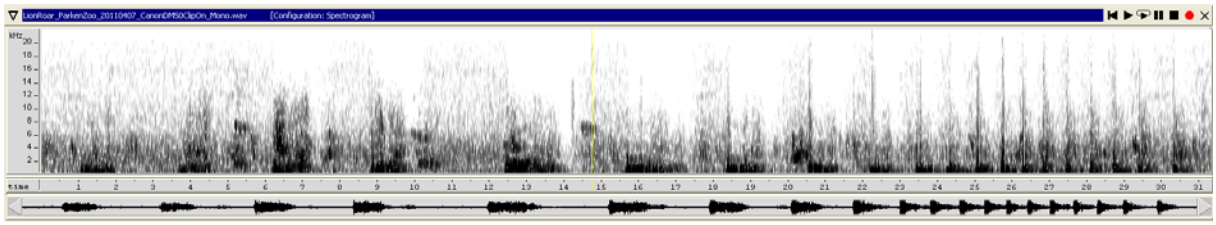


Figure 2. Spectrogram and waveform (excerpt) of lion roaring sequence recorded at Parken Zoo, Eskilstuna, Sweden. Canon DM50 clipon stereo microphone. Duration: 31 seconds.

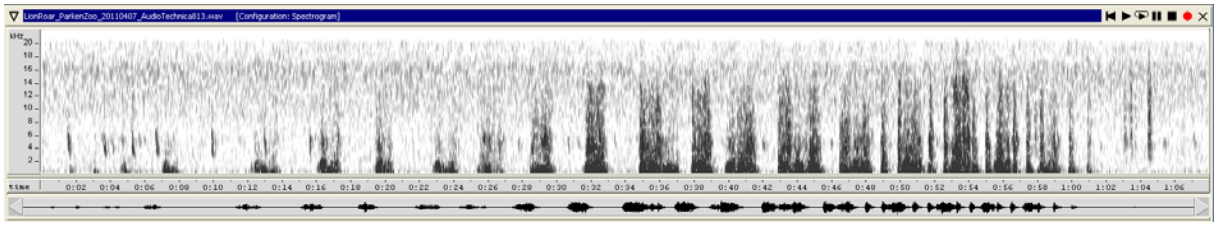


Figure 3. Spectrogram and waveform (excerpt) of lion roaring sequence recorded at Parken Zoo, Eskilstuna, Sweden. Audiotechnica AT813 external mono microphone. Duration: 67 seconds.

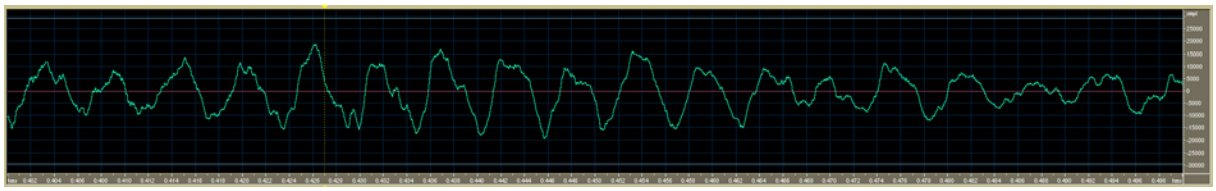


Figure 4. Waveform of lion roaring sequence ("outro" phase) recorded at Parken Zoo, Eskilstuna, Sweden. Audiotechnica AT813 external mono microphone. 18 distinct peaks – in a 100 ms sequence – give an estimated fundamental frequency of about 180 Hz.

Data post-processing

Audio tracks were extracted and converted into wav files (44.1 kHz, 16 bit, mono) with TMPGEnc 4.0 Xpress.

Analysis tools

Spectrogram and waveform analyses were carried out with WaveSurfer and Cool Edit.

Results

The film clips recorded at Parken Zoo resulted in two passages where the lions were caught on film while roaring; see *Plate 3* and *Plate 4*. This enabled comparison between acoustic and visual data (see Ananthakrishnan et al., 2011).

Spectrographic analysis

The three spectrograms shown in *Figure 1*, *Figure 2* and *Figure 3* all reveal the periodic phase characteristics of the roaring sequences. Despite the different acoustic characteristics between the microphones and the different recording setting, all three spectrograms reveal both low frequency components and a higher frequency component around 4 kHz.

Fundamental frequency analysis

A waveform passage is shown in *Figure 4*, and as is clearly seen there are 18 distinct peaks in the 100 ms long window. This gives an approximate fundamental frequency (F_0) of about 180 Hz, which is in accordance with the results reported by Pfefferle et al. (2007:3950), where mean F_0 in males was 194.55 Hz and 206.57 in females.. Naturally, further analyses are required in order to will reveal what degree of variation and range that occur in lion roars.

Discussion

The primary goal of this paper is to provide information about data collection issues associated with animal sounds, highlighting the difficulties involved when trying to obtain controlled high fidelity recordings of animal vocalizations. Future research will focus on more detailed acoustic analyses on the data obtained, and we hope to complement our data with additional high fidelity recordings of uncontaminated recordings of individual lions, in order to facilitate e.g. vocal tract estimation studies (see Ananthakrishnan et al., 2011).

Acknowledgements

Thanks to Jennie Westander, Conny Gärskog and Helena Olsson at Parken Zoo. Thanks to Jacqui Kirk at ALERT. Also thanks to Miriam Oldenburg for help with the "roarcordings".

References

Ananthakrishnan, G., R. Eklund, G. Peters & E. Mabiza (2011). An acoustic analysis of lion roars. II. Vocal tract characteristics. *Proceedings of Fonetik 2011*, 8–10 June 2011, Royal Institute of Technology, Stockholm, Sweden. [This volume.]

Frey, R. & A. Gebler (2010). Mechanisms and evolution of roaring-like vocalization in mammals. In: S. M. Brudzynski (ed.): *Handbook of Mammalian Vocalization – An Integrative Neuroscience Approach*. Amsterdam: Elsevier Academic Press, 439–450.

Grinnell, J. & K. McComb (2001). Roaring and social communication in African lions: the limitations imposed by listeners. *Animal Behavior* 62:93–98.

Hemmer, H. (1966). Untersuchungen zur Stammesgeschichte der Pantherkatzen (Pantherinae) Teil I. *Veröffentlichungen der Zoologischen Staatssammlungen München* 11:1–121.

Leyhausen, P. (1950). Beobachtungen an Löwen-Tiger-Bastarden, mit einigen Bemerkungen zur Systematik der Großkatzen. *Zeitschrift für Tierpsychologie* 7:46–83.

Peters, G. (1978). Vergleichende Untersuchung zur Lautgebung einiger Feliden (Mammalia, Felidae). *Spixiana* (Suppl.) 1:1–283.

Peters, G. & H. H. Hast (1994). Hyoid structure, laryngeal anatomy, and vocalization in felids (Mammalia: Carnivora: Felidae). *Zeitschrift für Säugetierkunde* 59:87–104.

Pfefferle, D., P. M. West, J. Grinnell, C. Packer & J. Fischer (2007). Do acoustic features of lion, *Panthera leo*, roars reflect sex and male condition? *Journal of the Acoustical Society of America* 121:3947–3953.

Schaller, G. B. (1972). *The Serengeti Lion – A Study of Predator-Prey Relations*. Chicago: Chicago University Press.

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

Weissengruber, G., G. Forstenpointner, G. Peters, A. Kübber-Heiss & W. T. Fitch (2002). Hyoid apparatus and pharynx in the lion (*Panthera leo*), jaguar (*Panthera onca*), tiger (*Panthera tigris*), cheetah (*Acinonyx jubatus*) and domestic cat (*Felis silvestris* f. *catus*). *Journal of Anatomy* 201:195–209.

An acoustic analysis of lion roars. II: Vocal tract characteristics

G. Ananthakrishnan [1], Robert Eklund [2,3,4], Gustav Peters [5] & Evans Mabiza [6]

[1] Centre for Speech Technology, KTH, Stockholm, Sweden
[2] Voice Provider, Stockholm, Sweden
[3] Department of Cognitive Neuroscience, Karolinska Institute, Stockholm, Sweden
[4] Department of Computer Science, Linköping University, Linköping, Sweden
[5] Forschungsinstitut Alexander Koenig, Bonn, Germany
[6] Antelope Park, Gweru, Zimbabwe

Abstract

*This paper makes the first attempt to perform an acoustic-to-articulatory inversion of a lion (*Panthera leo*) roar. The main problems that one encounters in attempting this, is the fact that little is known about the dimensions of the vocal tract, other than a general range of vocal tract lengths. Precious little is also known about the articulation strategies that are adopted by the lion while roaring. The approach used here is to iterate between possible values of vocal tract lengths and vocal tract configurations. Since there seems to be a distinct articulatory changes during the process of a roar, we find a smooth path that minimizes the error function between a recorded roar and the simulated roar using a variable length articulatory model.*

Introduction

The roar is a distinct mammalian vocalization made by only five species of *Felidae*. Researchers suggest that the ability to roar is made possible due to the specialized hyoid apparatus present in these mammals (Weissengruber et al., 2002). Acoustic-articulation modeling has been applied on several mammalian vocalizations in order to estimate the approximate vocal tract length of the animal producing the sound (Hauser, 1993; Taylor and Reby, 2010). The purpose has often been to correlate the estimated length of the vocal tract to the size of the animal to see if larger vocal tract lengths meant relative size dominance. The estimates were further correlated with the social behavior and mating roles of these vocalizations. Most of these methods applied the source-filter theory (Fant, 1970; Titze, 1994) to obtain inferences regarding the vocal tract characteristics. Here the properties of the larynx control the source signal characteristics, while the vocal tract configuration controls the filter characteristics. Since articulation data for mammals have not been very easy to obtain, most of these methods assume a uniform vocal tract for the mammals when they produce the sound and use the formant dispersion method (Titze, 1994; Fitch, 1997).

The lion roaring sequence usually consists of three different phases (Peters, 1978). The first phase is a series of low-intensity calls similar to 'mews'. The second phase, builds up to the climax with calls of increasing duration (shortening again towards the climax). Finally the sequence ends with a series of 'grunt' like sounds. In this study, we are interested in the second phase which is tonal in nature and has the maximum intensity in the entire sequence. Henceforth we only refer to the second phase by the word 'roar'.

Figure 1 shows the spectrogram of a prototypical roar of a female lion. It is clear that there is change in the formant structure also illustrated in *Figures* 2 and 3, showing the Spectral Envelopes varying over time and the average spectral slices for the two parts of a single roar respectively. This change in formant structure indicates that there is a corresponding change in the vocal tract dimensions during the process of producing the roar. Change in the quality of vocalizations have also been observed in other animals to where the vocalization includes protrusion of lips or jaw movement (e.g., Harris et al. 2006). Some species of fallow deer (*Dama dama*) are known to lower their larynx during the call (Vannoni et al., 2005).

Given this observation of changing formant structure during the roar, the uniform tube assumption can no longer be valid. One can suppose that that the filter (vocal tract) undergoes

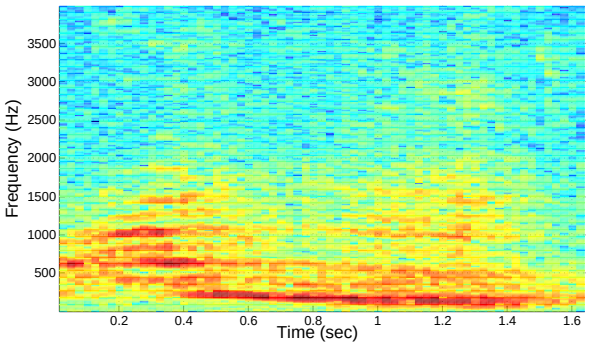


Figure 1: The spectrogram of a typical lion roar (in this case, a female lion's).

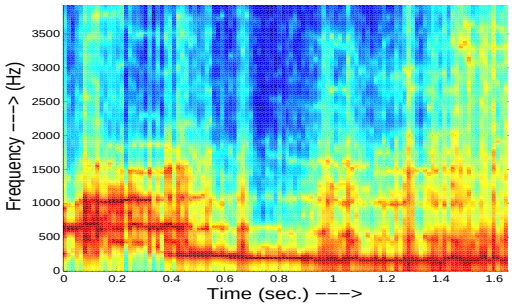


Figure 2: Illustration of the temporal changes in the formant structure, and therefore vocal tract configuration.

some change. However, one does not know what kind of change the vocal tract undergoes, whether it is the lowering of the larynx or changing of the vocal tract area function, or a combination of both.

Theory and Methods

The method proposed in this paper uses a Variable Linear Articulatory Model (VLAM) which allows the articulatory synthesizer developed by Maeda (1979) to be operated at different vocal tract lengths. Although this synthesizer has been designed for human-voices, the source-filter theory as shown previously by Taylor and Reby (2010) can be applied to other mammal vocalizations too. However, since the vocal tract area functions of a lion are largely unknown, we iterate over a range of values and select a configuration which best matches the spectral envelope of the recording of a lion roar. The several steps in the process are described below

1. The lion roar signal is segmented into overlapping windows, using the 'Hann' window function. Each window length is 30 ms in duration and successive windows are 5 ms apart.

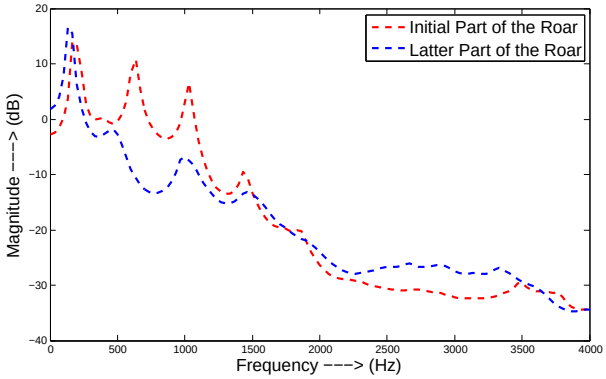


Figure 3: The Spectral Envelope, estimated using LPC analysis, from the beginning and the ending of one lion roar. This indicates that there is some change in the vocal tract configuration during the roar.

2. Linear Prediction Coefficients (LPC) were calculated for each window and then a Fast Fourier Transform (FFT) was applied, to the calculated transfer function so as to obtain the spectral envelope. The number of LPC parameters was set to 21, so as to obtain around 9 to 11 formant peaks within 4000 Hz. This was estimated based on the approximate dimensions of the Vocal-Tract Length (VTL) of a Lion, which is around 35 to 40 cm.

3. The spectral envelope for each window was converted to the decibel (dB) scale and normalized so as to limit the largest formant peak to 0 dB. We also subtracted the mean spectral slope from detected formants, so as to remove the effect of voicing in the estimates of the vocal tract shape.

4. We divided the vocal tract into three equal regions called the Jaw Section, Oral Section and the the Pharyngeal Section. The cross-sectional areas of the three sections were called JawSec, OralSec and PharSec respectively. We performed smoothing and linear interpolation on the three sections in order to approximate a 40 cylindrical tube model.

5. Using the VLAM simulations, we simulated the spectral transfer function, given different combinations of values for the four parameters VTL, JawSec, OralSec and PharSec. The spectral transfer function for each configuration was compared with the spectral envelope of the waveform for each time window to find the Euclidean distance between the two spectra.

6. Since several combinations of VTL and area functions can contribute to largely similar spectral characteristics Atal et al. (1978), we apply a smoothing function on the estimated vocal tract parameters. The movement being a muscular motion, a minimum jerk trajectory is the expected type of movement (at least for humans) Viviani and Terzuolo (1982). We thus apply a minimum jerk smoothing with multiple hypotheses Ananthakrishnan and Engwall (2011). The hypotheses are the 10 vocal tract configurations with minimum estimation error for each frame. These hypotheses are weighted by the inverse of the estimation error.

Data and Experiments

The data we used were recordings of lion roars made at two locations, namely, at the Antelope Park (Gweru, Zimbabwe), and Parken Zoo (Eskilstuna, Sweden). The equipment used at the Antelope Park was a DM50 electret stereo condenser shotgun microphone with a 150–15,000 Hz frequency range and a sensitivity of -40 dB. The estimated distance between the microphone and the lions varied from about four meters to ten meters, with the microphone pointing towards the general direction of a group of nine male lions (most of them born in 2006) in an open enclosure. Although there were other roars, we only considered the loudest roars which we assumed to be from the nine males mentioned above. The recordings at the Parken Zoo were made with two Canon HG-10 HD camcorders. One camera used the same microphone (DM50) as described above, while the other camera used an Audiotechnica AT813 cardoid-pattern, condenser mono microphone, with a frequency range of 30–20,000 Hz and a sensitivity of -44 dB. There where three lions, one male and two females. The male was 12 years old and weighed around 180 kilograms, while the females weighed around 165 kilograms was were around 14 years old. Further details of the data collected are mentioned in Eklund et al. (2011).

The waveforms were initially sampled at 44100 Hz, but were later sub-sampled to 8000 Hz to ensure compatibility with the VLAM model which estimated the vocal tract spectral transfer function in the frequency range of 0 to 4000 Hz. The waveforms were manually segmented to extract the second part of the roaring sequence, i.e. the tonal roar. We used a range of possible vocal tract lengths, ranging from 16 cm to 54 cm. The area functions for the three vocal tract

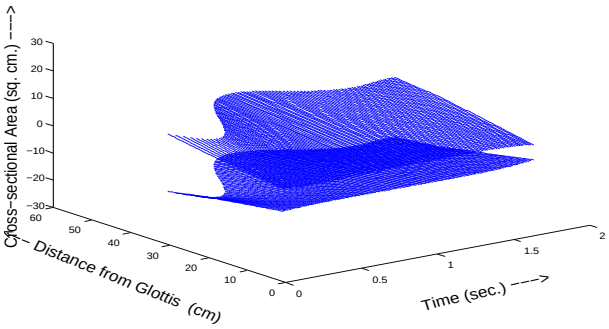
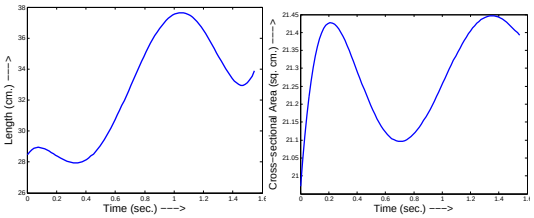
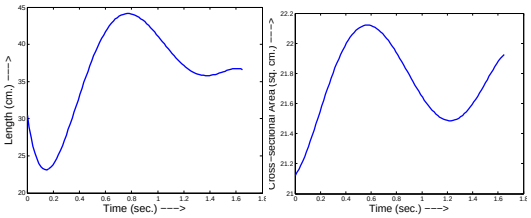


Figure 4: Illustration of how the vocal tract area function changes with respect to time during the course of a roar.



(a) Vocal Tract Length, Male (b) Jaw Section, Male



(c) Vocal Tract Length, Female (d) Jaw Section, Female

Figure 5: Illustration of how the vocal tract length and Jaw cross-sectional areas change with respect to time during the course of a roar.

sections were iterated between 8 to 24 sq. cm. These estimates were then compared with the videos sequences wherever available.

Results and Conclusions

Figures 4 and 5 indicate the estimated vocal tract shapes and VTLs over time for the female lion. This shows that vocal tract of a lion, approximates a frustum of a cone, rather than a uniform cylinder. The plots also indicate that, the roar involves a lengthening of vocal tract and then then a stabilization during the course of the roar. The range of variation is from 28 cm to 38 cm for the male lion and from 25 cm to 45 cm for the female lion. This may be effected by lowering the larynx, achieved during lifting up of the head. The female lion shows a larger variation in VTL during the course of the roar. The results

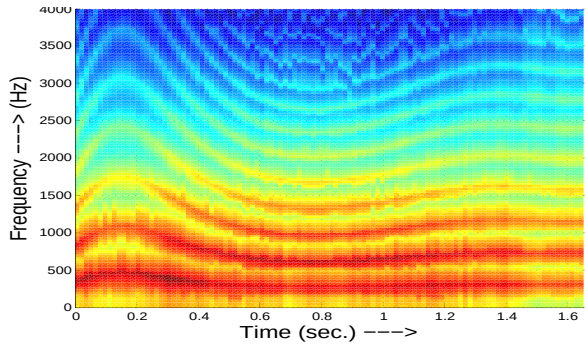


Figure 6: Illustration of the estimated spectral envelope of the lion roar.

also shows a slight decrease in the jaw area, especially for the female. In the videos that were recorded, the lions lifted their head up each time they roared. The jaw saw an increased opening followed by a reduction in the opening during the roars. The prediction from the estimates fit well with the observation about the VTL and the JawSec.

Dynamic analysis of animal vocalizations in order to extract the vocal tract characteristics is a very preliminary attempt in this paper. Some interesting observations have been uncovered in this study. The first being, the general shape of the vocal tract being more conical rather than cylindrical. Secondly, there seems to be a clear indication of larynx lowering, which is similar to the observations on fallow deer vocalizations (Vannoni et al., 2005). However, the female vocal tract is expected to be smaller than the male vocal tract given the differences in overall sizes. The mean VTL of the female lion's is estimated to be around 36 cm and is longer than the male lion's, estimated to be around 32 cm, which is rather unintuitive. Anatomical evidence for a male lion's vocal tract suggests a length of 38 cm Weissengruber et al. (2002). Estimating the mean VTL obscures the fact that the change in VTL for the female lion is also larger than the male lion's. This does not give any indication of what the static and normal lengths would be. Although some observations can be verified using video sequences, other observations need further data and analysis before make strong conclusions.

Future work would include analyzing physical, biological and ecological reasons for this type of motion during the roar, as well as other acoustic properties of the roar. Initial observations point to an increased roughness in the latter part of the roar, likely to be influenced by the voice source. This would also be an interesting investigation.

Acknowledgement

The authors would like to thank Jennie Westander and Conny Gärskog at Parken Zoo. Also thanks to Miriam Oldenburg for help with the "roarcordings" at Zimbabwe and Eskilstuna.

References

Ananthakrishnan G and Engwall O (2011). Mapping between acoustic and articulatory gestures. *Speech Communication*, 53(4):567–589.

Atal B S, Chang J J, Mathews M V and Tukey J W (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555.

Eklund R, Peters G, Ananthakrishnan G and Mabiza E (2011). An acoustic analysis of lion roars. I: Data collection and spectrogram and waveform analyses. In *Proc. Fonetik 2011*, this volume. Stockholm, Sweden.

Fant G (1970). *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*. Mouton De Gruyter.

Fitch W T (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *J Acoust Soc Am*, 102:1213–1222.

Harris T, Fitch W, Goldstein L and Fashing P (2006). Black and white colobus monkey (*Colobus guereza*) roars as a source of both honest and exaggerated information about body mass. *Ethology*, 112(9):911–920.

Hauser M D (1993). The evolution of nonhuman primate vocalizations: effects of phylogeny, body weight and social context. *American Nature*, 142:528542.

Maeda S (1979). An articulatory model of the tongue based on a statistical analysis. *J of Acous Soc Am*, 65:S22.

Peters G (1978). *Vergleichende Untersuchung zur Lautgebung einiger Feliden (Mammalia, Felidae)*. Zoologische Staatssammlung M ”unchen.

Taylor A M and Reby D (2010). The contribution of sourcefilter theory to mammal vocal communication research. *Journal of Zoology*, 208:221–236.

Titze I R (1994). *Principles of vocal production*. Englewood Cliffs: Prentice-Hall.

Vannoni E, Torriani M and McElligott A G (2005). Acoustic signaling in cervids: a methodological approach for measuring vocal communication in fallow deer. *Cognition, Brain, Behavior*, IX:551–565.

Viviani P and Terzuolo C (1982). Trajectory determines movement dynamics. *Neuroscience*, 7(2):431–437.

Weissengruber G, Forstenpointner G, Peters G, Kübber-Heiss A and Fitch W T (2002). Hyoid apparatus and pharynx in the lion (*Panthera leo*), jaguar (*Panthera onca*), tiger (*Panthera tigris*), cheetah (*Acinonyx jubatus*) and domestic cat (*Felis silvestris f. catus*). *Journal of anatomy*, 201(3):195–209.

A comparative acoustic analysis of purring in four cats

Susanne Schötz[1] *& Robert Eklund*[2,3,4]
[1] *Humanities Lab, Centre for Languages and Literature, Lund, Sweden*
[2] *Voice Provider, Stockholm, Sweden*
[3] *Department of Cognitive Neuroscience, Karolinska Institute, Stockholm, Sweden*
[4] *Department of Computer Science, Linköping University, Linköping, Sweden*

Abstract

This paper reports results from a comparative analysis of purring in four domestic cats. An acoustic analysis describes sound pressure level, duration, number of cycles and fundamental frequency for egressive and ingressive phases. Significant individual differences are found between the four cats in several respects.

Introduction

The domestic cat is one of the most popular pet animals in the world, and virtually everyone is familiar with its trademark "purring" sound. Contrary to what might be believed, it is not known exactly how purring is produced, and there is a surprising lack of studies of purring, even descriptive.

This paper compares a number of acoustic characteristics of purring in four domestic cats, with focus on sound pressure level, duration, number of cycles and fundamental frequency of ingressive and egressive phases.

The domestic cat

There are 35 to 40 felid species in the world today (Sunquist & Sunquist, 2002), and the domestic cat (*Felis catus*, Linneaus 1758) is by far the most well-known and common cat with an estimated number of 600 million individuals (Driscoll et al., 2009). It was long suggested that the cat was first domesticated in ancient Egypt around 3600 years ago, but it is now believed that domestication took place 10,000 years ago in the Fertile Crescent. The closest relative of the domestic cat is considered to be the African wildcat (*F. silvestris lybica*) (Driscoll et al., 2007; Driscoll et al., 2009). Today around 60 breeds of domestic cats are recognized (Menotti-Raymond et al., 2008).

Although varying considerably in size and weight, a domestic cat normally weighs between 4 and 5 kilos, and is around 25 centimeters high and 45 centimeters long. Males are significantly bigger than females, and are on average 20% heavier than are females (Pontier, Rioux & Heizmann, 1995).

Purring

As mentioned above, it is not known exactly how purring is produced, and the term as such has been used quite liberally in the literature. In a major review paper Peters (2002) employed a strict definition of purring as a continuous sound produced on alternating (pulmonic) egressive and ingressive airstream. Given this definition, purring is only found in the "purring cats" (i.e. all felids but the non-purring/"roaring cats" lion, tiger, jaguar, leopard; whether or not the non-roaring snow leopard can purr remains unsettled) and in the Genet.

A number of different purring theories are found in the literature. McCuiston (1966) suggested that purring was hemodynamic and that the sound consequently emanated from the bloodstream running through the thorax. This theory was proven wrong by Stogdale & Delack (1985). Moreover, both Frazer Sissom, Rice & Peters (1991) and Eklund, Peters & Duthie (2010) reported that purring maximum amplitude occurs near the mouth and nose. It has recently been suggested that purring "is caused by rapid twitching of the vocalis muscle, whereas the large pads within the vocal folds of *Pantherinæ* might impede rapid contractions of this muscle and thus make it difficult to purr" (Weissengruber et al., 2008:16; see also Weissengruber et al., 2002).

Contrary to what is often believed, cats do not exclusively purr when they are content, but also when they are hungry, stressed, in pain or close to dying, and behaviourists have suggested that the function purring serves is to signal that the cat does not pose a threat (Eldredge, Carlson & Carlson, 2008:297).

Previous research

There is a surprisingly small number or papers devoted to felid purring, and several of these papers are also impressionistic in character. One of the first papers exclusively devoted to purring the domestic cat was Moelk (1944), but the focus of her paper is a classification of different kinds of purr and how they are used, and no acoustic analysis is presented.

Frazer Sissom, Rice & Peters (1991) reported that domestic cats purr at a frequency of 26.5 Hz, while Eklund, Peters & Duthie (2010) reported the figure 22.6 Hz.

Remmers & Gautier (1972:359) reported that egressive phases in purring cats had a duration of 730 ms, while ingressive phases had a duration of 690 ms.

Data collection

Continuous calm purring was collected from the four domestic cats Donna (D; female, age 6 months, 3.0 kilos), Rocky (R; male, 6 months, 3.6 kilos), Turbo (T; male, 6 months, 3.6 kilos), and Vincent (V; male, 16 years, 5.2 kilos).

All cats were recorded in a quiet home environment using a Sony DCR-PC100E digital video camera recorder with an external Sony ECM-DS70P electret condenser stereo microphone. This microphone is small in size, and could easily be held close to the muzzle without scaring or disturbing the cat.

Figure 1 shows the microphone positions during the recording sessions with the four cats.

Videos are available at **http://purring.org**

Figure 1. The microphone positions of all four cats during data collection.

To be able to identify egressive and ingressive phases in the recorded audio files, the first author kept her hand on the side of the cats' chests during the recording session while saying the words "in" and "out" according to the expanding (in-breath) or collapsing (out-breath) rib cage several times during the recording sessions.

Method

Data post-processing

All videos were transferred to iMovie, and audio files (wav, 44.1 kHz, 16 bit, mono) of about 70 seconds for each cat were extracted with Extract Movie Soundtrack. The waveforms were normalised for amplitude with Audacity, and low-pass filtered copies were created with Praat (10–40 Hz, smoothing at 10 Hz). These copies were used together with the original normalised waveform, spectrogram and Praat's pitch analysis to facilitate manual segmentation and counting of respiratory cycles per phase.

Figure 2 shows an example of the manual segmentation in Praat.

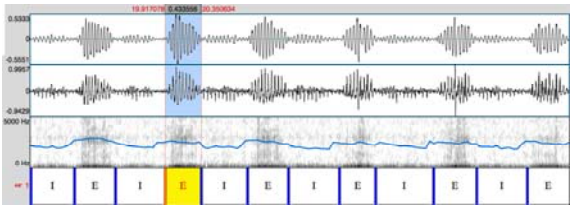


Figure 2. Manual segmentation of ingressive (I) and egressive (E) phases in Praat using the low pass filtered (top pane) and original (mid pane) waveforms as well as the original spectrogram and pitch contour (bottom pane).

The respiratory cycles per phase were labeled manually from the waveforms and counted with a Praat script. *Figure 3* shows an example of the procedure.

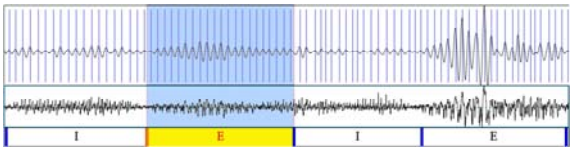


Figure 3. Manual labelling of cycles (pulses) per ingressive (I) and egressive (E) phases in Praat using the low pass filtered (top pane) and original (mid pane) waveform.

Egressive–ingressive identification

In order to ascertain that the egressive and ingressive phases were correctly identified, the parts of the recordings where the first author said "in" and "out" were located. Phases were then easily identified based on their distinct sound and waveform characteristics.

Analyses

Analyses were carried out with Praat. Statistics were calculated with SPSS 12.0.1.

Table 1. Summary Table. For all four cats results are given for sound pressure level (SPL), durations, cycles per phase, and fundamental frequency. Results are presented independently for egressive and ingressive phases, and statistical tests are performed on differences between egressive and ingressive phonation.

	Donna (D)		**Rocky (R)**		**Turbo (T)**		**Vincent (V)**	
Phonation type	Ingressive	Egressive	Ingressive	Egressive	Ingressive	Egressive	Ingressive	Egressive
No. phases analysed	39	39	40	40	61	61	61	61
Mean SPL (dB)	72.4	74.6	72.14	71.93	70.66	76.43	71.85	71.72
Mean SPL (dB) egr+ingr	73.48		72.03		73.52		71.78	
Standard deviation	0.8209	1.2974	0.9614	1.7693	1.96	3.20	1.0661	1.6260
Δ *t* test (paired-samples, two-tailed)	$p < 0.001$		$p = 0.427$		$p < 0.001$		$p = 0.426$	
Δ Wilcoxon (two related samples)	$p < 0.001$		$p = 0.249$		$p < 0.001$		$p = 0.224$	
Mean duration (ms)	673	587	819	756	604	511	511	484
Mean duration egr+ingr	632		788		558		498	
Standard deviation	120.80	82.70	169.23	130.05	58.90	45.09	85.10	69.72
Maximal duration	921	838	1038	997	773	634	719	614
Minimal duration	413	443	432	365	480	419	319	266
Δ *t* test (paired-samples, two-tailed)	$p < 0.001$		$p = 0.011$		$p < 0.001$		$p = 0.010$	
Δ Wilcoxon (two related samples)	$p < 0.001$		$p = 0.013$		$p < 0.001$		$p = 0.004$	
Mean no. cycles/phase	16.58	15.95	21.28	20.15	13.92	12.46	13.41	13.16
Mean no. cycles/phase egr+ingr	16.31		20.72		13.19		13.3	
Standard deviation	1.41	2.25	4.33	3.56	1.99	1.20	2.52	1.93
Maximal no. phases/cycle	22	22	29	28	21	15	18	17
Minimal no. cycle/phase	10	12	11	10	10	10	9	7
Δ *t* test (paired-samples, two-tailed)	$p = 0.178$		$p = 0.090$		$p < 0.001$		$p = 0.437$	
Δ Wilcoxon (two related samples)	$p = 0.132$		$p = 0.073$		$p < 0.001$		$p = 0.456$	
Mean fundamental frequency (Hz)	24.63	27.21	26.09	26.64	23.00	24.43	23.45	20.94
Mean frequency egr+ingr (Hz)	25.94		26.36		23.72		22.2	
Standard deviation	1.14	1.82	2.08	1.24	1.85	1.45	3.62	2.14
Highest fundamental frequency	27.5	33.2	33	29	27	28	28.8	24
Lowest fundamental frequency	21.6	24.2	23	24	19	20	18.2	17.1
Δ *t* test (paired-samples, two-tailed)	$p < 0.001$		$p = 0.174$		$p < 0.001$		$p < 0.001$	
Δ Wilcoxon (two related samples)	$p < 0.001$		$p = 0.067$		$p < 0.001$		$p = 0.002$	

Results

Summary results are presented in *Table 1* above.

I. Intracat analyses

We first analysed within-cat variation.

Amplitude

The normalised waveforms were used to extract the mean relative amplitude (SPL) in each ingressive and egressive phase for comparisons within each cat.

Mean relative SPL as derived from the normalised waveforms varied between 70.66 dB (T) and 72.4 (D) in the ingressive phase and between 71.72 (V) and 76.43 (T) in the egressive phase. For two of the cats (D/T), mean SPL was significantly higher in the egressive phases than in the ingressive ones, in contrast with Moelk (1944) and Peters (1981). However, no difference in mean SPL was observed for the other two cats (R/V).

Duration

Mean durations of the phases varied considerably between the four cats, ranging from 511 ms (V) to 819 ms (R) in the ingressive phase, and from 484 ms (V) to 756 ms (R) in the egressive phase.

Ingressive phases were significantly longer than egressive ones in all four cats, contrary to the results reported in Remmers & Gautier (1972:359).

Cycles per phase

The mean number of cycles per phase varied between 13.41 (V) and 21.28 (R) for ingressive phases and between 12.46 (T) and 20.15 (R) for egressive phases.

For all cats, the mean number of cycles per ingressive phase were higher than it was per egressive phase, thus replicating the results reported in Eklund, Peters & Duthie (2010) .

Fundamental frequency

All four cats showed fundamental frequencies that compare well to previous studies (Frazer Sissom, Rice & Peters, 1991; Eklund, Peters & Duthie, 2010). For the ingressive phase, mean F_0 ranged from 23.00 Hz (T) to 26.09 Hz (R), while the values for the egressive phase ranged from 20.94 Hz (V) to 27.21 Hz (D). Two of the cats (D/T) had significantly higher F_0 for the egressive phase as compared to the ingressive phase. One cat (V) showed the opposite pattern with significantly higher F_0 in the ingressive phase, while no significant difference was found for one cat (R).

II. Intercat analyses

Having performed within-cat analyses, we then turned to between-cat analyses. No intercat analyses of sound pressure level were performed since these were seriously affected by individual microphone positioning. All significance tests referred to are *t* tests (two independent samples, equal variances assumed, two-tailed).

Duration

All pair-wise comparisons revealed significant differences ($p < 0.001$) with the exception of T/V egressive duration ($p = 0.012$).

Cycles per phase

All pair-wise comparisons revealed significant differences ($p < 0.001$) with the exception of T/V number of ingressive cycles ($p = 0.305$) and number of egressive cycles ($p = 0.017$).

Fundamental frequency

All pair-wise comparisons revealed significant differences ($p < 0.001$) with the exception of D/V ingressive frequency ($p = 0.052$), T/V ingressive frequency ($p = 0.393$) and D/R egressive frequency ($p = 0.111$). With regard to combined fundamental frequency, all pairwise comparisons were significantly different with the exception or D/R ($p = 0.127$).

Discussion

To the best of our knowledge, this paper constitutes the first comparative and quantitative report of purring in domestic cats. As was the case in Eklund, Peters & Duthie (2010), previous research was both confirmed and contradicted. The lack of quantified reports in the literature makes far-reaching conclusions difficult, but our results hint at a certain degree of variation between individual cats in how purring is manifested, even if overall figures lie within the same general range.

Acknowledgements

Thanks to Gustav Peters for insightful comments.

References

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

Driscoll, C. A., M. Menotti-Raymond, A. L. Roca, K. Hupe, W. E. Johnson, E. Geffen, E. H. Harley, M. Delibes, D. Pontier, A. C. Kitchener, N. Yamaguchi, S. J. O'Brien & D. W. Macdonald (2007). The Near Eastern Origin of Cat Domestication. *Science* 317:519–523.

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, 2–4 June 2010, Lund, Sweden, 17–22.

Eldredge, D. M., Delbert G. Carlson & L. D. Carlson (2008). *Cat Owner's Home Veterinary Handbook*. Third edition. Hoboken, New Jersey: Wiley Publishing.

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

McCuiston, W. R. (1966). Feline purring and its dynamics. *Veterinary Medicine/Small Animal Clinician* 61:562–566.

Menotti-Raymond, M., Victor A. David, S. M. Pflueger, K. Lindblad-Toh, C. M. Wade, S. J. O'Brien & W. E. Johnson (2008). Patterns of molecular variation among cat breeds. *Genomics* 91:1–11.

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

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

Peters, G. (1981). Das Schnurren der Katzen (Felidae). *Säugetierkundliche Mitteilungen* 29:30–37.

Pontier, D., N. Rioux & A. Heizmann (1995). Evidence of selection on the orange allele in the domestic cat Felis catus: the role of social structure. *Oikos* 73(3):299–308.

Remmers, J. E. & H. Gautier (1972). Neural and Mechanical Mechanisms of Feline Purring. *Respiration Physiol*ogy 16:351–361.

Stogdale, L. & J. B. Delack (1985). Feline Purring. Compendium on the Continuing Education for the Practising Veterinarian 7(7):551-553.

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

Weissengruber, G. E., G. Forstenpointner, S. Petzold, C. Zacha & S. Kneissl (2008). Anatomical Peculiarities of the Vocal Tract In: H. Endo & R. Frey (eds.); *Anatomical Imaging*. Tokyo: Springer, chapter 2, 15–21.

Weissengruber, G. E., G. Forstenpointner, G. Peters, A. Kübber-Heiss & W. T. Fitch (2002). Hyoid apparatus and pharynx in the lion (*Panthera leo*), jaguar (*Panthera onca*), tiger (*Panthera tigris*), cheetah (*Acinonyx jubatus*) and domestic cat (*Felis silvestris* f. *catus*). *Journal of Anatomy* 201:195–209.

Imitation of bird song in folklore – onomatopoeia or not?

Åsa Abelin
Department of Philosophy, Linguistics and Theory of science, University of Gothenburg

Abstract

There are a number of expressions in Swedish and other languages, which describe the songs of different birds, e.g. the rose finch imitated as saying "Pleased to see you". These folk rhymes seem to both imitate the birds' songs and to describe some content connected to the bird. Swedish folk rhymes for the songs of different birds were analyzed in terms of sound structure and content. Imitation of the bird songs is reflected in both vowels and consonants of the folk rhymes, e.g. in speech sounds with energy on low frequencies, such as [m], [u], [o], or speech sounds with energy on high frequencies, such as [s], [t], [i]. Vibrant sounds are often transformed into [r]. Imitation also conserves the number of "syllables", i.e. the rhythmic structure of the bird songs. Intonation variation is also transformed into words like "falling", "down", "up". A special case of transformation of intonation variation is seen when the melody of the bird is interpreted as emotional or attitudinal, and transferred into words like "snälla" (please) and "fy" (shame on you).

The creation of folk rhymes could be seen as a type of folk etymology. In folk etymologies an incomprehensible series of sounds is heard and interpreted in terms of already existing words that fit the context, in the language one knows. Folk etymologies can be seen as a productive force in language development. A non-arbitrary connection between sound and content can aid memory and facilitate language learning.

Introduction

There are a number of expressions in Swedish and English, and probably in many other languages as well, that describe the songs of different birds. An example in English is the rose finch described as saying "Pleased to see you". A Swedish example is the chaffinch (bofink) saying "Snälla, snälla mamma får jag gå på bio ikväll klockan tio, klockan tio" (Please, please mom can I go to the movies tonight at ten, at ten) or "Trilla nerför trappan – nu är jag här"[1] (Fall down the stairs - now I'm here[2].) The Swedish ornithologist association http://www.sofnet.org/ has made a collection of these folk rhymes for birds.

There are different types of bird song: fixed pattern (the great tit), combination of patterns (the blackbird) and imitating (the parrot). The ones that have folk rhymes are mostly the birds with a fixed pattern.

The question is now what the origin of these folk rhymes for bird songs is. Is there onomatopoeia involved or are these folk rhymes solely invented from cultural beliefs about the birds and other associations? If we imitate – what is it that we imitate? Or do we hear what we want to hear when the signal is unclear to us, and, if so, what is it we want to hear in the case of birds?

Method

Approximately 130 Swedish folk rhymes for bird song were studied and a part of these were analyzed in detail according to sound structure and content. The folk rhymes were paired with recordings of the corresponding birds and a preliminary analysis in terms of phonemes, features and number of syllables of the rhymes was compared to an auditory analysis of sounds, number of syllables and intonation curves of the birds' songs. Exactly what the properties of the bird song that are perceived as syllables are, will

[1] Due to limited space the examples will not be written in IPA.

[2] The English translations are rough approximations to the wording of the rhymes, which sometimes have an uncommon grammar in Swedish.

not be analyzed here. An initial analysis of the content of the folk rhymes was made.

Examples of some of the folk rhymes are:

Domherren: Jul, jul, jul. Snö, snö, snö
(The bullfinch: Christmas ..., Snow, ...)
Skogsduvan: Gå då!
(The stock dove: Then go!)
Turkduvan: Men gå då!
(The collared dove: But please go!)
Ringduvan: Men gå då, ändå!
(The ring dove: But please go, anyway!)
Gransångaren: Salt sill, salt sill, salt sill
(The chiffchaff: Salt herring, ...)
Hönan: Ägg, ägg, upp i tak
(The hen: Egg, egg, up in the ceiling).

Analysis

The great tit

The sounds in rhymes for the disyllabic great tit (saying Edit! (a name); whiskey, whiskey, whiskey; tittut-tittut-tittut-tittut (peekaboo-)) are the following. The rhymes contain a majority of sounds with energy at high frequencies, e.g. [s], [t], front vowels with high F2, e.g. [i] (acute vowels and consonants in Jakobson's terminology).

The rhymes of the trisyllabic great tit (saying Här ska såås, här ska såås (We're going to seed), Var är du? (Where are you?), Kyss en skit! (Kiss a shit), Vintern tö, vintern tö (Winter thaw), Titta hit titta hit, jag är gul och svart och vit (look here look here, I am yellow and black and white) have the same type of sound inventory as the bisyllabic bird but these rhymes are trisyllabic.

The choice of words for the great tit are for example connected to early spring, when the bird appears again after winter, or to the colours of the bird. Other associations could be personal. But the words of the rhymes have been chosen to fit the sounds of the bird.

The chaffinch

The rhymes for the chaffinch (bofink) also have sounds on high frequencies as well as preserving the amount of syllables in the bird's song. Some other features are cross representational: the begging intonation of the chaffinch (starting high and then falling with modulations in the end) is translated into "Snälla, snälla" (please) or simply transformed from the falling intonation into the phrase "Trilla nerför trappan – nu är jag här" (Fall down the stairs – now I'm here.) Intonation becomes words.

The words chosen in rhymes for the chaffinch are thus, in these cases, describing the intonation of the bird's song.

The willow warbler

The song of the willow warbler (lövsångare) is similar to the long falling and varying melody of the chaffinch, and is interpreted as "Och vi som hade det så bra och så blev det så här" (We who had it so good and yet it became like this). The falling, complaining melody is transferred into a complaining sentence.

The doves

Another type of bird song is represented by the doves, the stock dove (skogsduva), collared dove (turkduva) and the ringdove (ringduva). The stock dove is said to call Gå då! (Then go!), Så kom (Come then), Skogis, skogis, skogis (Woody, woody), Ove, Ove, Ove (a name), Du sju (You seven), Ja tu (Me two), Du du som tog mina sju sju (You you, that took my two two). The rhymes are disyllabic just as the bird's call, and they are imitating the low frequency (grave) character by using vowels with energy on low frequencies, e.g. [u] and [o]. The consonants are more grave than for the finches. One rhyme associates to the woods (skog), which fits well with both the bird's sounds and the environment in which the bird lives.

The collared dove is trisyllabic and the rhymes are similar to the rhymes of the stockdove: Men gå då (But please go), Så kom då (So come then); Så gå då (So please go), Turkiet, Turkiet, Turkiet (Turkey), Kom Josef, kom Josef (Come Joseph). The latter associations go to the origin of the dove and are chosen so that they fit the dove's song.

The ringdove uses 5 syllables, where the second is usually stressed, saying: Men gå då, ändå, (But please go), Men gå då nån gång (But please go some time), Så kom då, ändå (So come then, anyway), Min älskling du är (My darling you are), Men ja har ju två, men ja har ju två (ringar) (But I have two, but I have two (rings)), Men RING då nån gång (But ring me sometime), Ja haar ju en ring (I do have a ring), Du tog sju för tu, din tjuv (you took seven for two, you thief), Jag vill ha smörgås, jag vill ha smörgås (I want a sandwich), Du tog min hustru

du (You took my wife, you), Du är tokig, du är dum (Your are crazy, you are stupid). The sound analysis is the same as for the other doves. There are many words with sounds which have energy on low frequencies as [o] in "gå" and "då". (There are also a lot of words with [ʉ] as in "du", "sju", "tjuv", which has energy also on high frequencies but which phonologically often counts as a back vowel.) There are also many nasals. The contextual associations go to the ring-like pattern on the neck of the dove, but there is also one rhyme using the homonymy of the word ring. The vowel of the word "ring" does not have energy on low frequencies, and so the need for a certain association sometimes overrides onomatopoeia in the choice of words.

The yellowhammer

Yet another interesting bird is the yellowhammer (gulsparv). It has a song with seven syllables and this is generally imitated in the rhymes: Sitt sitt sitt sitt sitt å skiiit (sit, ... and shit), Se se se se se se shiiit (see, ... shit), Vi-vi-vi-vi-visingsööö (Visingsö is the name of an island), 1 2 3 4 5 6 sjuuuu (seven), Fy på dig Nisse lille fyyyy (Shame on you little Nisse, shame on you), Nu är sommaren snart slut (Now the summer is soon over), Ett två tre fyra fem sex sjuuu, jag är liten jag är guuul (one, two, three, four, five, six, seven, I am little I am yellow), The sounds of the rhymes are generally on high frequencies, e.g. [i], [s]. The scolding content in the end of some of the rhymes, as in skit (shit), fy (shame on you) could be due to the outdrawn coarse voice quality in the end of the bird song, a type of emotional prosody that is interpreted by the listener and becomes transferred into words.

The common snipe

The [r] is imitated in words chosen for the rhymes imitating the common snipe (enkelbeckasin) in Herrarna, herrarna, herrarna (the men/boys), Grebba lilla, grebba lilla (little woman/girl). The common snipe has a distinctly vibrating call. Other words for men and women could have been used, but the onomatopoeic words, containing [r], were the ones chosen.

The chiffchaff

The chiffchaff (gransångare), in English having a name imitating its song, is going on with its Salt sill, salt sill, salt sill (salt herring). The clear sound of the rhyme is the [s], a sound on high frequencies.

The swallows

The different swallows with their long twitters are usually described as telling long stories, often about Virgin Mary. The sound imitiation is not obvious.

Comparison between languages

Comparing the different rhymes of the yellowhammer (gulsparv) shows some similarities between Swedish and English sound codification in folk rhymes. In English the bird has e.g. the following variations: Give me bread and some more cheeese, Bread and butter but no cheese. The structure which is similar to the Swedish rhymes is the extended final [i] and the rhymes having seven syllables.

The rhymes for the rose finch in English are Pleased to see you, Glad to see you, etc and in Swedish the rhymes are Se video, Köp en video, Skit i de du. In both languages the rhymes have 4 syllables and they end with [u]. [i] is also present in the rhymes, while the consonants are varying more between English and Swedish.

The content of the rhymes are different for these two birds.

Imitating sounds in making other sounds

Another way of imitation is described for the wood warbler (grönsångare): take a two-crown silver coin, spin it on a marble table and you get the falling little song (trudelutt) of the wood warbler. And for the dunnock (järnsparv): stir your fingers amongst iron pennies (ettöringar) and you get the song of the dunnock.

Conclusions and discussion

The sounds (vowels and consonants) of the rhymes for bird song seem to divide birds into birds with high pitched songs, imitated with e.g. [s], [t] and [i] and those with low pitched songs, imitated with e.g. [o], [u], [ʉ], nasals and voiced consonants. Vibrating sounds are imitated as [r]. The number of "syllables" in the bird song are often imitated with great precision. The intonation variation is sometimes transferred into words (falling down the stairs). The rising pitch of the hen is transformed into words: ägg, ägg, upp i tak (egg, egg, up in the ceiling).

A special case is when the meaning of the intonation or prosody sounds emotional and it

gets transferred into words (snälla, fy). This is the same phenomenon as in human expression of emotions; you can choose words or prosody for expression of emotions.

The choice of words, usually delimited by the need for certain sounds, is focussed on certain areas of common life, e.g. food, seasons, social relations (the rooster saying: upp alla nu, klockan är sju (up everyone now, it's seven o' clock)). Another common association is the physical appearance of the bird. On the other hand, many of the traits of folklore in general and for birds in particular might be detected in the different rhymes; the importance of the unusual, associative thinking, the importance of the first and the last, the importance of the meeting, the part in relation to the whole (cf. Tillhagen, 1977).

The creation of folk rhymes can be seen as a kind of folk etymology. In folk etymologies an incomprehensible series of sounds is heard and these sounds are interpreted in terms of already exisiting words that fit the context. Folk etymologies can be seen as a productive force in language development and it is common in children.

The salient sound properties seem to be rhythmic variations, high vs low frequency sounds and vibrant sounds. Crossrepresentential transformations from intonation to words or phrases are seen.

In general, the songs of the birds are imitated quite clearly, so onomatopoeia is an important factor in the creation of these folk rhymes. The choice of words can be seen as secondary; the contents associated to a certain bird can be expressed with different words but the words chosen are the most onomatopoeic.

The folk rhymes for bird songs is yet an example, albeit a small one, of motivated expressions in language and the interaction between sound and context in creating expressions. As shown in other studies (e.g. Kovics et al, 2010), a connection between sound and content can facilitates language learning. Iconicity is part of this (the connection between bird song and language sounds) and metonomy is another part (choosing imitating words that are appropriate in the physical or social context.) The creation of folk rhymes for bird song is partly a mnemonic trick to learn the different songs of birds, partly an irresistible process to interpret meaning when one listens to bird song.

References

Kovics V, Plunkett K, Westermann G (2010) The shape of words in the brain. *Cognition*, 114: 19–28

Tillhagen C-H (1978). Fåglarna i folktron. LTs förlag, Stockholm.

http://www.garden-birds.co.uk

http://www.sofnet.org/

http://sverigesradio.se/sida/gruppsida.aspx?programid=3275&grupp=6024

Articulatory modeling and front cavity acoustics

Björn Lindblom, Johan Sundberg, Peter Branderud & Hassan Djamshidpey*
*Department of Linguistics, Stockholm University and *Department of Speech, Music and Hearing, KTH Stockholm*

Abstract

Formant measurements were made on a physical replica simulating lip spreading and articulations with raised tongue blade. The data were successfully modelled numerically using traditional 2-D cross-sectional area functions. We found that spreading terminates the front cavity 'early' at the retracted mouth corner but adds a length correction representing the anterior residual part of the front cavity . For a raised tongue blade front cavity areas are incremented by amounts that depend on an interaction between the volume of the subapical space and the degree of lip opening.

The problem

In this contribution we address two outstanding problems in research on front cavity acoustics.

The first is the issue of determining where the vocal tract ends acoustically. Certain speech sounds are produced with retracted mouth corners. Typical examples are found among spread vowels such as [i], [e] and [ɛ] in which the spreading action gives the lip opening a curved surface and shifts the mouth corners to a point that is often located further back than the intersection between the front teeth and the midsagittal plane.

This situation creates a problem for models that represent the vocal tract as a series of cross-sections of variable areas and lengths: *Where does the vocal tract end acoustically?*

The second problem arises in the treatment of articulations with a raised tongue tip and blade. This gesture creates a space under the blade that is known to have an effect on the formant pattern under certain conditions (Stevens 1998).

The goal of this paper is twofold: (i) to present some new experimental data obtained from measurements on physical replicas of the vocal tract and (ii) to suggest principled ways of dealing with the two issues in the context of articulatory modeling.

Termination of the vocal tract

The first topic takes us back to a previous study in which spread and non-spread conditions were simulated with cylindrical tubes with & without notches. It was found that the formants of notched tubes could be accurately matched by the formants of unnotched tubes given appropriate equivalent lengths (Lindblom et al 2007).

HACMOD - a vocal tract replica

To study 3-D lip geometry under somewhat more realistic conditions we developed HACMOD. This is a hybrid device combining a 3-D replica of the front part of the vocal tract with the use of a copper tube, or plexiglass washers (Sundberg et al 1992). To create this model impressions and casts were first made of a human subject's anterior oral anatomy. From these casts, acrylic models of the jaws were then produced.

The formant frequencies of these resonators were measured by means of the STL ionophone whose sine-wave output was varied manually in frequency to identify the resonance peaks (Fransson & Jansson 1971). Special care was taken to achieve an airtight seal between the ionophone rod and the tube.

The modeling combined the HACMOD replica with various lip and back cavity configurations. Fig 1 shows a schematic of the set up used in the spreading experiment: a) Play dough lips (0.5 cm thick, manually varied opening area); b) HACMOD front cavity (2.5 cm deep; jaw opening [rel to clench] of 2.4 cm; mouth floor raised to the level of the lower teeth by filling lower jaw with lab putty); c) A cylindrical copper tube (inner diameter 2 cm, length 14 cm) was used as 'back cavity'.

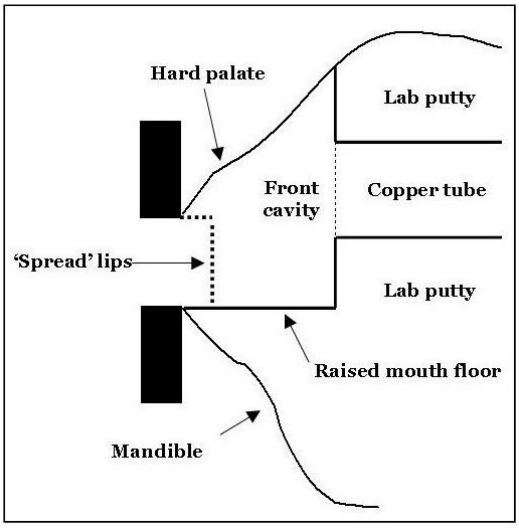


Figure 1. Set up used to simulate various degrees of 'spread lips' using HACMOD. The opening of the 'lips' is variable. The 'front cavity' is fixed consisting of HACMOD's upper and lower jaws of. A 'back cavity' is formed by connecting HACMOD with a copper tube.

To achieve a firm and air-tight tube-HACMOD connection, we attached a round plexiglass washer with a hole matching that of the tube's outer diameter at the posterior end of HACMOD which has a metal plate with a circular space for that purpose (see Figure 4C). The copper tube was then moved through this hole until 3 cm of its length was inside HACMOD. The front of the tube was plugged and the entire oral cavity was filled with lab putty. When the lab putty had hardened, a perpendicular cut was made at the anterior end of the copper tube and the plug was removed. In the experiments we used the posterior part of the lab putty as support for the copper tube.

These steps created a 3-D front cavity geometry with the upper teeth and palate at the top and a mouth floor flush with the lower teeth below. The back end of this front cavity was the perpendicular wall of the lab putty.

The lip opening was made rectangular with a vertical separation constant at 2.2 cm and width varying between 1 and 6 cm (uncurved). After the lip sheet had been placed on the surface of HACMOD those values changed and we ended up with a series of curved lip areas ranging from 3 to 21 cm^2. The retraction of the mouth corners was measured from the the surface of the lips in the midsagittal plane. This measure (the notch depth) ranged from 0 to 3 cm.

Spread lips and the acoustic termination of the vocal tract

To compute expected formant patterns and to address the issue of the termination of the vocal tract we made the following exploratory assumptions: (i) the area function for a spread vowel is generated in the traditional way up to the point where the mouth corner is located. (ii) the volume of the front cavity located anterior to the mouth corner is used to derive a length correction which is applied to the last section of the oral area function. It is calculated as the anterior front cavity volume derived by the area of the last section of the area function.

In accordance with these considerations area functions were generated and formants were computed for the entire set of measured configurations.

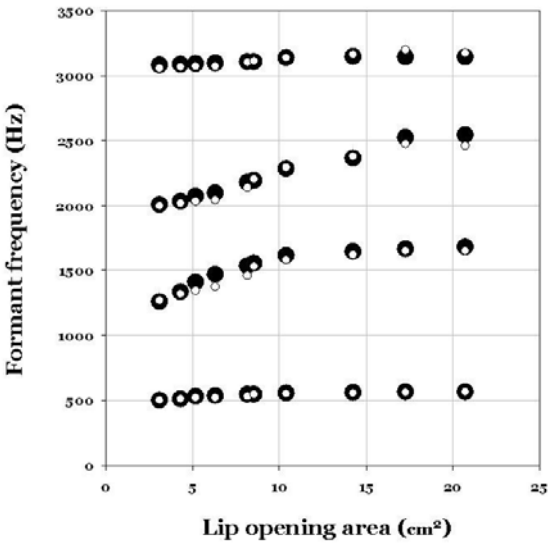


Figure 2. Measured and calculated formant frequencies for measured (solid dots and white circles respectively) as a function of lip opening area i.e., increased retraction of the mouth corners.

Length corrections were computed from estimates of front cavity volumes anterior to the mouth corner and from the last cross-section of the 'oral' area function. Curved lip areas were used as the last section of each area function. These tables were fed into WORMF, a program written by Johan Liljencrants and based on Liljencrants & Fant (1975). The results are presented in Figure 2.

Spreading increases F2 and F3 which show a clear dependence on the front cavity. F1 and F4 remain stable. The agreement between measured and computed values testifies to the the realism of the present proposal.

Role of the subapical cavity

The role of the sublingual space was studied using the set up shown in Figure 3. As in the spreading experiment we combined HACMOD with the cylindrical copper tube serving as the 'back cavity'.

A lab putty tongue conguration was fitted into the lower jaw component. To make it we made a cast of HACMOD's entire oral cavity. A cut perpendicular to the dental plane was made 2.5 cm from the upper front teeth. At the back another cut was made to imitate a raised blade. It created a space that began 7 mm behind the first cut and smoothly ended at the lower edge of the copper tube. The 'tip' was slightly truncated resulting in a constriction that was subsequently found to give a 0.2 cm^2 opening, not atypical at the moment of the release of a postdental stop.

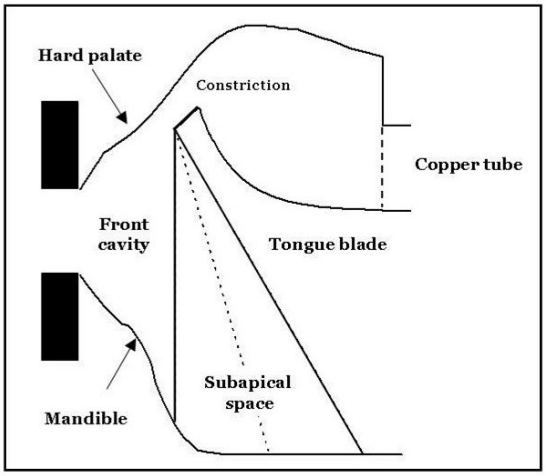
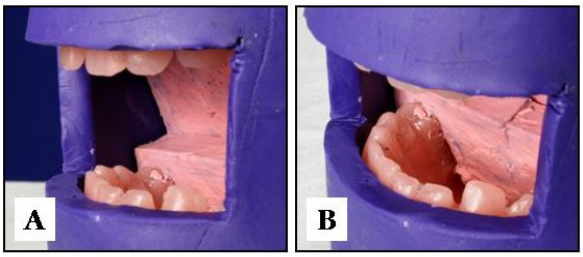


Figure 3. Simplified schematic illustrating the components used to investigate the effect of the subapical space. As in the first experiment we used a copper tube ('back cavity'), HACMOD's hard palate, teeth and mouth floor, and play dough lips. A raised tongue blade was carved out of lab putty material to create three different sublingual conditions.

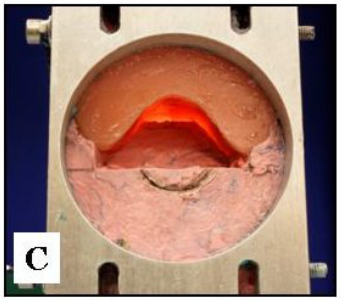
By weighing the pieces that had been removed and dividing their weights by the density (1.54 g/cm^3) of the lab putty, we were able to infer the volumes of all relevant cavities: The front space anterior to the constriction (including the upper jaw, the interdental space plus lower jaw space) had a total volume of 21.7 cm^3; The length of this front section [measured from front upper teeth to the constriction along the dental plane] was 2.5 cm. The post-constriction space was 2.3 cm long and had a volume of 8 cm^3. The back cavity was 13 cm long and had a cross-sectional area of 3.14 cm^2.

Figures 4 shows photographs of this set up. Panel A presents the front part at an angle that reveals the rectangular lip opening and the inside of HACMOD's right cheek. The underside of the 'tongue blade' shows up as a slant structure that connects to a narrow terrace-like part. From there a perpendicual cut down to the mouth floor. These details which differ slightly from the simplified schematic of Figure 3 were all taken into account in computing front cavity volumes.

In panel B we see another front view looking down into the space of the lower jaw.



Figures 4 A & B: Front views of set up.



Figures 4 C: Rear view of set up.

Panel C shows the rear view of HACMOD's metal plate, HACMOD's acrylic 'soft palate' contour and the lab putty forming the surface of the tongue. A light source was shone on the constriction and the hard palate.

The copper tube was fastened with the aid of a plastic washer inserted into HACMOD's circular space. Special care was taken to ensure that all surfaces in contact were tightly sealed.

Figures 5 and 6 summarize the results of this experiment. In Figure 5 the effect of varying the degree of spreading is presented. Figure 5 pertains to the condition of a front cavity with a perpendicular back wall. In other words no subapical cavity is present. Spreading is quantified in terms of retraction of the mouth corner (notch depth). F1, F2 and F4 stay relatively stable whereas F3 is clearly sensitive to the spreading gesture. [cavity affiliations: F1 whole system; F2 & F4 back cavity; F3 front cavity].

The measurements were compared with formant frequencies computed by applying the proposed spreading rule (thin lines) which assumes that the front cavity is terminated at the mouth corner with a length correction on the last cross-section. It is noteworthy that, whereas the calculations predict a front cavity resonance also for lip larger openings, no such formant could be observed in the measurements.

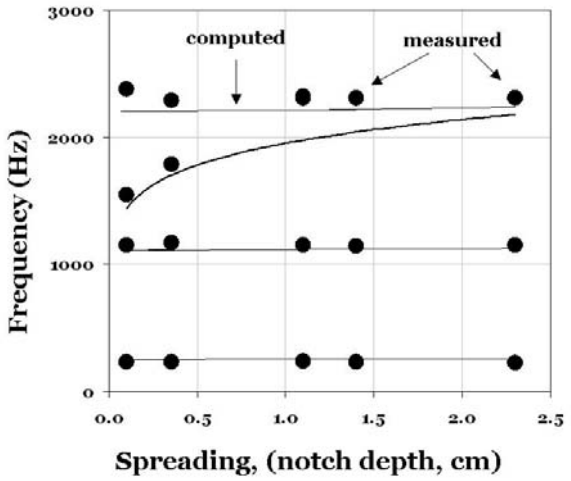


Figure 5. Formant data for condition with the front cavity's back wall perpendicular. The thin lines show calculated values.

Figure 6 presents a close up of the effect of spreading and the subapical cavity on F3 (the front cavity resonance). Only F3 was found to vary. F1, F2 and F4 were omitted showing no change. The comparison clearly demonstrates that F3 depends significantly both on spreading and the size of the subapical cavity.

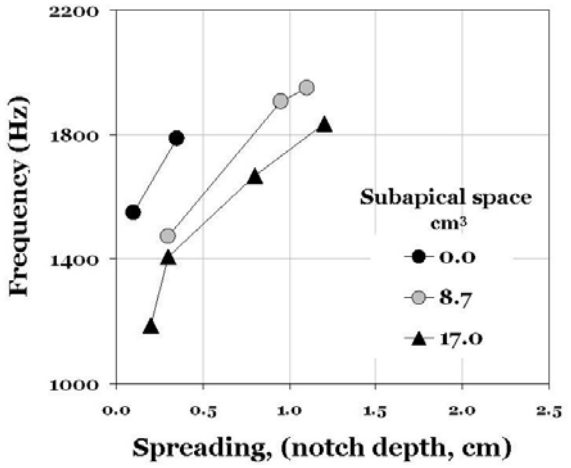


Figure 5. A close up on F3 (the front cavity resonance) which shows a clear dependence on both the size of the subapical cavity and the degree of spreading.

The data were modeled by computing formant frequencies from area functions obeying the spreading rule (front cavity terminated at the mouth corner with a length correction on the last cross-section) and with front cavity area increment derived from the volumes of the subapical space. The predicted values were systematically too low. We take that result as indicating that the acoustic role of the subapical cavity depends crucially on impedance conditions at the lips.

Summary

We found that spreading terminates the front cavity 'early' at the retracted mouth corner but adds a length correction representing the anterior residual part of the front cavity.

For a raised tongue blade, front cavity areas are incremented by amounts that depend on an interaction between the volume of the subapical space and the degree of lip opening.

References

Fransson F & Jansson E (1971). The STL-Ionophone: Transducer properties and construction *J Acoust Soc Am 58*:910-915.

Liljencrants J & Fant G. (1975): Computer program for VT-resonance frequency calculations, *STL-QPS*R 16(4):15-20. http://www.speech.kth.se/publications/

Lindblom B, Sundberg J, Branderud P & Djamshidpey H (2007): On the acoustics of spread lips, *Fonetik 2007*, *TMH-QPSR*, *50(1)*:13-16, http://www.speech.kth.se/publications/

Lindblom B, Sundberg J, Branderud P, Djamshidpey H & Granqvist S (2010). The Gunnar Fant legacy in the study of vocal acoustics, *10ème Congrès Français d'Acoustique*, Lyon, 12-16 Avril 2010

Stevens K N (1998): *Acoustic Phonetics*, MIT Press:Cambridge, MA.

Sundberg J, Lindblom B & Liljencrants J (1992): Formant frequency estimates for abruptly changing area functions: A comparison between calculations and measurements, *J Acoust Soc Am 91(6)*:3478-3482.

Age-related lip movement repetition variability in two phrase positions

Johan Frid[1], Susanne Schötz[1], Anders Löfqvist[2]

[1]Humanities Lab, Centre for Languages and Literature, Lund University
[2]Department of Logopedics, Phoniatrics and Audiology, Lund University

Abstract

This study examined the relationship between age and lip movement variability across repetitions of an utterance. We applied functional data analysis (FDA) to lip movement data of 15-20 repetitions of a short Swedish phrase from 37 Swedish speakers (19 females, 18 males, 5-31 years) collected with three-dimensional articulography. From each utterance, three different sub-units were extracted semi-automatically by locating consistent kinetic events in the lip movement functions. Results generally showed moderate negative correlations between age and amplitude variability. The longest possible sub-unit, given consistent kinematic events, showed the strongest correlation.

Introduction

This study compared age-related lip movement variability of Swedish speakers in three kinetically delimited slices of an utterance. A number of studies using acoustic analysis (e.g., Kent, 1976; Kent and Forner, 1980; Smith, 1978) and movement recordings (e.g., Sharkey and Folkins, 1985; Smith, 1995; Smith and McLean-Muse, 1986), have shown that lip movement variability across utterance repetitions decreases with age until adolescence. Some previous studies (Goffman and Smith, 1999; Sadagopan and Smith, 2008; Smith and Goffman, 1998) have used the spatiotemporal index (STI, Smith et al., 1995), which only provides a single metric of variability (cf., Lucero et al., 1997), incorporating both amplitude and phase. Others (Koenig et al., 2008; Lucero and Löfqvist, 2005) have used functional data analysis (FDA, Ramsay et al., 1996), where amplitude and phase variability are calculated separately. In earlier studies (Frid et al., accepted; Schötz et al., submitted), we have demonstrated that using kinematic landmarks to identify the speech segments to be analysed for variability showed more evident trends than using acoustic landmarks. We also found that the amplitude index of the FDA showed a higher age-related lip movement variability than the phase index of the FDA or the STI.

The purpose of the present study was to apply FDA to lip movements, and to compare different slices of the utterance. Our aim was to extend earlier findings of decreasing variability with age to see if utterance position affects speech movement variability. The long-term objective is to examine if children with atypical language development differ from typically developing children in terms of articulatory variability.

Method

To obtain as large lip movements as possible, we recorded the Swedish phrase *Mamma pappa barn* 'Mummy daddy children', which is short and can be spoken on a single breath. Lip movement data of 15-20 repetitions from 37 typically developed Swedish children and adults (19 females, 18 males, aged 5-31 years) were obtained along with a microphone signal using the Carstens Articulograph AG500. Sensors were placed on the upper and lower lip, and to correct for head movements also on the nose bridge and behind the right ear. Figure 1 shows the experimental set-up.

Landmark registration

Euclidean distances between the upper and lower lip sensors in three dimensions were calculated from the lip movement data, low-pass filtered at 25 Hz and used in the landmark registration. We delimited each token at consistent kinematic events using the first derivative of the distance function and located two points. To obtain four full cycles of opening-closing gestures of the lips, we set the onset point to the maximum velocity of the distance function in the opening phase during the transition from the first *m* to the first *a* in the

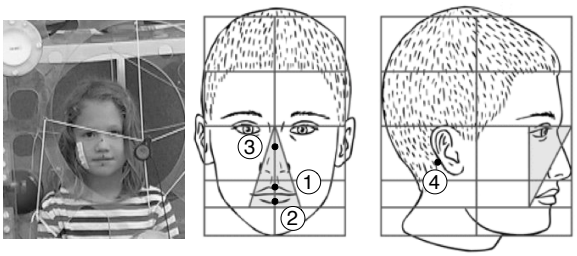


Figure 1: Experimental set-up with subject in the articulograph, and the sensor positions: upper and lower lip midsaggital on the vermilion border (1, 2), reference sensors on the nose bridge and behind the right ear (3, 4).

word *Mamma*. For the offset point we used the same transition from the *b* to the *a* in the word *barn*. An example of the kinematic landmark registration procedure environment is shown in Figure 2. Tokens with measurement errors or artefacts were excluded from further analysis.

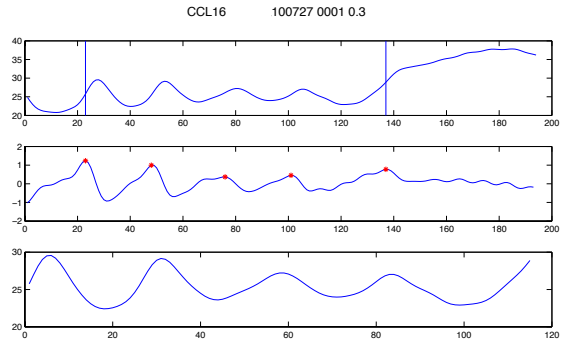


Figure 2: Lip distance function (top), its first derivative with marked velocity peaks (middle) and resulting trimmed and zoomed portion (bottom) of a token during kinematic landmark registration. The vertical lines indicate the positions of the onset and offset points described in the text above.

All tokens were further divided into two subsegments at the maximum velocity of the distance function in the opening phase during the transition from the first *p* to the first *p* in the word *pappa*. In the middle pane of Figure 2, this corresponds to the location of the third peak from the left. The sub-units contained two opening-closing cycles each. We will refer to the first sub-unit as **word1**, to the second as **word2**, and to the unit containing both words as **phrase**. Although these labels denote linguistics units to which the kinematic units are not aligned completely, we will still use them as a matter of convenience.

Functional data analysis (FDA)

The landmark delimited Euclidean distance functions were used as input to the FDA, a technique for time-warping 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), and Lucero and Löfqvist (2005). The procedure involves the following steps: (1) temporal normalisation of the signals from a number of tokens, (2) calculation of the mean signal, (3) alignment of individual signals to the mean signal using nonlinear time-warping, and (4) computation of one index of amplitude variability and one of temporal variability (phase). Each token was amplitude normalised by subtracting its mean and dividing by its standard deviation (see Koenig et al., 2008).

Results

In a previous study with the same data (Schötz et al., submitted), we found that amplitude variability showed a stronger correlation with age than phase variability. Therefore, we will only report the results for amplitude variability here. We analysed the relationships between the three speech units, the FDA amplitude index and age through correlations, scatterplots and linear regression models using the R statistical environment (R Development Core Team, 2011). FDA amplitude indices as a function of age for the three speech units are plotted in Figure 3, while Table 1 shows the statistical results of the correlation and linear regression analyses, including correlation coefficient, slope (β), significance level, coefficient of determination (R^2) and number of samples. The results show that age significantly predicted amplitude variability, and also explained a significant proportion of variance in amplitude variability in all the speech units.

	word1	**word2**	**phrase**
Correlation (*r*)	-0.49	-0.65	-0.66
β	-0.165	-0.279	-0.326
p	0.00219	< .001	< .001
R^2	0.24	0.42	0.44
n	37	37	37

Table 1: Results of correlation and linear regression analysis between age and the FDA amplitude variability index for the three speech unit conditions.

Paired-samples t-tests were conducted to compare the amplitude variability indices in the different

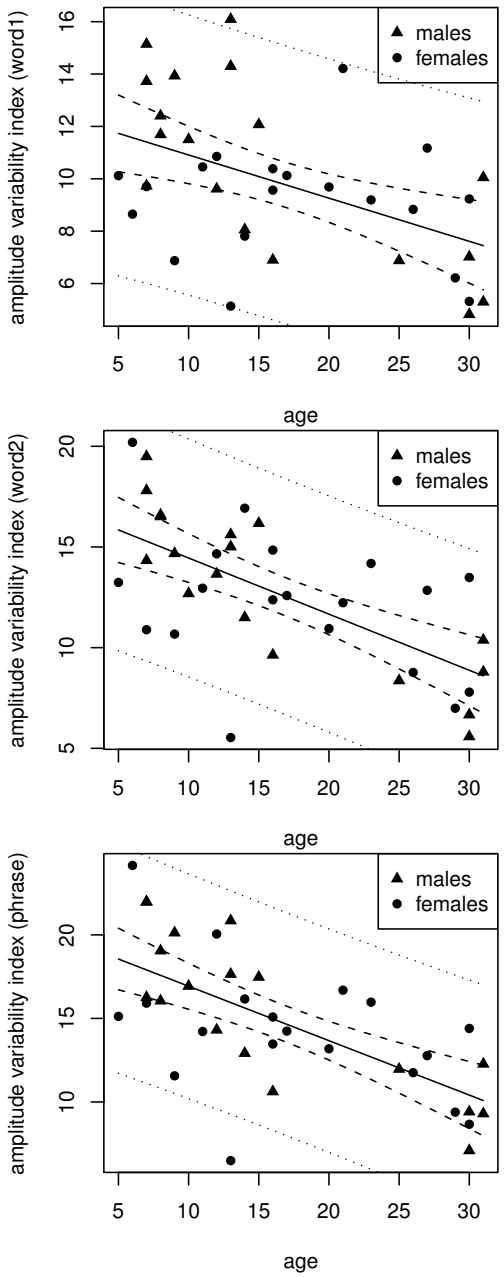


Figure 3: Amplitude variability as a function of age (solid lines), prediction intervals (dotted lines), and confidence intervals (dashed lines) for each of the three speech units conditions .

speech unit conditions. Means and standard deviations (SD) are given in Table 2. There was a significant difference in the scores for **word1** and **word2** conditions, t(36) = 6.49, $p < .001$. Furthermore, there were significant differences in the scores for **word1** and **phrase**, t(36) = 11.63, $p < .001$ and for **word2** and **phrase**, t(36) = 7.70, $p < .001$.

One of our motivations for splitting up the tokens into smaller units was to compare the results of each sub-unit with the result of the whole utterance. As the variability index scales differed in

	word1	**word2**	**phrase**
Mean	9.80	12.58	14.74
Median	9.70	12.85	14.40
SD	2.91	3.70	4.26

Table 2: Means, medians and standard deviations of the FDA amplitude variability index for the three speech unit conditions.

word1 and **word2**, they were rescaled using the means and standard deviations. We then calculated the correlation between age and the rescaled and combined amplitude variability indices. The two variables were correlated, $r(72) = -0.57$, $p < .001$, but the correlation was smaller than the one we obtained between age and **phrase** (-0.66).

Discussion and Future Work

The results for amplitude variability confirm the results of previous studies, i.e. that lip movement variability decreases with age. In (Frid et al., accepted) and Schötz et al. (submitted) we found higher correlations for amplitude than phase in both acoustic and kinematic landmarks. Koenig et al. (2008) reported the opposite pattern, with more variability for phase than amplitude. Those results were, however, based on records of airflow during fricative production, thus reflecting both articulatory and expiratory factors. The current results are based on articulatory movements alone. Similar developmental changes have been observed in non-speech motor activities such as reaching and finger tapping (Deutsch and Newell, 2003, 2004). The decrease of repetition variability with age is most likely due to a combination of factors. One factor may be cerebral and cerebellar development (Kent, 1976). Another one is practice, which leads to more stable motor performance. It is also likely that a developing and changing system will show increased motor variability during transitions, when a new mode of organisation is replacing an old one (Smith and Thelen, 2003).

In this study, the correlation between age and variability was almost the same for **word2** as for **phrase**, but weaker for **word1**. There are a few possible explanations for this. One is that **word2** has a longer duration than **word1**. The segmental content is also different: **word1** contains something like >amap< (one plosive and one nasal), wheras **word2** consists of the sequence >apab< (two plosives).[1] Another explanation may be that

[1]The 'greater than' and 'less than' are used to symbolise in- and outgoing transitions.

the phrase positions differ: **word1** is initial, while **word2** is medial in the phrase. It is possible that the initial position offers a potential anchoring point for articulation, and therefore obscures any age-related effects. It could also be that variability is revealed better in prominent words or sub-units. The utterance in this study was produced with a broad focus by all subjects, i.e. with the highest prominence on the final word *barn*, of which the *b* is included in **word2**.

Splitting up the utterance into sub-parts increased the sample size (by a factor of 2), but it did not yield a stronger relationship between age and amplitude variability. This is an interesting finding, which we would like to examine more thoroughly in the future. It would also be interesting to compare variability in different prosodic contexts. In further studies, we will also record not only more typically developed children, but also atypically developed children. Future work also includes an examination to see if children with atypical language development differ from typically developing children in terms of articulatory variability. We also want to examine the possible relationship of our results with cerebellar function as assessed by the blink reflex.

Acknowledgements

The authors 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. We are also grateful to J. Lucero for the use of his FDA MATLAB toolkit.

References

Deutsch K M and Newell K M (2003). Deterministic and stochastic processes in children's isometric force variability. *Dev Psychobiol*, 335–345.

Deutsch K M and Newell K M (2004). Changes in the structure of children's isometric force variability with practice. *J Exp Child Psychol*, 319—-333.

Frid J, Schötz S and Löfqvist L (accepted). Functional data analysis of lip movements: repetition variability as a function of age. In *Proc. of ICPhS XVII, Hong Kong, August 17-21, 2011*.

Goffman L and Smith A (1999). Development and phonetic differentiation of speech movement patterns. *J Exp Psychol: Hum Percept Perform*, 649–660.

Kent R (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *J Speech Hear Res*, 421–427.

Kent R and Forner L (1980). Speech segment duration in sentence recitation by children and adults. *J Phonetics*, 157–168.

Koenig L, Lucero J and Perlman E (2008). Speech production variability in fricatives of children and adults: Results of functional data analysis. *J Acoust Soc Am*, 3158–3170.

Lucero J and Löfqvist A (2005). Measures of articulatory variability in vcv sequence. *Acoust Res Lett Online*, 80–84.

Lucero J, Munhall K, Gracco V and Ramsay J (1997). On the registration of time and the patterning of speech movements. *J Speech Hear Res*, 1111–1117.

R Development Core Team (2011). R: A language and environment for statistical computing. Webpage: http://www.R-project.org, accessed on 11 Mar 2011.

Ramsay J, KG M, VL G and DJ O (1996). Functional data analysis of lip motion. *J Acoust Soc Am*, 3718–3727.

Sadagopan N and Smith A (2008). Developmental changes in the effects of utterance length and complexity on speech movement variability. *J Speech Hear Res*, 1138–1151.

Schötz S, Frid J and Löfqvist L (submitted). Lip movement repetition variability as a function of age: A comparison of two landmark registration strategies. In *Proc. of Interspeech, Florence, Italy, August 27-31*.

Sharkey S G and Folkins J W (1985). Variability in lip and jaw movements in children and adults: Implications for the development for speech motor control. *J Speech Hear Res*, 8–15.

Smith A and Goffman L (1998). Stability and patterning of speech movement sequences in children and adults. *J Speech Hear Res*, 18–30.

Smith A, Goffman L, Zelaznik H N, Ying G and McGillem C (1995). Spatiotemporal stability and patterning of speech movement sequences. *Exp Brain Res*, 439–501.

Smith B (1978). Temporal aspects of english speech production: A developmental perspective. *J Phonetics*, 37–67.

Smith B and McLean-Muse A (1986). Articulatory movement characteristics of labial consonant productions by children and adults. *J Acoust Soc Am*, 1321–1328.

Smith B L (1995). Variability of lip and jaw movements in the speech of children and adults. *Phonetica*, 307–316.

Smith L and Thelen E (2003). Development as a dynamical system. *Trends Cog Sci*, 343–348.

Exotic vowels in Swedish: a project description and an articulographic and acoustic pilot study of /iː/

Susanne Schötz[1], Johan Frid[1], Anders Löfqvist[2]

[1]Humanities Lab, Centre for Languages and Literature, Lund University
[2]Department of Logopedics, Phoniatrics and Audiology, Lund University

Abstract

This paper introduces the research project Exotic vowels in Swedish – an articulographic study of palatal vowels [VOKART]*, which aims at increasing the empirical knowledge of vowel production in general, and at extending our knowledge of the articulatory dynamics of palatal vowels in Swedish in particular. In a pilot study of the realisation of the vowel /iː/, we analysed articulatory and acoustic recordings of two male speakers of different regional varieties of Swedish– one South Swedish with regular [iː] and one East Central (Standard) Swedish with "damped" so called "Viby-coloured" [ɨː]. Results showed that [ɨː] was pronounced further back with an overall lower tongue position, but with a higher tongue tip position than [iː]. Acoustic analysis showed a lower F_2 for [ɨː] than [iː], indicating a more centralised vowel quality. These tentative results will be followed up with larger studies. In addition, one of the analysis tools being developed within the project is described briefly.*

General introduction

In a cross-language comparison, Swedish has a fairly rich vowel system with some particularly exotic and distinctive features. One such feature is that among the front, close vowels there are three contrastive long vowel sounds /iː, yː, ʉ̟ː/, characterised by a relatively small acoustic and perceptual distance, exemplified by minimal triplets such as /niː, nyː, nʉ̟ː/ ('you', 'new', 'now'). Specifically the contrast between /yː/ and /ʉ̟ː/, two similar but still phonemically distinct rounded vowels, is considered highly unusual and exotic among the world's languages. The acquisition of these two vowel sounds by Swedish children as well as adults learning Swedish typically presents a major difficulty (e.g., Johansson, 1973; Linell and Jennische, 1980). For these three vowels, the tongue articulation is assumed to be basically identical, but the documentation is incomplete.

Yet another exotic feature is the today fairly wide-spread realisation of /iː/ and /yː/ in Swedish with a "damped", "buzzing" so called "Viby-coloured" (named after the small town of Viby in Central Sweden) quality, generally phonetically transcribed as [ɨː]. This vowel quality has been recognised as a dialectal feature in several Swedish regions (Bruce, 2010), and is considered to be very rare among the world's languages (Ladefoged and Maddieson, 1996).

Phonetic investigations of vowels in different languages have been mainly acoustic (Ladefoged, 2003). Acoustic studies of Swedish vowels using formant frequencies include Fant (1959), Eklund and Traunmüller (1997), and Kuronen (2000). However, it does not seem possible to uniquely determine the underlying articulation of a vowel from its formant frequencies.

Object of study and goals

The general object of study of the project is the production of vowels, specifically the articulation of the Swedish long palatal vowels /iː/, /yː/, /ʉ̟ː/, /eː/ and /øː/. We will focus on three specific issues: (1) the crowding of vowels among the front, close vowels, particularly /yː/ and /ʉ̟ː/, (2) the diphthongisation of all five, long vowels, and (3) the special realisation of /iː/ and /yː/ vowels with a "damped" quality in contrast to the regular realisation of these vowel sounds. The major goal of our project is to describe and understand the articulation of Swedish palatal vowels, including their articulatory dynamics (diphthongisation). Furthermore, we will also elucidate the dialectal variation among Swedish vowels. For this purpose, we will restrict ourselves to studying vowels produced by speakers from the three metropolitan areas of Stockholm, Göteborg and Malmö, which represent East Central (Standard), West Central and South Swedish.

Material and method

Speech material containing vowels from at least 15 subjects from each of the three dialectal areas Stockholm, Göteborg and Malmö will be recorded. For the sake of completeness, we will record realisations of all Swedish vowel phonemes, and then focus our study on the palatal vowels. Articulographic and acoustic (16 kHz/16 bit) recordings of the vowels will be made in several consonantal contexts and different types of speech material. The recorded movements of each vowel will then be analysed using automatic methods. As indicated above, our focus will be on articulation, but also the formant frequencies (F_1, F_2, F_3) and their dynamics (possible diphthongisation) will be analysed and related to the articulatory trajectories. We will use the Carstens Articulograph AG500 to record the articulatory movements (at 200 Hz). This method is based on the principle that when a coil (sensor) moves inside a magnetic field, a voltage is induced in the coil, which is proportional to the distance between the coil and the transmitter coil generating the magnetic field. Articulography tracks movements in 3D of the discrete points where the sensors are attached to the tongue or to other articulators. We will record twelve sensors simultaneously. Figure 1 shows the positions of these.

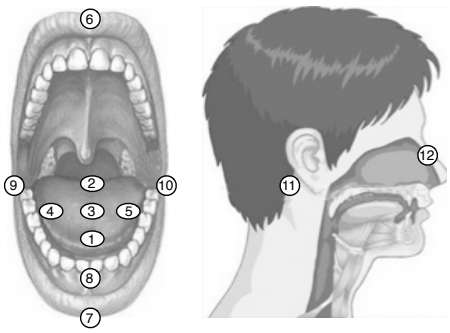


Figure 1: Positions of the 12 sensors to be used.

Our articulographic method is particularly suited for examining the assumed similarity in tongue position among the Swedish close front vowels. The magnitude of the lip opening (from larger to smaller) is regarded as the major distinctive feature for these vowels: unrounded (with spread lips) /iː/, outrounded /yː/ and inrounded /ʉ̟ː/ (Fant, 1959; Ladefoged and Maddieson, 1996).

Pilot study of two realisations of /iː/ in Swedish

Our interest in studying the damped realisation of /iː/ and /yː/ from a production point of view is related specifically to an old dispute represented in the Swedish linguistics and phonetics literature about its precise place of articulation (Bruce and Engstrand, 2006). The disagreement is about whether the point of major constriction for these vowel sounds, i.e. damped /iː/ and /yː/, is further front, as compared to their regular counterparts (front, close vowels), and basically alveolar (Noreen, 1903), or instead further back and rather central (Borgström, 1913). Moreover, the position of the tongue tip relative to the tongue dorsum in damped /iː/ and /yː/ has also been under debate (see e.g., Engstrand et al., 2000). It should be stressed that these views about the specific point of major constriction of these vowels are at best intelligent speculations, as adequate articulatory data seem to be lacking here. A fronted (alveolar) variant would seem to be more odd as a place of articulation of a vowel, while a retracted (central) variant would appear to be a vowel articulation which is less unusual and found in a fair number of the world's languages (Björsten et al., 1999; Engstrand et al., 2000). To learn more about the articulation and acoustics of the regular and damped realisations of Swedish /iː/, we conducted a small pilot study.

Data and method

Articulographic and acoustic recordings of /iː/ with two Swedish male speakers of different regional varieties – one South Swedish speaker with regular [iː] and one East Central Swedish speaker with damped [ɨː] – were made using the Carstens Articulograph AG500. Three repetitions of the two words /biːbel/ and /papiːpa/ with bilabial contexts and primary stress on the /iː/ vowel were recorded using a subset of the sensor positions shown in Figure 1. Sensors were placed on the tongue tip, tongue blade and tongue dorsum. We placed reference sensors on the nose bridge and behind the ear. Articulographic analysis was made with the MATLAB script Mview (Tiede, 2010), and Praat (Boersma and Weenink, 2011) was used in the the acoustic analysis.

Results

Articulographic analysis

The MATLAB script Mview enables examination of the data in three dimensions. Since the recorded sensors were aligned along the tongue on the same axis, we examined the tongue movement pattern at a midsagittal plane. We selected an articulatory measurement sample from one point in time from the steady-state portion of each vowel. Figure 4 shows the positions of the sensors at this time instant.

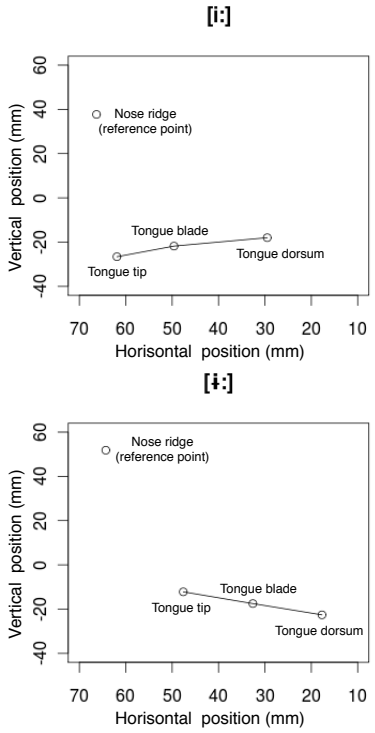


Figure 2: Midsagittal plots of four articulatory measurement sample points from one point in time of the steady-state portion of Swedish regular [iː] (top) and damped [ɨː] (bottom).

The tongue is positioned more forward in the mouth (relative to the nose bridge) for [iː] than for [ɨː]. Another difference is that for [iː] the tongue blade is lower than the tongue dorsum and the tongue tip is lower than any of the other two. For [ɨː] however, the height pattern is reversed; the tongue tip appears to be higher than the others. The overall position of the tongue also appear somewhat lower (again, relative to the nose bridge) for [ɨː].

Acoustic analysis

Table 1 displays mean values of the first four formant frequencies (F_1-F_4) of manually segmented steady-state portions of three repetitions each of regular [iː] and damped [ɨː]. F_1 appears to be almost identical in [iː] and [ɨː], while F_2 is 601 Hz higher in [iː]. F_3 and F_4 show slight differences; F_3 is 131 Hz higher in [iː], while F_4 is 282 Hz higher in [ɨː].

Figure 3 shows an F_1/F_2 plot with one Bark circles of regular [iː] and damped [ɨː], as well as the primary cardinal vowels pronounced by Daniel Jones (see e.g., IPA, 1999). [iː] appears to be quite similar to the second primary cardinal vowel [eː], i.e. a front close or close-mid vowel, while [ɨː] is positioned further back, like a central close or close-mid vowel.

Mean Fn (Hz)	[iː]	[ɨː]
F_1	332	330
F_2	2017	1416
F_3	2685	2554
F_4	3239	3521

Table 1: Mean formant frequencies (Hz) of the first four formants in the steady-state portion of [iː] and [ɨː] (three repetitions by one male speaker for each vowel realisation).

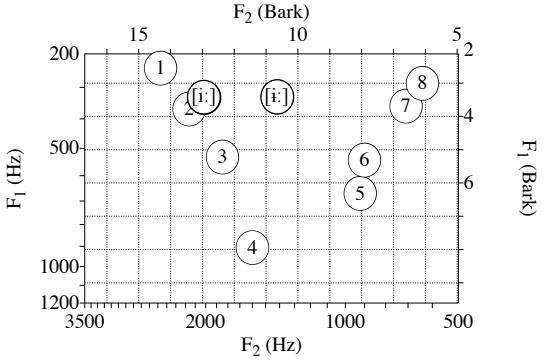


Figure 3: F_1/F_2 plot of mean formant values measured in the steady-states of regular [iː] and damped [ɨː] along with the eight primary cardinal vowels as pronounced by Daniel Jones. Each circle is one Bark in diameter.

Discussion

Articulatory as well as acoustic differences between regular and damped /iː/ were observed. We found a difference in tongue position that suggests that [iː] is pronounced further back, i.e. not as an alveolar, but more like a central palatal vowel. The shape of the tongue is also different in the two realisations, i.e. an downward slope from dorsum to tip for [iː], but an upward slope for [ɨː]. This supports the observations of [ɨː] by Noreen (1903), but not Borgström (1913). However, the tongue position and shape needs to be investigated further using more subjects of several Swedish dialects.

The acoustic analysis showed that the formant frequencies differ mainly in F_2, indicating that the main difference between the two realisations is that [ɨː] is pronounced further back than [iː]. This is in line with Engstrand et al. (2000). Both [iː] and [ɨː] seem closer to the second cardinal vowel [eː] than the first one [iː], suggesting that they are close-mid vowels. However, the South Swedish speaker used a slight diphthongisation [eiː], and although only the steady-state portion of the [iː] was used in the analysis, this may still have had some influence on the results. Our tentative results are based on a very limited mate-

rial and need to be followed up by larger studies with more subjects and vowel repetitions. Future work includes recording and analysing all Swedish vowels produced by at least 15 subjects each from three dialectal areas.

Exemplifying work in progress

We are currently developing tools for visualising the data. One way of plotting articulatory data is the common midsaggital view of the tongue from the side. Figure 4 shows an example of such a plot of the tongue profiles for one realisation of each of the Swedish long vowels /iː, yː, eː, øː, ɛː, ʉː, uː, oː, aː/ along with the palate contour for one speaker and one moment in time per vowel.

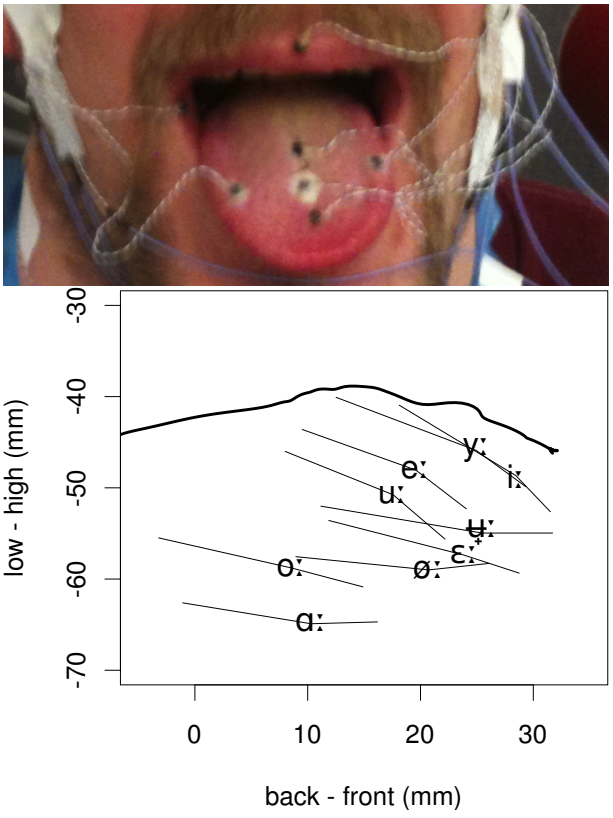


Figure 4: Sensors placed on a male south Swedish speaker and the corresponding midsagittal view of the palate contour and stylised tongue profile for one moment in time of the Swedish long vowels /iː, yː, eː, øː, ɛː, ʉː, uː, oː, aː/.

The palate profile was measured as the subject pressed the tongue tip sensor up against the palate and slowly pulled it backwards. The tongue profiles were reconstructed from sensors 1, 2 and 3 (see Figure 1) by simple linear interpolation. Future work includes developing visualisation tools for vowel dynamics in three dimensions.

Acknowledgements

This work is supported by a grant from the Swedish Research Council, grant no. 2010-1599. We also would like to thank Mark Tiede for his kind permission to let us use Mview.

References

Björsten S, Bruce G, Elert C C, Engstrand O, Eriksson A, Strangert E and Wretling P (1999). Svensk dialektologi och fonetik – tjänster och gentjänster. In *Svenska landsmål och svenskt folkliv*.

Boersma P and Weenink D (2011). Praat: doing phonetics by computer (version 5.2.17) [computer program]. Webpage http://www.praat.org/, visited 3-March-11.

Borgström M (1913). Askermålets ljudlära. In *Svenska landsmål och svenskt folkliv B*.

Bruce G (2010). *Vår fonetiska geografi*. Lund: Studentlitteratur.

Bruce G and Engstrand O (2006). The phonetic profile of swedish. *Sprachtypologie und Universalienforschung*, 12–35.

Eklund I and Traunmüller H (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, 1–21.

Engstrand O, Björsten S, Lindblom B, Bruce G and Eriksson A (2000). Hur udda är viby-i? experimentella och typologiska observationer. *Folkmålsstudier*, 83–95.

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

IPA (1999). *Handbook of the IPA*. Cambridge: Cambridge University Press.

Johansson F A (1973). *Immigrant Swedish Phonology*. Lund: Gleerups.

Kuronen M (2000). *Vokaluttalets akustik i sverigesvenska, finlandssvenska och finska*. Ph.D. thesis, Jyväskylä: University of Jyväskylä.

Ladefoged P (2003). *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Malden, MA: Blackwell.

Ladefoged P and Maddieson I (1996). *The Sounds of the World's Languages*. Oxford: Blackwells.

Linell P and Jennische M (1980). *Barns uttalsutveckling*. Lund: Liber.

Noreen A (1903). *Vårt språk I*. Lund:Gleerups.

Tiede M (2010). Mview: Multi-channel visualization application for displaying dynamic sensor movements. In development.

Audiovisual integration in binaural, monaural and dichotic listening

Niklas Öhrström, Heidi Arppe, Linnéa Eklund, Sofie Eriksson, Daniel Marcus, Tove Mathiassen and Lina Pettersson
Department of Linguistics, Stockholm University

Abstract
Audiovisual speech perception was investigated in three different conditions: (i) binaurally, where the same sound was presented in both ears, (ii) monaurally, where the sound was presented in one ear randomly, and (iii) dichotically, where the subjects were asked to focus on what was heard in the right ear. The results showed visual influence to be lowered in random monaural presentation as well as in dichotic presentation. Low visual influence to dichotic presentation, as compared with binaural presentation, supports the notion of an attentional component in audiovisual speech processing. Low visual influence in the random monaural presentation may be due to increased attention to the auditory modality because of uncertainty.

Introduction

This paper is concerned with cross modal integration in speech perception and whether or not the visual impact on auditory perception might be related to the amount of accessible attentional resources.

The McGurk effect (McGurk and MacDonald, 1976) shows that optical information about a speaker's speech gesture has an influence upon auditory speech perception, not only at low signal to noise ratios (Sumby and Pollack, 1954; Erber, 1969) but also when the acoustically conveyed speech signal is clear. In the study carried out by McGurk and MacDonald a face articulating /gaga/ presented together with an acoustically presented /baba/, was predominantly perceived as /dada/ by adult listeners. In a later study (Traunmüller and Öhrström, 2007a), incongruent audiovisual front vowels were presented. Perceived vowel openness correlated almost exclusively with vowel openness conveyed through the auditory channel, while perceived lip rounding correlated predominantly with lip rounding conveyed through the visual channel. An auditory (Swedish) /gig/ synchronized with a visual /gøg/ was accordingly perceived as /gyg/. An auditory /gyg/ synchronized with a visual /geg/ was perceived as /gig/.

Ever since the finding of McGurk and MacDonald (1976), the nature of audiovisual integration has been debated. Some theorists have claimed the effect to occur at an early level of speech processing, which means that optical information influences the phonetic percept (e.g. Traunmüller and Öhrström, 2007b, Colin et al., 2002), while others claim it to occur later: E.g. according to the fuzzy logical model of perception (Massaro, 1998), information in each modality is supposed to be processed and evaluated in parallel before integration and decision making (i.e. concept matching) take place.

Automaticity is another issue connected to timing of integration. Intuitively, an early integration approach would leave less space for endogenous attention to have impact on the incoming signal before it is integrated. Signs of automaticity have been shown in many studies (e.g. Green et al., 1991, Rosenblum and Saldaña, 1996, Hietanen et al. 2001). Massaro (1984) also claimed audiovisual integration to be robust to lack of attention. Considering audiovisual perception of emotions, Vroomen et al. (2001) claimed integration to be independent of attention. However, in later studies where distractors were used (Tiippana et al., 2004) and according to the dual task paradigm (Alsius et al., 2005; Alsius et al., 2007), presence of an attentional component has been demonstrated in audiovisual integration.

The present study aims to further investigate the robustness of audiovisual integration in speech perception using a (i) dichotic listening task, in which endogenous attention is kept on what is heard in the right ear, and (ii) a monaural task, in which sound is presented in one ear only, while the listener won't know in which one in advance. In the first task, attentional re-

sources are supposed to be consumed by focusing. The audiovisual integration would be inhibited if dependent of available attentional resources. The second task is concerned with uncertainty, where the listener won't know in which ear the sound will appear next (i.e. possibly attention consuming). Listening to one ear is however equivalent with a decrease in sound intensity of about 3 dB. This could in contrast potentially lead to more auditory confusions and more visual influence.

Method

Participants

In total 30 subjects, 25 female and 5 male, volunteered as perceivers. They were all native speakers of Swedish. They were all right-handed, reported normal hearing and normal or corrected vision. Their mean age was 24.5 years (SD = 5.6 years).

Speech materials

A right ear advantage (REA) test was made to ensure that the listeners' preference would be on the right ear. It was a subset of a test used by Söderlund et al. (2009), originally constructed by Hugdahl (2002). It consisted of the syllables /ba/, /ga/ and /da/ presented in congruent and incongruent dichotic fashion. There were a total of nine REA stimuli, each presented three times in random order.

The stimuli in the following experiments were a further edited subset of the visual, audio and audiovisual stimuli used in Traunmüller & Öhrström (2007a). There were two speakers, one male and one female.

In the first block the visual stimuli showed the speaker while pronouncing the syllables /gig/, /gyg/, /geg/ and /gøg/. Each token was presented twice in random order, thus giving a total of 16 presentations.

Block 2 consisted of auditory and incongruent audiovisual stimuli. A summary of these stimuli is shown in table 1. Each token in the second block was presented twice in random order, thus giving 48 presentations in total.

Block 3 consisted of stimuli corresponding to those in block 2, but presented in one ear at a time. Stimuli were randomized in such a way the listener couldn't predict in which ear the next sound would appear. Each token in block 3 was presented once, giving a total of 48 presentations.

Block 4 consisted of incongruent dichotic auditory and audiovisual stimuli. There were dichotic incongruences concerning vowel openness but not roundedness. The stimuli of block 4 are shown in table 2. Each dichotic token were presented twice, thus giving a total of 48 presentations.

Table 1. Stimuli presented in the second experimental block. A = acoustically presented stimulus, V = optically presented stimulus.

A	V	A	V
/gig/	-	/gyg/	-
/gig/	/gyg/	/gyg/	/gig/
/gig/	/gøg/	/gyg/	/geg/
/geg/	-	/gøg/	-
/geg/	/gyg/	/gøg/	/gig/
/geg/	/gøg/	/gøg/	/geg/

Table 2. Stimuli presented in the forth experimental block. A_{left} = acoustically presented stimulus in the left ear, A_{right} = acoustically presented stimulus in the right ear V = optically presented stimulus.

A_{left}	A_{right}	V	A_{left}	A_{right}	V
/gig/	/geg/	-	/gyg/	/gøg/	-
/gig/	/geg/	/gyg/	/gyg/	/gøg/	/gig/
/gig/	/geg/	/gøg/	/gyg/	/gøg/	/geg/
/geg/	/gig/	-	/gøg/	/gyg/	-
/geg/	/gig/	/gyg/	/gøg/	/gyg/	/gig/
/geg/	/gig/	/gøg/	/gøg/	/gyg/	/geg/

Experimental procedure

Three listeners were participating at a time. They were seated at approximately an arm's length distance from a computer screen and wore somewhat isolating headphones (Deltaco, stereo dynamic, HL-56). They were given instructions in both written and spoken form. The subjects wrote their answers on prepared response sheets in a forced choice design.

In the initial REA-test, the listeners listened to the incongruent dichotic stimuli. The listeners were asked to report what they had heard and choose between <ba>, <da> and <ga>.

The order of the following blocks varied across subjects to avoid context effects. In the experimental blocks, the nine Swedish long vowels appeared as response alternatives.

In block 1 (optic stimuli), the subjects were asked to report what vowel they had seen through speech reading.

In block 2 (binaural stimuli), the subjects were asked to report what vowel they had heard,

while watching the articulating face when shown.

In block 3 (monaural stimuli), the subjects were asked to report what they had heard while watching the articulating face when shown on screen. They weren't aware of in which ear the sound would appear next.

In block 4 (dichotic stimuli), the subjects were asked to report what they had heard in their right ear while watching the articulating face when shown on screen.

Results

According to the initial REA-test, a majority of the subjects responded mostly in accordance with what was presented in the right ear. This tendency was not however overwhelming: on average 53.6% (SD = 9.1%).

In the following, relative visual influence on perceived rounding will be calculated according to equation 1:

Equation 1:

$$\text{Rel.infl.} = (AV_{round} - A_{round}) / (V_{round} - A_{round})$$

AV_{round} = Proportion of audiovisual tokens perceived as a rounded vowel.
A_{round} = Proportion of auditory (only) tokens perceived as a rounded vowel.
V_{round} = Proportion of visual (only) tokens perceived as a rounded vowel.

Example: If an optic /i/, paired with an avoustic /y/ is perceived as rounded to a 60% extent, then $AV_{round} = 0.6$. If the acoustic /y/ in single mode is completely perceived as rounded, then $A_{round} = 1$. If the optic /i/ = is completely perceived as unrounded, then $V_{round} = 0$. The relative visual influence on the perceived rounding would then be 0.6.

Five subjects were excluded in the following analysis because of too small differences, ($|V_{round}-A_{round}| \leq .4$), leading to incomparable results and unreliable measures.

For block 1, the visual responses regarding roundedness are shown in table 3.

For block 2, the responses to auditory and audiovisual binaural stimuli regarding roundedness are shown in table 4. An intended /i/, produced by the female speaker was often categorized as /y/ (42.3%) and even as /ʉ/ in some cases (5.8%). This skewness is also present in block 3 and 4.

For block 3, the responses to monaurally presented stimuli are shown in table 5.

For block 4, the responses to dichotically presented stimuli are shown in table 6. As could be seen in table 2, intended/presented rounding didn't differ across ears.

Table 3. Confusion matrix for visually perceived roundedness (block1). "0"=unrounded, "1"=rounded. Rows: intended, columns: perceived rounding (%).

Stimulus	0	1
0	95.2	4.8
1	2.9	97.1

Table 4. Confusion matrix for perceived roundedness among auditory and audiovisual stimuli, binaurally presented (block2). "0"=unrounded, "1"=rounded responses to visual stimuli. Rows: presented, columns: perceived vowels (%).

Stimulus			
Aud	Vis	0	1
0	*	83.7	16.3
1	*	1.9	98.1
0	1	57.1	42.9
1	0	26.2	73.8

Table 5. Confusion matrix for perceived roundedness among auditory and audiovisual stimuli, monaurally presented (block3). "0"=unrounded, "1"=rounded responses to visual stimuli. Rows: presented, columns: perceived vowels (%).

Stimulus			
Aud	Vis	0	1
0	*	86.5	13.5
1	*	4.3	95.7
0	1	73.6	26.4
1	0	20.7	79.3

Table 6. Confusion matrix for perceived roundedness among auditory and audiovisual stimuli, in dichotic mode (block4). "0"=unrounded, "1"=rounded responses to visual stimuli. Rows: presented, columns: perceived vowels (%).

Stimulus			
Aud	Vis	0	1
0	*	86.5	13.5
1	*	8.7	91.3
0	1	72.8	27.2
1	0	21.4	78.6

The relative visual influence was calculated according to equation 1 for each subject in each condition. The averages across subjects are shown in figure 1. Paired samples t-tests revealed that the visual influence is significantly

lower in the monaural and dichotic condition as compared with the binaural condition: $t(24) = 4.89$, $p < .005$ (2-tailed); $t(24) = 2.71$, $p < .05$ (2-tailed).

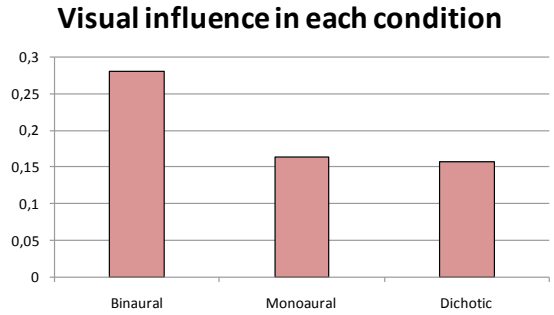


Figure 1. Visual influence on rounding in each condition. Averages across subjects.

Discussion and conclusion

The results of this study are in accordance with earlier studies (Alsius et al., 2005, Alsius et al., 2007, Tiippana et al., 2004), that there is an attentional component involved in audiovisual speech processing. Still, the issue about automaticity in audiovisual speech processing isn't yet clarified: We have shown that integration is inhibited when a competing task consumes attentional resources, but can we disregard from visual information, without looking away, even when attentional resources are available? Just asking the subjects to focus on what is heard vs. seen isn't a satisfactory approach since the two answers will reflect two different percepts, one vocal and one gestural (facial) (Traunmüller & Öhrström, 2007b).

The results in block 3 (monaural stimuli) is of particular interest. Only one ear involved, would intuitively evoke more confusions in auditory mode than binaural stimuli, due to degraded sound input. This would, according to Sumby & Pollack (1954) and Erber (1969), leave more space for visual influence. Instead there were slightly less confusions and visual influence significantly lower than for binaural stimuli. This may be due to the experimental design where, the listeners weren't aware of in which ear the next sound would appear. This uncertainty may cause attention to be drawn to the auditory modality. The visual influence in this study is substantially lower than that obtained in Traunmüller & Öhrström (2007a). This may be due to the experimental design, where visual stimuli were mixed together with auditory and audiovisual stimuli in the same experimental block, forcing the subjects always to attend to the speakers' face.

References

Alsius A, Navarra J, Campbell R, & Soto-Faraco S (2005). Audiovisual integration of speech falters under high attention demands. *Curr Biol*, 15: 839-843.

Alsius A, Navarra J & Soto-Faraco S (2007). Attention to touch weakens audiovisual speech integration. *Exp Brain Res*, 183.3: 399-404.

Colin C, Radeau M, Soquet A, Demolin D, Colin F & Deltenre P (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin Neurophysiol*, 113.4: 495-506.

Erber NP (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *J Speech Hear Res*, 12: 423-425.

Green KP, Kuhl PK, Meltzoff AN & Stevens EB (1991). Integrating speech information across talker, gender and sensory modality: Female faces and male voices in the McGurk effect. *Percept Psychophys*, 50: 524-536.

Hietanen JK, Manninen P, Sams M & Surakka V (2001). Does audiovisual speech perception use information about facial configuration?. *Eur J Cogn Psychol*, 13.3: 395-407.

Hugdahl K & Davidson RJ (2003). Dichotic listening in the study of auditory laterality. In: Kenneth Hugdahl, eds, *The Asymmetrical Brain*. Cambridge. MA US: MIT Press, 441-475.

Massaro DW (1984). Children's perception of visual and auditory speech. *Child Dev*, 55: 1777-1788.

Massaro DW & Stork DG (1998). Speech recognition and sensory integration. *Am Sci*, 86: 236-244.

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

Rosenblum LD & Saldaña HM (1996). An audiovisual test of kinematic primitives for visual speech perception. *J Exp Psychol Hum Percept Perform*, 22: 318-331.

Sumby WH, Pollack I (1954). Visual contribution to speech intelligibilty in noise. *J Acoust Soc Am*, 26.2: 212-215.

Söderlund G, Marklund E & Lacerda F (2009). Auditory white noise enhances cognitive performance under certain conditions: Examples from visuospatial working memory and dichotic listening tasks. In: *Fonetik 2009*, 160-164.

Tiippana K, Andersen TS & Sams M (2004). Visual attention modulates audiovisual speech perception. *Eur J Cogn Psychol*, 16.3: 457-472.

Traunmüller H & Öhrström N (2007a). Audiovisual perception of openness and lip rounding in front vowels. *J Phon*, 35: 244-258.

Traunmüller H & Öhrström N (2007b). The auditory and visual percept evoked by the same audiovisual stimuli. In: *AVSP 2007*, L4-1.

Vroomen J. Driver J & de Gelder B (2001). Is cross-modal integration of emotional expressions independent of attentional resources?. *Cogn Affective Behav Neurosci*, 1.4: 382-387.

A novel Skype interface using SynFace for virtual speech reading support

*Samer Al Moubayed and Jonas Beskow**

sameram@kth.se
beskow@kth.se
**Authors in alphabetical order*

Department of Speech, Music and Hearing
KTH Royal Institute of Technology, Stockholm, Sweden.

Abstract

We describe in this paper a support client interface to the IP telephony application Skype. The system uses a variant of SynFace, a real-time speech reading support system using facial animation. The new interface is designed for the use by elderly persons, and tailored for use in systems supporting touch screens.
The SynFace real-time facial animation system has previously shown ability to enhance speech comprehension for the hearing impaired persons. In this study we employ at-home field studies on five subjects in the EU project MonAMI. We present insights from interviews with the test subjects on the advantages of the system, and on the limitations of such a technology of real-time speech reading to reach the homes of elderly and the hard of hearing.

Introduction

Speech and sounds are important sources of information in our everyday lives for communication with our environment, be it interacting with fellow humans or directing our attention to technical devices with sound signals. For hearing impaired persons this acoustic information must be enhanced, supplemented, or even replaced by cues using other senses.

We believe that the most natural modality to use is the visual, since speech is fundamentally audiovisual and these two modalities are complementary.

The strong bonding and redundancy between the human face and the speech signal has been extensively investigated. The most basic redundancy, that is a result of a physical process is lip movements. Lip reading has been shown to enhance speech comprehension, for example, in different noisy environments (Summerfield (1992)).

On the other hand, there is a growing number of hearing impaired persons in the society today. With this growing number, there is also a growing number of research projects which aim to to develop the next generation of assisting devices that will allow this group - which predominantly includes the elderly - equal participation in communication and empower them to play a full role in society. Such projects are the H@H (Hearing at Home, Beskow et al. (2008b)), and MonAMI (Beskow et al. (2008a)).

For a hearing impaired person, it is often necessary to be able to lip-read as well as hear the person they are talking with that is in order to communicate successfully. Often, only the audio signal is available, e.g. during telephone conversations or certain TV broadcasts.

One of such supportive technologies for lip reading is the SynFace system (Beskow et al. (2004)). SynFace was developed as a real-time facial animation system that receive audio signal as input, and generates facial movements using a 3D animated face. These movement are then synchronized with the audio signal and showed to the listener in combination with the audio. SynFace, in several experiments, has been shown to enhance speech comprehension for normal and hard-of-hearing persons (Salvi et al. (2009)).

Visual speech reading support technologies such as SynFace, however, have not yet been employed in the everyday life of persons in need and there is little research on the integration of these technologies in smart homes and devices, so that they can efficiently and independently be used by people.

This study aims at integrating the SynFace system in the popular, widely used, Skype IP telephony system. We describe in this article the structure of the SynFace client interface of Skype, and discuss the different properties that advance on making SynFace an integrated solution for elderly. Further on, we describe field studies on five elderly subjects.

The paper is structured as follows. Section 2 describes briefly the structure of the SynFace system. Section 3 describes the SynFace interface for Skype and how it is tailored for the use by elderly. Section 4 describes field studies and Section 5 reports on advantages and limitations of the technology, and recommendations and insights from interviews with the test subjects.

The SynFace System

SynFace is a lip-synchronized talking head that is optimized as a visual reading support. The system was originally developed under the European Commission project Teleface (Beskow et al. (1997) as a virtual facial support system for the hearing impaired, and was later developed to support several languages such as English, Swedish, Dutch and German.

SynFace employs a 3D animated face model controlled by articulatory oriented parameters. An online control model drives the system based on time-stamped phonetic input. A phonetic recognizer is used in order to drive the face movement from the incoming speech signal. The constraints on the recognizer for this application are speaker independence, task independence and, above all, low latency. The recognizer is based on a hybrid of hidden Markov models and recurrent neural networks. Special effort has been put in reducing all sources of latency in the processing chain. This was achieved by limiting the look-ahead in neural networks, decoder and the control model for the hearing impaired.

Figure 1 shows a snapshot of one of the 3D animated agents used in SynFace.

Architecture

In this project, Skype, a popular IP telephony system, is used as a back end for SynFace.

Skype, indeed popular, still lacks an intuitive interface for the less experienced users in computer software, and requires reasonably adequate knowledge in computer based telephony to be able to use the system. In addition, the system provides a good environment to deploying and testing speech reading support using real-time computer animation for the use in IP telephone, since other services such as speech coding, compression, echo and noise reduction are advanced and robust. Figure 2 shows a snapshot of the Skype interface.

Figure 1: One of the animated faces in the SynFace system.

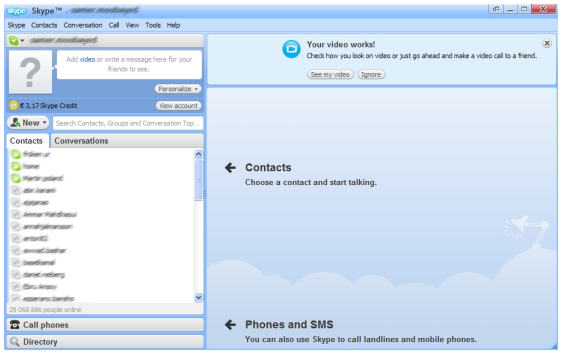


Figure 2: A snapshot of the the Skype IP telephony user interface.

The choice for building a new interface for Skype is made so to operate under touch-screen functionality, which is becoming more and more accessible using smart devices and laptops. The interface is also made very minimalist, so to provide basic functionality that allows subjects to perform the task with a minimal number of operations, while being intuitive and responsive.

Figure 3 shows a screen-shot of the client interface. The interface loads all the contacts from the subject Skype account when the subject starts the system. Every contact is represented by a button, that includes a contact picture in case the Skype contact is associated with a profile picture. On clicking on the button, a call is started and a hang-up button is shown so the call can be terminated. Whenever a call is started, the SynFace animated face starts moving in sync with the audio signal.

The audio signal is delayed by a 200 ms period to allow for animation, and re-synchronized with the animation. Figure 4 shows the flow of in-

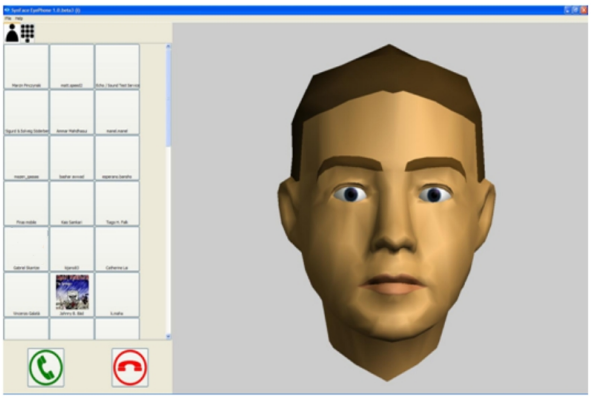


Figure 3: A snapshot of the new interface, including the SynFace facial animation system.

formation and the underlying structure of the system.

In this implementation, the incoming audio from Skype is transferred to a visual audio channel, instead of to the speakers, using the Virtual Audio Cable system (http://www.ntonyx.com/vac.htm). The audio is then transferred to the SynFace system. In the SynFace system, the audio is fed into the phoneme recognizer, and the output is used to generate the facial animation. After that, the audio and the animation is synchronized and played in a delay of 200 ms to the speakers.

Field Studies

The new interface is tested in the context of the MonAMI project. The system is targeted for test at two sites of assisting home centers for the elderly in Stockholm. A pitch presentation is given to all the interested residents and asked for participating in long term test period of one week, where the system will be installed at their own apartments, and used as an alternative to their home phones. At both sites, five subjects in total have subscribed to the tests.

The pitch presentation of the system is made using a 22" screen all-in-one Asus eee-top computer. All the subjects showed dissatisfaction about the size of the setup, and required a smaller hardware due to space limitations at their own apartment. Large display screen, although provide higher quality image, but limit the mobility of the system due to its size and weight. For the subscribed users, a 13" light weight touch screen laptop is used for the studies. Figure 5 shows a picture of the computer with the system deployed.

Before the tests, an interview is carried out with each of the subjects about their use of telephones, their use of previous computer technology, and their expectations of the system. A telephone based hearing test is also carried out to measure the hearing ability of the subject, and found that all the subjects have normal or near normal hearing ability.

The same interview was carried out after one week of the use of the system. During the test, a telephone and on-site technical support was offered for any difficulties using the system with the help of two students.

Results

From the pre-test interviews, it was clear that all the test subjects had very little knowledge about the contribution of facial and lip movement to speech understanding, and hence were reluctant to use the system to its potential. However, several subjects reported during the final interview that they realized the contribution of lip movements from conscious observation to SynFace and other mediums such as TV.

One of the main hurdles for the tests were technical barriers between the subjects and the usage of computers. Although the system was installed on a dedicated machine, and required no knowledge on how to operate a computer, the need to deal with a hardware had very often required assistance. Nonetheless, a positive reaction to using the system was seen with the subjects with a little better computer skills.

In regards to the system interface, subject very seldom required any help to operate the system, and two subjects have used the system for more than 40 calls during the test week. Regarding SynFace lip movements, users have shown high satisfaction by the animation of the face and by the synchronization of the face and the audio signal, however, they believed that they are in no need for the service. We believe this is due to several effects such that the subjects did not have hearing impairment, and did not believe that information from the face can help to better understand the speech signal.

One interesting insight was that, before conducting the test, the subjects have shown discomfort towards the look of the SynFace facial design, and demanded a more human-like appearance of the face. However, from the post-test interviews, all the test subjects shown comfort in looking the face and seen themselves in no need for a more human-like facial appearance.

Discussion and Conclusions

This paper presents the structure and the interface of a virtual speech reading support system that is integrated with the IP telephony application Skype. The interface is developed to provide the users of Skype with automatically generated

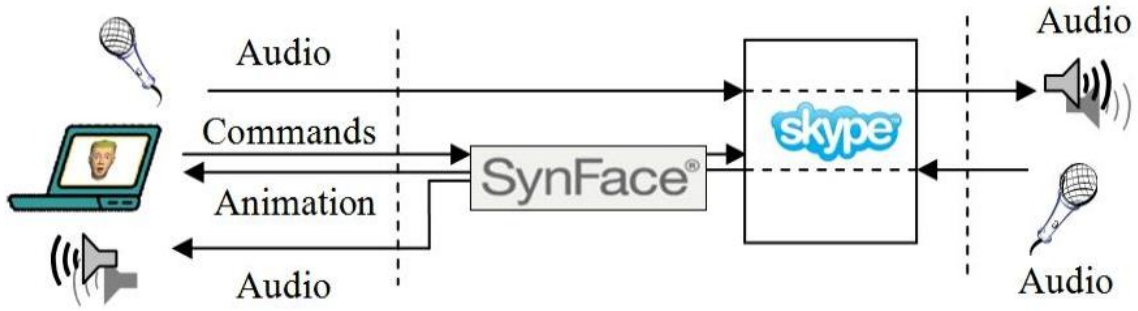


Figure 4: A data flow chart of the system.

Figure 5: The computer setup used for the field studies.

facial movements supporting the audio signal of the conversant.

The visual support is provided using the SynFace real-time speech driven facial animation system. This interface is tailored for the use by the elderly at a home setting and for a long term purpose, and hence structured to be accessible and intuitive, and avoids as much as possible experience in how to operate a hardware device or a software, which we believe to have been hurdles in the way for the wide use of this technology by the intended audience, namely the elderly and the hard of hearing.

The field study, which have been carried by 5 elderly subjects during one week, provide insights on the technology and the limitations of the system to be used independently by the targeted audience.

The findings from the study show that, although the current state of the virtual speech reading support technology is mature enough to be accepted by the elderly, many difficulties still persist in the face of this technology to take of, and should be strongly taken into consideration. Such considerations are the size, weight, mobility and operability of the hardware, and intuitive, responsiveness and clarity of the user interface.

We also believe that, since the subjects who have taken part in the study did not suffer from any hearing impairment, this lessened the need of the subjects to exploit the system to its potential, since they did not rely on the facial animation to enhance their speech comprehension.

Acknowledgement

This work is partially funded by MonAMI, an Integrated Project under the European Commission's Sixth Framework Program (IP-035147), and by Stiftelsen Promobilia (grant #8038).

References

Beskow J, Dahlquist M, Granström B, Lundeberg M, Spens K and Öhman T (1997). The teleface project-multimodal speech communication for the hearing impaired. In *Proceedings of Eurospeech'97*. Citeseer.

Beskow J, Edlund J, Granström B, Gustafson J, Jonsson O and Skantze G (2008a). Speech technology in the european project monami. *FONETIK 2008*, 33.

Beskow J, Granström B, Nordqvist P, Al Moubayed S, Salvi G, Herzke T and Schulz A (2008b). Hearing at home–communication support in home environments for hearing impaired persons. *Proceedings of Interspeech Brisbane, Australia*.

Beskow J, Karlsson I, Kewley J and Salvi G (2004). Synface–a talking head telephone for the hearing-impaired. *Computers helping people with special needs*, 627–627.

Salvi G, Beskow J, Al Moubayed S and Granström B (2009). Synface: speech-driven facial animation for virtual speech-reading support. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009:3.

Summerfield Q (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions: Biological Sciences*, 335(1273):71–78.

Anticipatory lip rounding– a pilot study using The Wave Speech Research System

Gabrielsson, D., Kirchner, S., Nilsson, K., Norberg, A., Widlund, C.
Speech-language Pathology Students at Karolinska Insitutet
Supervisor: Cortes, E.
Department of Linguistics, University of Stockholm

Abstract
Previous research on anticipatory lip rounding has generated two views on this phenomenom. On the one hand there is the look ahead model which postulates that the initiation of lip rounding is governed by the structure of the utterance and on the other there is the temporally locked model that suggests that the initiation occurs at a set time before a rounded vowel. This current study aimed to replicate parts of an earlier one by Lubker and Gay (1982), who examined temporal aspects of anticipatory lip rounding. The purpose of the replication was to compare data obtained from electromyography (EMG) and magnetometry with data from the Wave System (NDI). A number of subjects recorded Swedish speech material. Movement coordinates from the upper and lower lip were obtained in the x, y and z dimensions. However, only the x dimension was further analyzed since it represented an anterior-posterior lip movement. While processing data it became apparent that this dimension alone was insufficient for determining when the lip rounding was initiated. The analysis did, however, provide an answer to how the subjects organized their lip rounding in terms of look ahead versus temporally locked models.

Introduction

Research in the field of lip rounding has generated two views on this phenomenom. One view, suggested by Henke (1966) amongst others, postulates that speakers apply a so called look ahead model (Abelin et al., 1980). This model assumes that anticipatory lip rounding is governed by the length of the consonantal string preceding a rounded vowel, given that all consonants are unmarked for labiality (McAllister, 1978). The other view, called the temporally locked model, suggests that the initiation of lip rounding occurs at a set time before the rounded vowel and is therefore not influenced by the length of the consonantal string. This view is supported by Harris and Bell-Berti (1979; 1981; 1982) and Gay (1977).

Almost all previous research has, despite differing views, used EMG as primary means of measurement. Since EMG registers electrical activity in muscles which does not necessarily indicate an actual movement, this means of measurement has been questioned (Lubker & Gay, 1982). Due to this insufficiency it was of interest to use technique that aimed to measure actual movement rather than electrical muscle activity. Hence, Lubker and Gay (1982) complemented their EMG measurements with magnetometry, where a magnet was attached to the upper lip to register its articulatory movements in the x and y dimension. However, data obtained from EMG only marginally corresponded with the movements registered by magnetometry. Therefore it was still of interest to perform a similar experiment with improved technique, which this replication aimed to do.

The experiment was performed with a prototype of The Wave Speech Research System (NDI) that registers the small movements associated with speech by using an electromagnetic field. There were several reasons for using Lubker and Gay's (1982) study for this replication. For one, the authors represented both views on the nature of anticipatory lip rounding, giving them a critical attitude towards previous research, and their speech material was comprehensive, containing both nonsense words and real Swedish utterances.

Method

Subjects and speech material

Five female subjects, aged 23 to 45, participated in the experiment. All of them where native speakers of Swedish except for one who had Italian as her native language but was an advanced speaker of Swedish. All of the subjects were familiar with the nature of the experiment.

Since the experiment aimed to replicate parts of an earlier study by Lubker and Gay (1982) parts of their speech material was used. The material consisted of nine Swedish utterances and twelve nonsense words that followed Swedish phonotactic rules, see Table 1. The participants read the material once before recording and then during recording the subjects read it twice without pause, with the instructions to articulate as naturally as possible. Four reference sounds /ɐ/, /i/, /u/ and /m/ were produced at the end of every repetition with the instructions to articulate as distinctly as possible, with the purpose of obtaining reference data to match with the other utterances.

Table 1. The speech material in orthographic and phonetic representation. Parentheses: Speech sounds not pronounced by all subjects.

Orthographic	IPA
ditt skrot	['dɪt:skəˈru:t]
ett danskt skåp	[ɛˈtɛn:s(k)t'sko:p]*
vilket tyder	[vɪlkəˈty:dɛr]
ditt sot	['dɪt:'su:t]
ett danskt skrutt	[ɛˈtɛn:s(kt)'skrɵt:]*
ditt skot	['dɪt:'sku:t]
det också	[de:(t)'ɔksɔ]*
Stockholm	['stɔkˌ(h)ɔlm]*
ett norskt skåp	[ɛt:'nɔʂ(kts)'ko:p]*
asto	[ɐˈstu:]
asksto	[ɐsk'stu:]
ato	[ɐˈtu:]
aksto	[ɐk'stu:]
oto	[ʊˈtu:]
oksto	[ʊk'stu:]
osto	[ʊˈstu:]
osksto	[ʊsk'stu:]
ata	[ɐˈtɒ:]
aksta	[ɐk'stɒ:]
asksta	[ɐsk'stɒ:]
akstra	[ɐk'strɒ:]

Procedure

The recording was made using the Wave System where two sensors were placed mid-sagittally on the vermillion border on the upper and the lower lip. To register the movements of the head a reference sensor was placed on a headgear which was worn by the subject. The Wave System subtracted the information obtained by the headgear sensor from the information registered by the lip sensors leaving only movements caused by articulation. The subjects were placed in profile in front of a generator that created an electromagnetic field. Simultaneously with the registration of the lip movements an acoustic recording was made using a microphone placed at breast height on the subject. The Wave System synchronized the acoustical data with the movement data by adding a synchronizing signal to the pure acoustical signal.

Data processing

Data from three out of five subjects were selected for further processing. The first elimination was due to the fact that the lower lip sensor could not be attached to one of the subjects. The second elimination was due to one of the subjects not being a native speaker of Swedish.

From the remaining data, the second reading was chosen for further analysis. The acoustical recording was transcribed using Wavesurfer which provided an onset time for all phonemes, which subsequently could be matched with the movement data. The Wave System generates movement data in three dimensions, x, y and z and the torsion for each one of them. The coordinates for these dimensions were converted into diagrams to visualize the lip movements during the utterances. The diagrams of the x dimension were chosen for thorough analysis since this dimension represented the spatial direction horizontally parallel to the generator, given that the subject was seated with her left ear towards the generator. In other words, the x dimension represented the anterior-posterior protrusion of the lips. The aim was to locate the initiation of the protrusion movement for comparison with the acoustical onset of the rounded vowel.

Results

The information from the x dimension alone was insufficient for deciding when anticipatory lip rounding was initiated but the data does show the movement of the lips from a posterior to an anterior position. A higher value on the y axis in the diagrams indicates a more posterior

position of the lips. Figure 1 shows the reference sounds /ɐ/, /i/, /u/ and /m/. Figure 2 shows the movement from a spread to a rounded vowel in the utterance “ditt skrot”. All figures show utterances produced by three different subjects.

Figure 3 shows nonsense words with varying length of the consonantal string and this figure demonstrates that movement of the lower lip in the x dimension is initiated at different points in time before the rounded vowel.

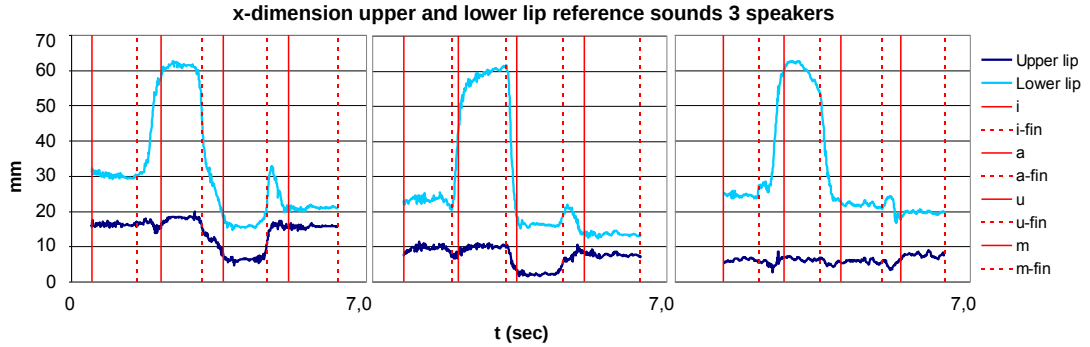


Figure 1. Left to right: Subject 1-3 producing the reference sounds /ɐ/, /i/, /u/ and /m/. Red vertical continous lines indicate the acoustical onset of a speech sound. Red vertical dotted lines indicate the end of a speech sound. X-axis: time in seconds. Y-axis: posterior-anterior movement in mm. High values on the y axis indicate a more posterior position.

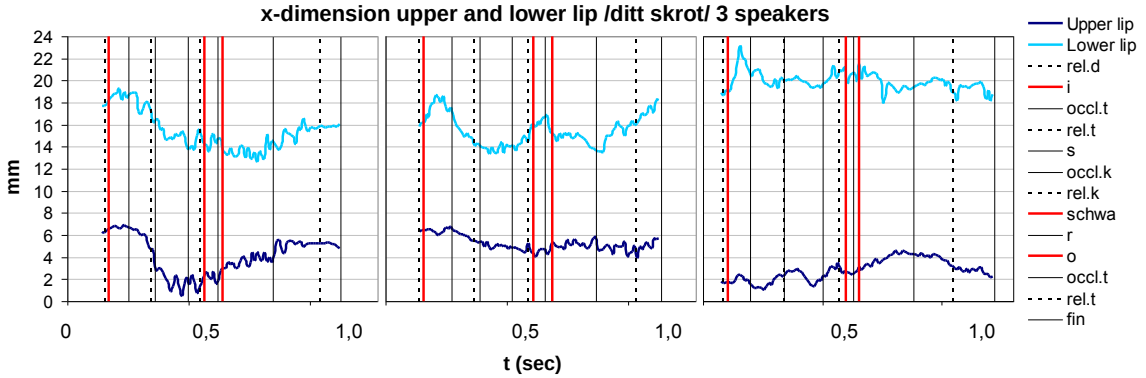


Figure 2. Left to right: Subject 1-3 producing ”ditt skrot”. Red vertical lines indicate the acoustical onset of vowels. Black vertical dotted lines indicate the release in plosives. Black continous lines indicate the occlusion in plosives and the initiation of remaining speech sounds. X-axis: time in seconds. Y-axis: posterior-anterior movement in mm. High values on the y axis indicate a more posterior position.

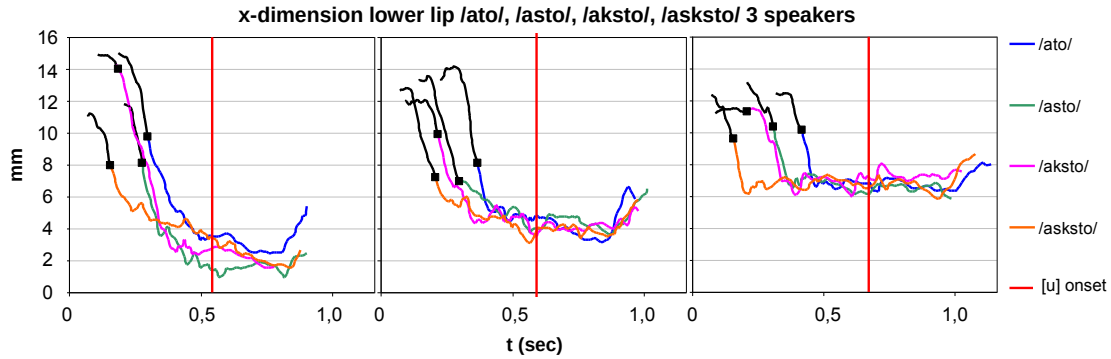


Figure 3. Left to right: Subject 1-3. The movement of the lower lip in the x dimension for the nonsense utterances /ato/ (blue line), /asto/ (green line), /aksto/ (pink line), /asksto/ (orange line) synchronized in time with the acoustic onset of [u] (red vertical line). The black part of the lines: [ɐ]. Black squares: onset of the first consonant. X-axis: time in seconds. Y-axis: posterior-anterior movement in mm. High values on the y axis indicate a more posterior position.

Discussion

The aim of this study was to perform a partial replication of an earlier study by Lubker and Gay (1982) and compare results based on EMG measurements and magnetometry with results obtained by the Wave System. The decision to only analyze the x dimension generated problems with deciding when anticipatory lip rounding began because when rounding it is not only the degree of protrusion that impacts the movement but other aspects including jaw opening and the torsion of the lower lip seem to have an influence as well. Attempts were made to use data from the reference sound /m/ to create a base line as it ought to be articulatory neutral with regards to protrusion. This attempt further illustrated the difficulties in studying one dimension alone, as the expected neutral reference sound /m/ produced a larger movement of the lower lip towards an anterior position than did the reference sound /u/ for subject 2 and 3, see Figure 1. It was discussed whether the initiation of the anticipatory lip rounding could be set at the turning point when a posterior movement becomes an anterior one. This, however, might also include movements other than actual protrusion. For example in Figure 2 the turning point occurs during /i/ but the movement towards /u/ also includes /ə/. Hence, this movement indicates the transition from /i/ to /ə/ rather than the initiation of the protrusion. With this in mind it would have been arbitrary to choose the starting point on a visual basis alone without a theoretical foundation. One other possible alternative that might have created a base line in the x dimension could have been to place another sensor mid-sagittally on the front side of the upper and lower row of teeth, something to keep in mind for future research in the field.

The results did show that the subjects did not initiate the lip rounding at the same time for all four nonsense words, see Figure 3. In other words, lip rounding is not temporally locked for these three speakers in these utterances but is governed by the length of the consonantal string which would provide some support to the look ahead model.

Worth noting is the possible influence of the technical equipment on the subjects' articulation, since the sensors can make them aware of their lip movements. This could result in the subjects not using their normal movement pattern.

Despite the problems with this experiment the Wave System has potential, as it in comparison to the EMG method measures actual movement in three dimensions whilst EMG technichally only measures one dimension; whether muscle activity is present or not. However, three dimensions generate a large amount of data that currently demands manual processing. Since a greater number of subjects is desirable in these types of studies it would be of great value if a dedicated software that translates the movement coordinates to comprehendable graphical representations were to be developed. This would enable users to take full advantage of the system with regard to its three dimensions, which potentially could shed new light over this field of research.

References

Abelin, Å., Landberg, I. & Persson, L. (1980) A Study of Anticipatory Labial Coarticulation In the Speech of Children. *Papers from the Insitute of Linguistics*, University of Stockholm, II, 2–18

Bell-Berti, F. & Harris, K. S. (1979) Anticipatory coarticulation: some implications from a study of lip rounding. *Journal of the Acoustical Society of America* 65, 1268-70

Bell-Berti, F. & Harris, K. S. (1981) A temporal model of speech production, *Phonetica* 38, 9-20

Bell-Berti, F. & Harris, K. S. (1982) Temporal patterns of coarticulation: lip rounding, *Journal of the Acoustical Society of America* 71, 449-454

Gay, T. (1977) Cineflourographic and Electromyographic Studies of Articulatory Organization, In: M. Sawashima and F.S. Cooper, eds, *Dynamic Aspects of Speech Production*, Japan: University of Tokyo, 87-105

Henke, L. (1966) Dynamic articulatory model of speech production using computer simulation. Diss. Massachusetts Institute of Technology

Lubker, J. & Gay, T. (1982) Anticipatory labial coarticulation: Experimental, biological and linguistic variables *Journal of the Acoustical Society of America* 71(2)

McAllister, R. (1978) Temporal asymmetry in labial coarticulation. *Papers from the Insitute of Linguistics*, University of Stockholm 35, 1-29

NDI: Northern Digital Inc. Europe, GmbH, Life Sciences Division: Radolfzell, Germany.

Coarticulation: A universal phonetic phenomenon with roots in deep time

Björn Lindblom[1] & Peter MacNeilage[2]
[1]Department of Linguistics, Stockholm University, Sweden, [2]Department of Psychology, University of Texas at Austin, USA

Abstract

Coarticulation is a universal feature of spoken languages. Many decades of experimental phonetic research have produced a large literature on the topic. However, despite many important contributions, we still lack an answer to the perhaps most fundamental question about coarticulation: Where does it come from?

In keeping with the Frame/Content theory of speech evolution (MacNeilage 2008) our analysis subsumes speech movements in the class of continuous cyclic motor behaviors (swimming, walking, breathing and chewing) and views them as sharing the discrete positional control seen across species in 'precision walking' (Grillner 2006) and reaching (Georgopoulos & Grillner 1989).

In speech the well-known conservatism of evolution is evident both in its syllabic organization - which is built on the oscillatory motion of the mandible - and its use of discrete spatial targets –in the production of phonemes.

Once this organizational framework is assumed, a natural explanation of coarticulatory overlap presents itself. The overlap arises from the fact that the responses of articulatory structures to discrete segmental goals are slower than the rate at which the open and close states of the syllabic jaw cycle occur.

Explaining coarticulatory overlap

When we speak, our motor system coordinates a large number of neuro-muscular components. The movement between two consecutive phonemes is rarely a single one-parameter trajectory. The kinematics of a CV syllable is better pictured as a time chart specifying a long list of actions to be performed by articulatory, phonatory and respiratory structures.

Even the simplest utterance is a multi-channel event. An example makes that evident. Consider the syllable [ku] spoken in isolation. There is no lip activity specified for the stop. Nevertheless, we find that the lip rounding for [u] is in progress during the closure, the articulatory movements for [k] overlapping or being coarticulated with those for [u].

The elementary fact highlighted here is that coarticulation is manifested in a temporal overlap between any two channels recruited by different phonemes. The often cited representation of this fact is the schematic used by Joos in his classic monograph on acoustic phonetics (1948). It shows the beginning of the phrase *Wo ist ein Hotel*? represented as a series of 'innervation waves' overlapping in time.

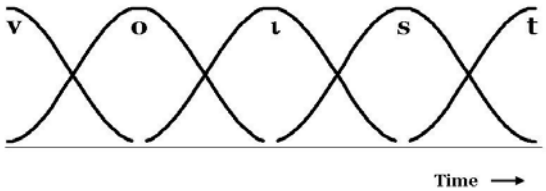


Figure 1. Coarticulatory overlap illustrated by Joos's classic "Overlapping innervation waves" for the German phrase 'Wo ist ein Hotel?'

Why do the movements for the adjacent phonemes have to overlap? Could things be otherwise?

For one thing, reaching articulatory goals for a phonetic segment takes time. For all articulators to be in position for [u] shortly after the release of [k], the rounding and other movements must be initiated well in advance. As suggested in Joos's diagram every segment gives rise a wax-and-wane pattern, a period of anticipation followed by a de-activation phase.

Second, different channels show different temporal properties. Along a continuum from

slow to fast they line up with breathing at the slow end, the jaw and the tongue body at intermediate positions and the tongue tip and phonatory mechanisms at the fast end. Although those characteristics do not by themselves automatically entail coarticulatory overlap, they are certainly part of the phenomenon of coarticulation.

It would seem that to explain coarticulation it is necessary to understand the origin of the temporal overlap of consecutive gestures. That is the problem we will try to address in this contribution.

Liberman's view of coarticulaiton

Alvin Liberman gave the question of the origin of coarticulation a great deal of thought. His thinking had been influenced by the difficulties he and his team at Haskins Laboratories had experienced in constructing reading machines for the blind (Liberman et al 1967). The perception of speech is special, he had concluded. Phonemes can be transmitted at rates of 15 per second or higher, but humans can only identify auditory items at rates around seven per second or lower.

He ended up viewing coarticulation as an evolutionary development that had emerged in response to a tacit demand for a more rapid rate of communication. Without coarticulation, he argued, we would be speaking no faster than we can spell. In his view, coarticulation is what makes the high rates of phonemes per second possible. In parallel, auditory perception – allegedly not capable of handling such high rates – co-evolved with production and came to include a phonetic module specialized for speech perception and for decoding coarticulated signals.

A similar view has been expressed by Phil Lieberman (1991).

Our own approach to this issue is different. Rather than viewing coarticulation as an innovation for increasing transmission rate, we prefer tracing the roots of this process to how speech motor control has been shaped, step by step, by building on existing mechanisms.

From babbling to speech

First, in broad strokes, a few facts about infant speech production. At 6-8 months normal children engage in canonical babbling: The child phonates and moves its jaw rhythmically up and down while leaving the rest of the vocal tract inactive (MacNeilage 2008). The acoustic result roughly resembles [ba ba ba] although the 'segments' and 'syllables' implied by this transcription should not be taken to mean that they are part of the child's motor program. The received view is that they are fortuitous consequences of the jaw movement, the up position causing the lips to close and the down position creating an open vowel-like vocal tract. This phonetic development is a very robust milestone (Vihman 1996).

How does the child go from babbling to speech? How does the transition from pseudo segments to real segments occur?

Clues from non-speech movements

It has been suggested (MacNeilage 2008) that the oscillatory nature of canonical babbling is a strong indication that it is based on a type of neural machinery that evolved a very long time ago to serve locomotor and vegetative purposes. This work connects with a vast field of research on cyclic motor behaviors such as breathing, chewing, and various modes of locomotion such as swimming, walking and running (Grillner 2006).

Numerous studies show that these rhythmic behaviors involve central pattern generators (CPG's) which produce sequences of cyclically alternating motions:

"… *in all animals, vertebrates or invertebrates, movements are controlled by CPG networks that determine appropriate sequences of muscle activation ….. Each animal is endowed with a broad repertoire of CPGs, located in different regions of the central nervous system …and available for differential activation, thus providing animals with a distinctive set of solutions to accommodate their widely divergent patterns of behavior.*" (Grillner 2006).

CPG networks generate the undulatory pattern of swimming seen in lampreys and salamanders. They respond to sensory information about the environment. Their behavior is highly flexible and not limited to 'rigid stereotypies'.

Chewing is also considered to be based on CPG's::

"*The rhythmical activation of the various muscle groups involved in mastication is generated by a CPG in the brain stem. Sensory feedback, particularly from intraoral mechanoreceptors, modifies the basic pattern and is particularly important for the proper coordination of tongue, lips, and jaws'.* (Lund 1991).

Chewing is generally described as involving the coordination of masticatory, facial, lingual, neck and supra- and infra-hyoid muscles – a description that could equally well be applied to speech. Since evolution tends to base its innovations on existing capacities, the similarity between speech and chewing strongly suggests that in ancient times rhythmic mechanisms like those in chewing came to be used in speech and are still present in the development and adult use of spoken language (MacNeilage 2008, Hiiemae & Palmer 2003).

Discrete target control: 'precision walking' and reaching

Of crucial importance to the present argument is the fact that the CPG control associated with basic behaviors is maintained throughout phylogenetic history and is seen to combine with *discrete target control* in more evolved forms of locomotion.

"*… locomotor movements that require visuomotor coordination with a precise foot placing— like walking up a ladder, when each foot must be accurately placed on each rung—is difficult not only in the decorticate state but also after transection of the corticospinal tract*" (Georgopoulos and Grillner, 1989).

"*…the problem of visually-guided feet placement is an interesting topic since it involves the superposition of discrete and rhythmic movements. For instance, when specific feet placements are required during walking (e.g. when a cat walks over a branch or when we cross a river walking on stones), rhythmic signals from the CPGs need to be modulated such that the feet reach specific end positions*." (Ijspeert 2008).

"*During this type of ''precision walking", neurons in motor cortex become strongly activated in precise phases of the movements ….. The same neurons are also activated during reaching tasks. These corticospinal neurons are thus involved in the precise placement of the limb, whether in locomotion or other motor tasks. The accurate placement of the foot during locomotion in complex terrain can best be considered as a dynamic reaching movement superimposed on the locomotor movement itself*". (Grillner 2006).

"*Locomotion and reaching have traditionally been regarded as separate motor activities. In fact, they may be closely connected both from an evolutionary and a neurophysiological viewpoint. Reaching seems to have evolved from the neural systems responsible for the active and precise positioning of the limb during locomotion; moreover, it seems to be organized in the spinal cord. The motor cortex and its corticospinal outflow are preferentially engaged when precise positioning of the limb is needed during locomotion and are also involved during reaching and active positioning of the hand near objects of interest. All of these motor activities require visuomotor coordination, and it is this coordination that could be achieved by the motor cortex and interconnected parietal and cerebellar areas.*" (Georgopoulos & Grillner 1989)

A framework for speech

The above observations allow us to propose a framework within which we could seek an answer to how the step from babbling to speech, specifically from pseudo-units to the genuine segments and syllables of adult speech may be taken.

In the top row of Figure 2 examples of rhythmic movements such as breathing, chewing and locomotion are listed. Swimming and walking in the lamprey are undulatory and cyclic respectively. They belong in the top cell of the first column. which exemplifies control by CPG networks. In the right hand cell of the top row we find the type of locomotion that involves exact foot placement: 'precision walking'. The binary division into two categories is admittedly a simplification, since species differ by degree in their ability to perform precise limb positioning.

The bottom cells extrapolate from the knowledge summarized in the top row to the development of speech. Canonical babbling is assumed to be driven by the CPG networks also serving chewing. There is no active control behind the quasi-segmental/syllabic output. As the child explores its vocal tract auditorily and motorically, it gradually acquires a better mastery of its vocal system and becomes able to coordinate the basic jaw movement with articulatory activity. In other words, rather than just letting the jaw close the vocal tract at the lips as in canonical babbling, it may reinforce that occlusion by active muscular action or, with time, also succeed in making it elsewhere, perhaps somewhere along the hard palate, by using the tongue instead. Similarly, the tongue may actively leave its rest position during the open vocal tract which would vary the quality of

the 'vowel' part. Since this mode ('precision talking') relies on a form discrete target control it is a close parallel to 'precision walking'

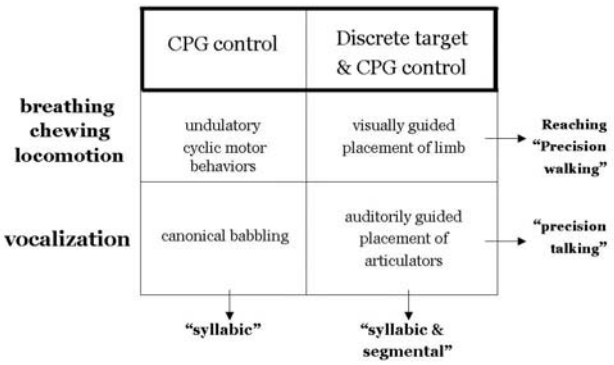


Figure 2. Top row summarizes our review of non-speech motor mechanisms. Bottom row suggests how those mechanisms come into play in speech.

Origin of coarticulatory overlap

In accordance with Frame/Content theory we have claimed that the syllabic organization of speech presents a further example of a cyclic motor behavior (the oscillatory motion of the mandible).

We have also maintained that, as seen in precision walking and reaching, this motion is modulated by discrete target control as exemplified in the production of phonemes.

When those assumptions are made, it becomes possible to address the question asked in the introduction: *Where does the phenomenon of coarticulation come from?*. Our answer is that coarticulatory overlap arises from the fact that the responses of articulatory structures to discrete segmental goals are slower than the rate at which the open and close states of the syllabic jaw cycle occur.

Acknowledgement

The authors are grateful to Giampiero Salvi of TMH/KTH for bringing Ijspeert's work to our attention.

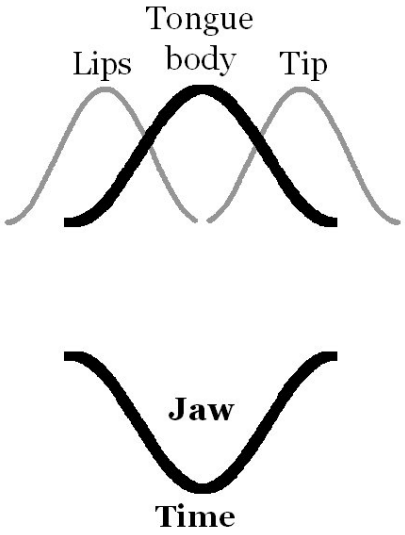


Figure 3. Schematized kinematic traces drawn according to the format of the Joos diagram. The responses of articulatory structures to discrete segmental goals (top traces) are slower than the rate at which the open and close states alternate in the syllabic jaw cycle (bottom trace).

References

Georgopoulos A P & Grillner S (1989): Visuomotor coordination in reaching and locomotion", *Science 245*:1209–1210.

Grillner S (2006): Biological pattern generation: The cellular and computational logic of networks in motion, *Neuron 52*:751–766..

Hiiemae & Palmer (2003): Tongue movements in feeding and speech, *Critical Reviews in Oral Biology and Medicine 14*:413-429.

Ijspeert A J (2008): Central pattern generators for locomotion control in animals and robots: a review, *Neural Networks 21/4*:642-653.

Joos M (1948): "Acoustic phonetics", *Language 24*:2, supplement.

Liberman A M, Cooper F S, Shankweiler D P & StuddertKennedy M (1967): Perception of the speech code, *Psychological Review 74*:431-461.

Lieberman P (1991): *Uniquely human*, Harvard University Press:Cambridge, USA.

Lund 1991): *Critical Reviews in Oral Biology and Medicine 142*(1):33-64

MacNeilage P (2008). *The origin of speech*, Oxford University Press.

Vihman M M (1996): *Phonological development*, Blackwell.

Phonetic transcriptions as a public service

Michaël Stenberg
Centre for Languages and Literature, Lund University

Abstract

This paper, which highlights one aspect of a thesis work in progress, focuses phonetic transcriptions aimed at a general public. It compares various systems used, and discusses what social circumstances may promote a demand for this kind of transcription. Public service transcriptions appear in encyclopedias, foreign and second language textbooks, phrasebooks for travellers, dictionaries or databases for broadcasters, and even in advertising. Sweden, a country with a relatively high and even level of education, experienced an increasing need for such transcriptions in the second half of the 20th century. An interesting—and particularly Swedish—application are the so-called respelled pronunciations introduced in 1959 by Systembolaget (the Swedish Alcohol Retail Monopoly) in its product catalogues, with the view of making customers feel less awkward asking for wines at the counter. In more recent years, the advent of the Internet has brought an abundance of IPA as well as respelled transcriptions, some of them elaborate and scientifically done, others, however, of doubtful quality.

Keywords: phonetic transcription, general public, IPA, respelled transcription, social circumstances

Introduction

Phonetic transcriptions appear in many contexts and may thus be directed to a variety of target groups. In scientific works, such as recordings of dialects or advanced pronunciation dictionaries, they are aimed at a narrow circle of scholars or specialists, likewise when used within the logopedics an phoniatrics professions for describing details of atypical speech. Often, however, transcriptions serve a more general public: they feature in encyclopedias, foreign and second language textbooks, phrasebooks for travellers, pronunciation dictionaries/databases for broadcasters, and even in advertising.

Transcriptions in general encyclopedias in Swedish language—central to my thesis work—at the beginning exclusively used letters from the Swedish alphabet to denote speech sounds. Over time, special characters were introduced, most of them originating from the phonetic alphabets created in the late 19th century, e.g. *Det svenska landsmålsalfabetet* (The Swedish dialect alphabet, 1879) and IPA (1888). In Sweden, there exists a well-established tradition of presenting pronunciation in encyclopedias, at least for entry headwords. In the Swedish speech community, 'correct' pronunciation, notably of loanwords and foreign proper names, has usually contributed to enhancing a person's prestige. Conversely, mispronouncing a loanword, a well-known foreign name or trademark, sometimes may have a stigmatizing effect.

Background

The national romantic currents in 19th century Europe aroused an overall interest in folklore. In the Nordic countries, folk music and dialects began to be collected, No means of acoustic recording was yet available—Edison's phonograph was patented in 1877, but came onto the market only in 1890—thus, informants' speech had to be recorded by hand *in situ.* Rough phonetic transcription systems already existed, but now, linguists were urged to design alphabets that could cope with the fine details of pronunciation. In Sweden, those efforts resulted in the above-mentioned Landsmålsalfabetet—expressly conceived for capturing Swedish dialects—being published in 1879 by linguistics professor Johan August Lundell in Uppsala. It soon got its counterparts in the neighbouring countries: *Norvegia* by Johan Storm, and *Dania* by Otto Jespersen, one of the founders of the International Phonetic Association (IPA). In Finland, the renowned linguist and ethnologist Emil Nestor Setälä constructed an alphabet dedicated to recording the varieties of spoken Finnish.

Phonetic transcriptions differ

The Nordic phonetic alphabets for dialects have a high scientific level, and contain a wide range of characters. Thus, the Swedish sinologist Bernhard Karlgren—a student of Lundell's—could use Landsmålsalfabetet when researching into various Chinese dialects. Since its characters are cursive, designed to be handwritten, it resembles shorthand. By contrast, the characters of the International Phonetic Alphabet (IPA) are based on antiqua (roman) typefaces in order to facilitate use in movable-type printing. Some of the ordinary cast metal type pieces could easily be turned upside down, thus resulting in ə, ʌ, ɔ, etc.

The primary aim of the IPA was helping students of foreign languages to achieve a better pronunciation, and it became widely used for this purpose before the First World War. However, all language learners were not in secondary education or at university; many were studying on their own. Without a live model for pronunciation, it might prove difficult for them to find out what the IPA signs symbolized. Persons not accustomed to studying could also find it troublesome to learn to use the IPA. Therefore, both before and after the advent of the IPA, so-called respelled (or respelt) transcriptions have flourished. Early examples can be found in English textbooks for emigrants to the U.S., e.g. that by Jungberg (1869), recent ones in many phrase books for travellers. Berlitz nowadays to some extent are using the IPA, but still their English-speaking public are provided with transcriptions of this kind for French:

Comment allez-vous?
komahng talley voo
Nous avons un enfant
noo zavawng zang nahnfahng.

Standardization and norms

At the same time as the above-mentioned collecting of dialects was going on, in some European countries, national standard pronunciation models were being constructed for use within theatre, education and broadcasting (Siebs 1898, 1931). They often became the norm and served as bases for pronunciation dictionaries. In Sweden, no such codification took place, even though in the theatre and primary education, there was a strong admonition to allow nothing but *rikssvenska* ('National Swedish'), a concept with an unclear definition, but corresponding to an orally transmitted variety of spoken Swedish alleged to be 'free from dialect features'. This situation prevailed well into the 20th century, whereas in Norway, the opposite policy was adopted: teachers were instructed to encourage dialects and were deprived of all rights to interfere with children's pronunciation.

Sweden had to wait long before the first dedicated pronunciation dictionaries were published, viz. those by Hedelin (1997) and Garlén (2003).

Social circumstances matter

In countries like the U.S., where people are able to study and attain high social status without any knowledge of foreign languages, social pressure for 'correct' pronunciation of foreign names or expressions is not very present. In Sweden, the situation is quite different. Higher education is not accessible to those mastering only Swedish —and has never been. From the 19th century and up to the end of the Second World War, German was dominant as foreign language in secondary schools, but was then superseded by English. In the 1950s, English was introduced as a mandatory subject on the compulsory school curriculum, so that all pupils born c. 1945 or later would receive at least three years of English language education.

Of course, this meant a general rise of the educational level, but also an increasing gap between generations. Middle-aged people who had not studied English sometimes felt excluded and encountered more difficulties finding employment, especially within the service sector, and occasionally were ridiculed—even by their own children—for being incapable of pronouncing English. The same adversity affected immigrated people without English knowledge. Having realized this, Swedish authorities in the early 1970s started up beginners courses in English for adults all over the country, entirely free of charge.

The post-war boom in English promoted a demand for phonetic transcriptions, and the IPA came into use on a larger scale than ever. It featured not only in English textbooks and dictionaries, but also in French, and—to a lesser extent—German ones. Even geography schoolbooks for 11–13 year olds contained footnotes with IPA pronunciation advice.

The current status of the IPA in Sweden is hard to evaluate. On the one hand, its use at school seems to have diminished, on the other

hand, square brackets and IPA characters have come into fashion in the advertising agencies who use them on billboards and shop signs, sometimes to inform about pronunciation, but more often in an erroneous or nonsense way, just because of their decorative value.

Public service transcriptions

English and French, unlike Finnish, Turkish and Spanish, or even German and Swedish, have a relationship between writing and pronunciation that is far from direct, Given this fact, reading an English or French phrase aloud as if it were Swedish, i.e. applying the human Swedish text-to-speech algorithm, will yield a bizarre result, hardly understandable for native speakers of the languages in question. Since knowledge of French in present-day Sweden is scarce compared to that of English, there is less social pressure for mastering French pronunciation. Nevertheless, for centuries French has been considered a language of prestige in Sweden, and mispronouncing a French name would count as a more serious error for a newsreader to commit than distorting a Portuguese, Dutch or even a Finnish name.

Phonetic transcriptions for broadcasters

In Great Britain, the BBC has always been caring about pronunciation. Founded in the 1940s, its Pronunciation Unit employs phonetics experts who help announcers and newsreaders to find the most suitable pronunciation for every single instance. Many years of work has resulted in publications like *BBC Pronouncing dictionary of British names* (2nd edn 1983) and *Oxford BBC Guide to Pronunciation* (2006). In those, two parallel transcription systems are used: IPA and a respelling system, the latter of which yields a more anglicized result, apt for those not familiar with foreign languages. An example:

Sven-Göran Eriksson,
Swedish football manager
(sven yoer-an ay-rik-son).

In Sweden, the public service broadcasting companies *Sveriges Radio, Sveriges Television* and *UR* (Swedish Educational Broadcasting Company) together maintain a similar service, giving advice via a periodical leaflet, *Språkbladet,* and a database, *Dixi,* the latter accessible to staff only, on the three companies' common intranet.

Systembolaget's transcriptions

Following the abolition of the rationing book in 1955, *Systembolaget* (the Swedish Alcohol Retail Monopoly) witnessed a sales increase, in particular for strong beverages like aquavit and vodka. In accordance with its instructions not to maximize profit but rather promote public health, the company decided to launch a campaign to encourage customers to substitute wine for traditional strong liquor. However, the board envisaged a problem: a major part of the wines had French names; thus the presumptive new customers were likely to feel embarrassed at the counter, or even refrain from trying wine and stick to buying their customary bottle of *'Renat'* ('Absolutely Pure Alcohol', 40 % by volume).

The proposed solution was to include respelled phonetic transcriptions in the product catalogue. The company's sales manager, who had no training in phonetics, took on the task of composing the transcriptions, which made their debut in February 1959. Here are two examples:

Le Vallon Hanappier
(lö vallå´ng annapje´)
Château Paveil de Luze
(schatå´ pave´j dö ly´s).

Several years later, the phonetician Claes-Christian Elert, then Senior Lecturer and *Docent* at the Institute of Linguistics at Stockholm University, was called upon to give his opinion of the company's pronunciation advice. So he did (Elert 1967a), and some corrections and improvements were made. Later, another professional phonetician took over the responsibility for the transcriptions. They remained until the mid-1990s; their disappearance was probably due to the introduction of self-service outlets.

Conclusion and further work

Hopefully, the ongoing IT development will favour an extended use of the IPA, even in the United States, where some reluctance to it seems to linger. Anyway, there will probably still be a need for transcription systems more convenient for the casual user.

In my future work, by interviewing users of encyclopedias etc., I will endeavour to investigate the issue of optimizing transcriptions for various kinds of works, in order to make them render the best service to all users, not only specialists, but also a more general public.

References

BBC Pronouncing dictionary of British names, 2nd edn (1983). Pointon, G E (ed) Oxford: OUP.

Elert, C-C (1967a). En expert om våra uttalsråd. In: *Bouquet, Systembolagets personaltidskrift* 2 (1967) 32–34.

Elert C-C (1967b). Uttalsbeteckningar i svenska ordlistor, uppslags- och läroböcker. In: *Språkvård* 1967:2. Stockholm: Nämnden för svensk språkvård. (Also in: Studier i dagens svenska in: *Skrifter utgivna av Nämnden för svensk språkvård* 44. Stockholm 1971: Läromedelsförlagen.)

Garlén C (2003). *Svenska språknämndens uttalsordbok: 67 000 ord i svenskan och deras uttal.* Stockholm: Svenska språknämnden: Norstedts ordbok.

Hedelin P (1997). *Norstedts svenska uttalslexikon*. Stockholm: Norstedts.

Jungberg, C G (1869) Utvandrarens tolk: Praktisk lärobok i engelska språket. Stockholm: Ebeling & K:i.

Olausson L & Sangster C (2006). *Oxford BBC Guide to Pronunciation.* Oxford: OUP.

Parkvall M (2009). *Lagom finns bara i Sverige och andra myter om språk.* Stockholm: Telegram Bokförlag.

Rosenqvist, H (2004). Markering av prosodi i svenska ordböcker och läromedel. In: Ekberg, L & Håkansson, G (eds) *Nordand 6. Sjätte konferensen om Nordens språk som andraspråk.* Lund.

Siebs Th (1898). *Deutsche Bühnenaussprache.* Köln: Verlag Albert Ahn.

Siebs Th (1931). *Rundfunkaussprache.* Berlin: Reichs-Rundfunk-Gesellschaft.

Viner och spritdrycker: Prislista D, 1 februari 1959 (1959). Stockholm: Nya Systembolaget Återtryck.

Contrastive analysis through L1-L2map

Preben Wik[1], Olaf Husby[2], Åsta Øvregaard[2], Øyvind Bech[2], Egil Albertsen[2], Sissel Nefzaoui[2], Eli Skarpnes[2], Jacques Koreman[2]
[1] Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
[2]Department of Language and Communication Studies, NTNU, Trondheim, Norway

Abstract
This paper describes the CALST project, in which the primary aim is to develop Ville-N, a computer assisted pronunciation training (CAPT) system for learners of Norwegian as a second language. Ville-N makes use of L1-L2map, a tool for multi-lingual contrastive analysis, to generate a list of language-specific features. These can be used to tailor pronunciation and listening exercises. The tool can also be used for other target languages.

Introduction

CALST (Computer-Assisted Listening and Speaking Tutor) is a project headed by the Norwegian University of Science and Technology (NTNU) in collaboration with The Royal Institute of Technology (KTH), the University of Oslo (UiO), The Adult Education Centre (EVO) in Trondheim, and. The project's aim is to develop a computer-assisted pronunciation training (CAPT) system for Norwegian as a second language (NSL). The project has two main goals:

1) To develop a Norwegian CAPT system, Ville-N, based on Ville, the virtual language teacher for Swedish, developed at KTH (Wik, 2004). The system should be able to cater for L2-learners with a wide range of communicative abilities, ranging from foreign university students at NTNU and UiO who are proficient in English, to illiterate users at EVO with no English skills. The program should be used with minimal requirements on instruction or supervision.

2) To create a database and tools for contrastive phonetic and phonological analysis for all relevant L1-L2 pairs, partly as a foundation for future research in second language acquisition (SLA), and partly in order to better tailor exercises used in Ville-N.

Ville-N will be used to complement pronunciation teaching in the Norwegian courses both for foreign students and employees at NTNU and UiO, as well as in several courses for teachers of Norwegian as a Second Language. In addition, the collaboration in the project with EVO widens the target group to also include L2-learners that are illiterate and from other language and social backgrounds than what is found at the universities.
Several extensions to the Swedish Ville have been made in order to accommodate the different needs of the Norwegian system.

Norwegian dialects

The Norwegian language situation is quite different from that of Sweden and many other language communities in that there is no accepted pronunciation standard in Norwegian. Although there is a common form taught in adult second language classrooms called Bokmål (Urban East Norwegian, UEN), different dialects are used both in formal and informal situations. This creates a serious problem for L2-learners of Norwegian since there are large pronunciation differences in the various dialects and often different words are used to express the same meaning. This dialectal variation cannot easily be addressed in standard language courses, and a need has thus been identified to better equip L2-learners to deal with everyday communicative situations where variation in the speakers' dialect is typical

Multiple speakers

To address this problem recordings of multiple speakers have been made. One male and one female speaker of the dialects in the following regions of Norway have been recorded: Northern Norway, Trøndelag, Western/Southern

Norway and Southeastern Norway (UEN). All in all 8 different voices will be used in the program, accompanied by 8 different embodied conversational agents (ECAs), to give each voice a personality.

The learner can select an ECA from the GUI and listen to and practice one specific dialect, or train across dialects in the same exercise and let the program select the target voice.

As reported by several researchers, including McAllister (1998), it is good to listen to many different speakers in order to achieve better listening comprehension. Multiple-talker models have been reported as particularly effective to improve perception of novel contrasts (c.f. Logan et al., 1991; Probst et al., 2002), as the inherent variability allows for induction of general phonetic categories or other L2 specific salient features.

Learners will also be able to run the exercises with one specific dialect in mind, which is useful when selecting a role model for pronunciation in production exercises. The advantage of having both male and female voices in each dialect becomes apparent in this case to allow learners to choose a role model with the same gender as themselves.

Hyper-articulated recordings

The native speakers who made the recordings were informed to speak as they normally do with normal speed, reductions and coarticulations, in order to offer the learners spoken utterances that are as close to authentic speech as possible. It is however difficult for beginners to assimilate reductions and coarticulations in the early phases of learning, and it is often common practice to speak slower and clearer when speaking to L2 learners who are in an early phase of development. To incorporate such considerations in the program two options were considered: To make it possible for the learners within the program to slow down the speech samples by manipulating the acoustic signal, or to make double recordings, one normal and one slow hyper-articulated version of all the recordings. The difference between slow and normal speech is not uniformly distributed, as for example plosives are not stretched in the release burst but only in the occlusion. Long vowels are typically exaggerated, coarticulations will be reduced or removed, and other aspects of the speech such as the lexical stress will be affected differently with stressed syllables being stronger and more emphasized in a hyper-articulated version. It was hence decided that to ensure the best possible quality in the learning material, exaggerated, hyper-articulated versions of all recordings were made even if it would include more work and a larger set of recordings.

The learners have the option to choose to do the exercises using either type of recording, or a learner can choose to use normal recordings and select a: “Say again” button, to get a repetition of the last spoken word in a hyper-articulated version.

Wordlists

A wordlist of basic vocabulary from the course books used by the participating institutions has been created and categorized into semantic categories. Approximately 1000 words were selected from the aggregated wordlists and divided into 43 categories. The criterion for the selection of the base vocabulary was also that it should satisfy the A1 and A2 vocabulary range of CEFR, (Common European Framework of Reference for Languages)

All words have then been visualized. Approximately 30% of the images are from “UVic's Language Teaching Clipart Library”. For words where no appropriate image was found complementary drawings have been made by a local artist in the same artistic style as the drawings from the internet, in order to get a consistent and coherent set of images.

English translations, transcription and the inflection of words have been added, and as mentioned above, sound files have been recorded for the aggregated wordlist, in four dialects, with one male and one female speaker for each dialect, and in both normal and hyper-articulated versions.

The L1-L2map

The contrastive analysis hypothesis (CAH) as presented in Lado (1957), claims all problems in learning a foreign language can be explained from transfer problems induced by the learner’s native language. It is today generally accepted that the claims made by the CAH are too strong, and that there are other factors which determine the difficulty language learners have with acquiring new sounds (Eckman, 1977; Odlin, 1989; Flege, 1995; Major, 2001). This does however not mean that CAH should be

completely rejected or abandoned. As stated by Ellis (1994)

"*The problem with CAH is that it is too simplistic and too restrictive. The solution as many researchers have come to recognize, lies not in its abandonment but in careful extension and revision"*

The second task in the CALST project is to design and evaluate a revised and extended contrastive analysis tool called L1-L2map. Since the problematic aspects of the L2 are not the same for all learners, the aim of the CAPT system is to assist learners (and teachers) in identifying the problematic aspects for each individual, and work on these contrasts with special exercises. The idea with L1-L2map is to make the CAPT system informed by L1 specific filtering.

The L1-L2map will serve as a platform for researchers with a phonetic background to encode language data and make it available in a format that can be used by CAPT creators. L1-L2map is designed as a wiki with two levels of users.

First of all as a generally accessible tool, where any user can access the data and browse and compare the phonological features of different languages. A group of specialists will have administrative rights and the responsibility for inserting feature data about languages that they have phonetic-linguistic competence for. Each language is thus encoded individually by an expert in that particular language, and the contrastive analysis is performed by 'superimposing' the data for two languages on top of each other.

The L1-L2map can carry out an automatic contrastive analysis where the source language can be chosen from a large number of languages, with (one of the dialects of) Norwegian as the target language. The first version of the L1-L2map is based on the UPSID database, which contains 451 languages (Maddieson, 1980). This number has been increased to more than 500 languages in L1-L2map. As shown in Figure 1, the result of the contrastive analysis is displayed in four tabs.

The choice of tabs is based on the IPA representation: "Consonants", "Consonants (other)", "Vowels" and "Diphthongs". The "Consonants" tab presents pulmonic consonants, while non-pulmonic consonants and affricates are presented in "Consonants (other)". The latter is only shown if relevant for the L1-L2 comparison. For the sake of simplicity rows in the consonant tables (representing manner of articulation) are only visualized if used in at least one of the languages. A fifth, "Language information" tab presents some general information, following the information given in UPSID. A lay-out which is very similar to that used in IPA was used in order to provide language experts using the system with a simple and recognizable lay-out.

As shown in Figure 1, a color scheme is employed in the visualization of the data. The first language chosen is blue and assumed L1 status, the second language chosen is red and assumed L2 status, and the features they have in common (where there is an overlap) are green. This way it becomes apparent for the L2 learner which features are different from their L1 (red), and thus needs attention since they are absent in the L1.

Extensions to UPSID

The UPSID database only lists sounds that are distinct phonemes in any given language. It is not enough to do a simple comparison of which sounds constitute phonemes in each language. A number of extensions are in the process of being added to the L1L2-Map. These include positional restrictions, syllable structure (phonotactics), tone, stress, and timing.

Positional phonemic restrictions

The positions in which sounds occur in syllables must be taken into account. Difficulties that could be predicted from CAH if position is part of the description, will otherwise go by unnoticed. For example in Mandarin, only two consonants ([ŋ] and [n]) are allowed at the end of a syllable, even if many other consonants appear in syllable initial position. A consequence of this is that many consonants which can occur at the end of Norwegian syllables present a difficulty for learners of Norwegian with Mandarin as L1.

The L1L2-map presented in Figure 1 is the complete set of consonants in both Norwegian and Mandarin, taken from the UPSID data, where positional restrictions are not taken into account. In the next version of L1-L2map, this pane will be divided into three separate panes, with each pane only displaying the phonemes that are allowed in initial medial, and final position respectively.

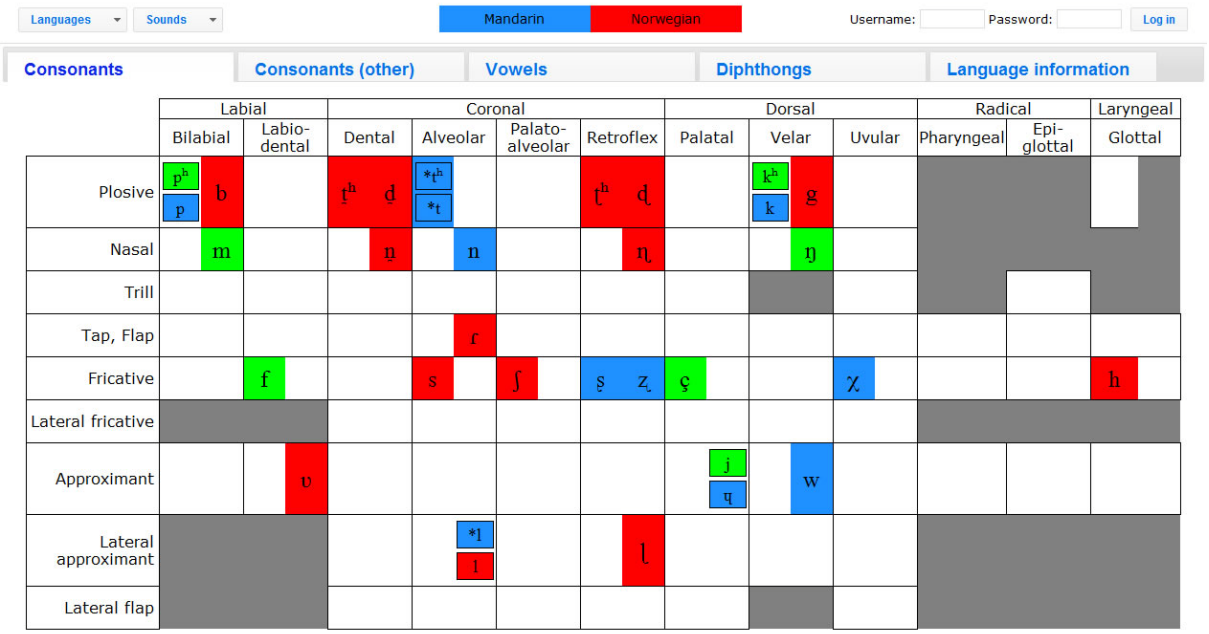


Figure 1 Visualization of the consonant part of the L1-L2map. Here exemplified with Mandarin as L1 in blue, Norwegian as L2 in red. Overlapping phonemes displayed in green.

Syllable structure (phonotactics)

Not only positional restrictions, but also the phonotactic constraints of languages (restrictions on the permissible combinations of phonemes) will be encoded in the L1L2-map. Norwegian has a relatively complex syllable structure comparable to that of Swedish, and constitute a difficult part of acquiring the language for many learners.

Permissible consonants and consonant clusters in onset and coda will be encoded into separate lists in the database, and when a contrastive analysis is performed between the learners L1 and the L2 clusters that are allowable combinations in the L2 but missing from the learners L1 will be extracted.

L1-L2map is a useful tool for developers of CAPT systems for any language, as well as for language teachers. The tool allows the insertion of new languages and/or dialects, and outputs useful information about the L2 phones which L2 learners need to acquire.

The L1-L2map is available online: http://calst.hf.ntnu.no/l1-l2map

Acknowledgements

We are grateful for financial support from Norgesuniversitetet, project number P54/2009, and to Henning Reetz for making the scripts for his UPSID interface (web.phonetik.uni-frankfurt.de/upsid.html) available.

References

Eckman, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning, 27*(2), 315-330.

Ellis, R. (1994). *The Study of Second Language Acquisition.* Oxford University Press.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 233-277.

Lado, R. (1957). *Linguistics across cultures.* University of Michigan Press.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/and/l: A first report. *Journal of the Acoustical society of America, 89*(2), 874-886.

Maddieson, I. (1980). *UPSID: UCLA phonological segment inventory database.* Phonetics Laboratory, Department of Linguistics.

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology.* Lawrence Erlbaum.

McAllister, R. (1998). Second language perception and the concept of foreign accent. In *STiLL-Speech Technology in Language Learning*.

Odlin, T. (1989). *Language Transfer.* Cambridge University Press.

Probst, K., Ke, Y., & Eskenazi, M. (2002). Enhancing foreign language tutors-In search of the golden speaker. *Speech Communication, 37*(3-4), 161-173.

Wik, P. (2004). Designing a virtual language tutor. In *Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004* (pp. 136-139). Stockholm University.

Tone restricts F_0 range and variation in Kammu

Anastasia Karlsson[1], *Jan-Olof Svantesson*[1], *David House*[2], *Damrong Tayanin*[1]
[1]*Dept. of Linguistics and Phonetics, Lund University;* [2]*Dept. of Speech, Music and Hearing, KTH, Stockholm*

Abstract

The aim of this study is to investigate whether the occurrence of lexical tones in a language imposes restrictions on its pitch range. We use data from Kammu, a Mon-Khmer language spoken in Northern Laos, which has one dialect with, and one without, lexical tones. The main finding is that speakers of the tonal dialect have a narrower pitch range, and also a smaller variation in pitch range.

Introduction

There is recurrent interest in comparing the F_0 range of different languages in the broad context of investigating language-specific use of F_0. There has been general speculation that different pitch ranges and other characteristics of F_0 can comprise a part of the phonetic structure of a language and thus differ systematically between languages (see Traunmüller & Eriksson (1993) and Keating & Kuo (2010) for reviews). One question concerns the influence of lexical tone on intonation, and this has generated the hypothesis that tone languages may have an overall larger F_0 range than non-tonal languages by virtue of the additive effect of the lexical tones being superimposed on the intonation contour. Several studies have supported this hypothesis, while in other studies no difference in pitch range between tonal and non-tonal languages was found. In some studies, the opposite tendency has been observed where tone languages display a smaller F_0 range.

In many of the studies supporting the hypothesis, Standard Chinese has been compared with English. In a study of broadcast news speech (Yuan & Liberman, 2010), it was found that Standard Chinese has a wider pitch range and more F_0 fluctuations than English. This is discussed in terms of the effect of lexical tones.

In Zhang & Tao (2008), where a bilingual Chinese-English corpus was used to develop a mixed-language speech synthesis system, the pitch range of the English words was larger in the bilingual corpus than in the English one. These results are discussed in terms of the influence of the Chinese lexical tones on the corpus.

In Keating & Kuo (2010), Standard Chinese was found to have a larger pitch range than English in single-word utterances. However, this effect was not seen in prose passages. These results highlight the effect of speech material. Eady (1982) found no difference in F_0 standard deviations between English and Standard Chinese.

Another interesting and relevant area of study is the modification of F_0 which takes place in infant directed speech. Grieser & Kuhl (1988) reported an exaggeration of F_0 range in infant directed speech in Standard Chinese. However, in a study comparing infant directed speech in Australian English to Thai (Kitamura et al., 2001) it was found that F_0 range was more exaggerated in Australian English than in Thai. These results are discussed in terms of restriction on pitch excursions in infant directed speech due to lexical tone.

Lexical tone can thus be seen to either restrict F_0 range or enhance it, varying across language, speech material, and speaking style. By investigating a language in which lexical tone is a characteristic of one dialect but absent from another dialect, we aim to study the effect of lexical tone on F_0 range.

Kammu is a Mon-Khmer language spoken by some 600,000 people, mainly in Northern Laos, but also in adjacent areas of Vietnam, Thailand and China. One of its main dialects has lexical tones (high or low) on each syllable, while the other main dialect lacks lexical tones. The tones have developed by the merger of voiceless and voiced initial consonants. Other differences between the dialects are marginal, and speakers of different dialects understand each other without difficulty (Svantesson, 1983; Svantesson & House, 2006).

Earlier studies of Kammu have shown a compressed F_0 range in the tonal dialect as compared to the non-tonal dialect in spontaneous speech (Karlsson et al., 2011), as well as in planned speech (House et al., 2009;

Karlsson et al., 2010). In this study we make a more systematic study of F_0 range differences in a planned speech material.

Method

Recordings of 14 speakers of the tonal dialect and 9 speakers of the non-tonal dialect were used in this investigation. The subjects were recorded in Laos and Thailand using a portable Edirol R-09 digital recorder and a lapel microphone. The utterances were digitized at 48 kHz sampling rate and 16-bit amplitude resolution and stored in .wav file format. Most of the speakers were recorded in quiet hotel rooms. One speaker was recorded in his home and one in his native village.

Since Kammu is an unwritten languge, the material was presented written in Lao or Thai, and the speakers were asked to translate it into Kammu. Almost all Kammu speakers in Laos and Thailand are bilingual and have received their school education in Lao or Thai, respectively. Thus there was some variation in the recorded material. The resulting utterances were checked and transcribed by one of the authors, Damrong Tayanin, who is a native speaker of Kammu.

The following sentences from our material were used in this investigation (given in both dialect forms):

Táʔ Kàm kùuŋ táaj ʔòʔ.
Taʔ Kam guuŋ taaj ʔoʔ.
Mr Kàm saw my brother.

Táaj ʔòʔ kùuŋ táʔ Kàm.
Taaj ʔoʔ guuŋ taʔ Kam.
My brother saw Mr Kàm.

Màʔ mʌʌt kʰɔ́ɔŋ ʔòʔ? Táʔ Kàm mʌʌt kʰɔ́ɔŋ ʔòʔ.
Maʔ mʌʌt kʰɔɔŋ ʔoʔ? Taʔ Kam mʌʌt kʰɔɔŋ ʔoʔ.
Who took my things? Mr Kàm took my things.

Kàə məh məʔ? Kàə məh kɔ́ɔn ʔòʔ.
Gəə məh məʔ? Gəə məh kɔɔn ʔoʔ.
Who is he? He is my child.

Kìi məh máh? Kìi məh kláaŋ.
Gii məh hməh? Gii məh klaaŋ.
What is this? This is an eagle.

Kìi məh máh? Kìi məh tàaŋ.
Gii məh hməh? Gii məh daaŋ.
What is this? This is a lizard.

The underlined words were used in the investigation. For each of these words, the maximum and minimum F_0 value was measured, using the *Praat* program, and the F_0 range over the word was computed as the difference (in semitones) between the maximum and the minimum.

Results

Table 1. Means and standard deviations of the F_0 ranges (semitones).

word	mean	sd	N
Kàm/Kam	2.20/1.88	1.72/1.32	16/15
kùuŋ/guuŋ	1.09/1.77	1.15/0.97	16/15
ʔòʔ/ʔoʔ	1.40/1.85	1.07/1.32	16/15
kùuŋ/guuŋ	1.45/1.90	2.05/0.74	32/18
Kàm/Kam	1.33/3.76	1.21/3.50	32/18
màʔ/məʔ	1.36/1.58	1.44/1.63	24/16
ʔòʔ/ʔoʔ	0.39/2.70	0.44/2.29	7/16
Kàm/Kam	2.08/4.23	1.63/3.29	19/15
ʔòʔ/ʔoʔ	0.38/1.87	0.56/1.50	19/15
kàə/gəə	0.68/1.33	0.91/0.92	28/20
màʔ/məʔ	1.42/1.76	1.35/1.19	28/20
kàə/gəə	0.29/1.29	0.46/1.09	28/18
kɔ́ɔn/kɔɔn	0.55/0.61	0.57/0.69	28/18
ʔòʔ/ʔoʔ	1.79/2.03	1.38/1.90	28/18
kìi/gii	0.03/1.86	0.17/1.59	28/23
máh/hməh	1.84/4.23	2.04/2.95	28/23
kìi/gii	0.07/1.65	0.22/1.61	29/14
kláaŋ/klaaŋ	3.88/6.66	3.34/3.68	29/14
kìi/gii	0.25/0.32	0.388/0.393	30/20
máh/hməh	2.59/3.87	2.41/3.01	30/20
kìi/gii	0.27/0.91	0.44/1.35	34/15
tàaŋ/daaŋ	2.89/3.68	1.88/2.55	34/19

The results of the measurements are shown in Table 1, which shows the mean and standard deviation of the range for each word in the material, shown in the table in the same order as they are presented above. The number of repetitions of each word is shown as well. The word or number before the slash refers to the tonal dialect, and those after the slash refer to the non-tonal dialect.

It can be seen from the table that except for one word (the very first word *Kàm/Kam*), the mean of the range is greater for the non-tonal dialect than for the tonal dialect. In 21 cases of 22, the non-tonal speakers have greater mean range that the tonal ones, and a binomial test gives a highly significant result ($p < 0.0001$). The standard deviation is larger for the non-tonal dialect in 18 cases of 22 (exceptions are *Kàm/Kam* (1) *kùuŋ/guuŋ* (1), *kùuŋ/guuŋ* (2) and *məʔ/məʔ* (2)), giving a significant result of the binomial test ($p = 0.0043$).

Discussion

The results show that the F_0 range over a word (measured in semitones) is, on the average, larger in the non-tonal dialect than in the tonal dialect. Furthermore, there is greater variation in the ranges in the non-tonal dialect than in the tonal dialect, as the standard deviations show. This is consistent with earlier findings (House et al., 2009; Karlsson et al., 2010; 2011; Karlsson, 2011), which also show, in different situations, that the F_0 range is smaller in the tonal than in the non-tonal dialect. These results are also in line with those found for infant-directed speech in Kitamura et al. (2001) where F_0 range was more exaggerated in Australian English than in Thai. It could be that in more engaged speech, e.g. infant-directed and spontaneous, lexical tones become more restrictive in their influence on the intonation contour.

Our result is opposite to what was found for Chinese when compared to English (Yuan & Liberman, 2010; Zhang & Tao, 2008) where the presence of lexical tones results in an expanded F0 range. One explanation could be that Kammu has a simpler tone system with only two level tones while Chinese has a more complex system with contour tones. In Kammu the difference between the low and high tone is often relatively small (Svantesson & House, 2006) which may also restrict the use of large pitch excursions.

In Karlsson et al. (2010), we present data that strongly suggest that the intonational systems of the two Kammu dialects are basically identical, but also that there is a prosodic hierachy, where lexical tone is stronger than sentence accent, which in its turn is stronger than focal accent. It seems to be necessary to uphold the contrast between the lexical tones in the tonal dialect, and when this conflicts with other uses of F_0, such as for marking sentence or focal accent, this may be inhibited.

These restraints on the use of F_0 for intonation may be the explanation for the results found here, that there is generally a smaller pitch range in the tonal dialect than in the non-tonal dialect, and also for the fact that there is less variation in the range.

References

Eady S J (1982). Differences in the F0 patterns of speech: Tone language versus stress language. *Language and speech,* 25: 29-42.

Grieser D L, Kuhl P K (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental psychology,* 24: 14-20.

House D, Karlsson A, Svantesson J-O, Tayanin D (2009). The phrase-final accent in Kammu: effects of tone, focus and engagement. *Proceedings of Interspeech 2009,* Brighton, 2439-2442.

Karlsson A (2011). Prosodic features of Kammu tonal and non-tonal dialects: read and spontaneous speech. In: Endo M, Saitô Y, eds, *Tone, accent and intonation in eastern Eurasian languages.* The 18th Meeting of the Linguistic Circle for the Study of Eastern Eurasian Languages, Aoyama Gakuin University, Tokyo, 19-28.

Karlsson A, House D, Svantesson J-O, Tayanin D (2010). Influence of lexical tones on intonation in Kammu. *Proceedings of Interspeech 2010,* Makuhari, Japan, 1740-1743.

Karlsson A, House D, Svantesson J-O, Tayanin D (2011). Comparison of F_0 range in spontaneous speech in Kammu tonal and non-tonal dialects. *Proceedings of ICPhS 2011.*

Keating P, Kuo G (2010). Comparison of speaking fundamental frequency in English and Mandarin. *UCLA working papers in phonetics,* 108: 164-187.

Kitamura C, Thanavishuth C, Burnham D, Luksaneeyanawin S (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant behavior and development*, 24: 372-392.

Svantesson J-O (1983). *Kammu phonology and morphology.* Lund: Gleerup.

Svantesson J-O, House D (2006). Tone production, tone perception and Kammu tonogenesis. *Phonology,* 23: 309-333.

Traunmüller H, Eriksson A (1993). F0-excursions in speech and their perceptual evaluation as evidenced in liveliness estimations. *Perilus,* 17: 1-34.

Yuan J, Liberman M (2010). F0 declination in English and Mandarin broadcast news speech. *Proceedings of Interspeech 2010,* Makuhari, Japan, 134-137.

Zhang Y, Tao J (2008). Prosody Modification on Mixed-Language Speech Synthesis. *Proc. 6th International Symposium on Chinese Spoken Language Processing, ISCSLP2008,* Kunming, 253-256.

Visualizing Prosodic Densities and Contours: Forming One from Many

Daniel Neiberg
TMH/CSC/KTH

Abstract

This paper summarizes a flora of explorative visualization techniques for prosody developed at KTH. It is demonstrated how analysis can be made which goes beyond conventional methodology. Examples are given for turn taking, affective speech, response tokens and Swedish accent II.

Introduction

The ability to empirically extracting the essence of a phenomenon is a prerequisite for understanding in natural sciences. The concept of essence originates from Aristotle's *essentia*, originally phrased as "to ti ên einai", literally "the what it was to be" (Cohen, 2011). It refers to the attribute or set of attributes that make an object or substance what it fundamentally is, and which it has by necessity, and without which it loses its identity. In phonetics, the prosodic essence is usually attributed to the average, standard deviation or slope of fundamental frequency, intensity or duration for specific segments of speech. The usefulness of such analysis is not only motivated by its success, but also to the acceptance within the community and compliance with the current paradigm of thought.

In recent years, a number of explorative visualization techniques of fundamental frequency (F0) and intensity have emerged. The common idea is to generate a single instance contour or density cloud out of many instances. Applications include exploring the prosody of backchannels and in the vicinity of thereof, turn-taking and variation of accent between dialects. As pointed out, the different techniques can be divided into methods which are intended to produce prosodic density maps and methods which are intended to produce contours. The underlying machinery can then roughly be divided into segment based vs. instantaneous approaches and parametric vs. non-parametric. Another division can made between generic approaches which takes an arbitrary feature as input vs. those which targets a specific part of the signal. A summary of these techniques are shown in Table 1.

This article briefly describes each of these techniques and gives examples. Finally, the usefulness and drawbacks are discussed.

Non-parametric Density Generating Methods

The Bitmap technique uses contours of F0 (in a semitone scale) or intensity and plots these with partially transparent dots. The accumulation of these dots from multiple instances forms a density cloud. Prior to plotting, the F0 or intensity is typically aligned to the start or end of a speech segment.

The technique was introduced in (Edlund, et al., 2009; Heldner, et al. 2009), and an example is given in Figure 1. Variations of the same technique has been explored by (Schötz, et al., 2010; Ward, et. al., 2010).

The Bitmap technique is a non-parametric density generating method, which is also generic in the sense that any acoustic feature can be used for input.

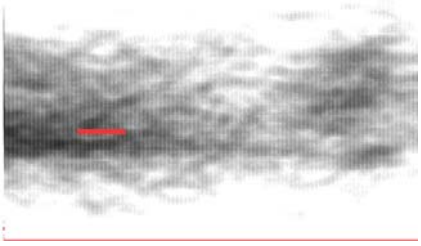


Figure 1. Bitmap plot of the 1000ms of the interlocutor's speech preceding each of 3054 backchannels as approximated by VSU-Low. The line shows the 26th percentile of the speakers' pitch from 860 to 700ms prior to the VSU-Low, which is an elicitation cue. Reproduced from (Edlund, et al., 2009) with permission.

Table 1. Summary of methods for visualizing one density or contour from multiple instances.

Method	*Shows*	*Feature/target*	*Parameterization/model*	*References*
Bitmap	*Density*	*Generic*	*None*	*(Edlund, et al., 2009; Heldner, et al. 2009; Schötz, et al., 2010; Ward, et. al., 2010)*
TVCQCC	*Density*	*Full spectra*	*2D LI-DCT*	*(Neiberg, et al., 2010a)*
FFV	*Density*	*F0 delta spectra*	*HMM*	*(Laskowski, et al. 2008, Heldner, et al. 2009)*
TVCQCC	*Density*	*F0 spectra*	*2D LI-DCT*	*(Neiberg, et al., 2010b)*
LI-DCT	*Contour*	*Generic*	*1D LI-DCT*	*(Gustafson, et al. 2010)*
minJerk	*Contour*	*F0*	*2D LI-DCT / minJerk*	*(Neiberg, et al. 2011)*

Parametric Density Generating Methods

Parametric density methods introduce assumptions via a parametric model. Two main themes are identified: segment based and instantaneous approaches.

In the instantaneous approach, the target is the frame-by-frame variation of fundamental frequency. This variation can be computed directly by the Fundamental Frequency Variation spectrum (FFV) without using a conventional pitch tracker. By applying a Hidden Markov Model (HMM), sequences of FFV can be learned for the speech segments of interest. Such an approach has adopted in (Laskowski, et al. 2008) to discover sequences which trigger turn-taking. The same approach was also adopted in (Heldner, et al. 2009) to investigate prosody in different types of overlapped speech.

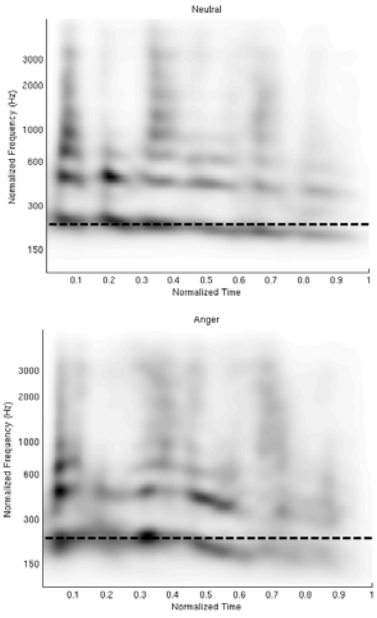


Figure 2. Spectral densities from each generated from affective speech from 40 speakers: Neutral on the top and Anger on the bottom. The dashed line is an arbitrary chosen mean F0.

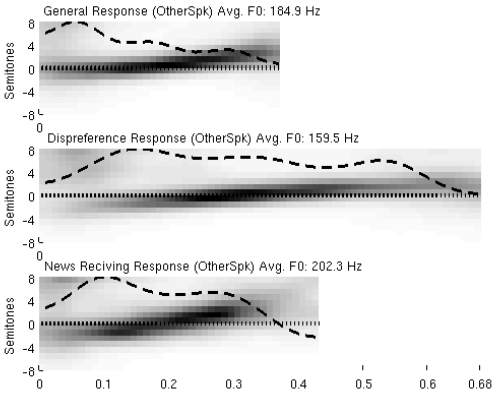


Figure 3. Three types of attitudes in response tokens. F0 is shown as densities relative to average F0 stretched to average duration. Dashed lines are normalized intensity. LDA analysis of the underlying TVCQCC gave that these attitudes can be classified based on prosody.

In the segmental approach, examined targets are the segmental variation of either the entire F0 normalized spectrum or the region around F0. The entire F0 normalized spectrum was explored as Time Varying Constant-Q Cepstral Coefficients (TVCQCC) for affective speech in (Neiberg, et al., 2010a). Two examples for Neutral and Anger are given in Figure 2, where each *prototypical spectrogram* is computed from 40 speakers. The time varying parameterization is based on a length invariant discrete cosine transform (LI-DCT). Since the basis functions are periodic, the interpolation in time gives good interpolation of the syllabic rhythm is speech. Further, the length invariance allows for separation of duration or speaking rate. The spectrograms are derived by taking the average of multiple instances in the TVCQCC domain, followed by inverse transformation. The procedure involves a compression by only

keeping the higher order coefficients where most of the variation is found.

A special case of TVCQCC was used to explore the prosody of conversational grunts, i.e. non-lexical conversational tokens such as filled pauses and backchannels in (Neiberg, et al., 2010b). This time only the part located ± 8 semitones from the per segment average F0 was kept. The procedure is shown in Figure 3, where news receiving responses and dispreference responses are contrasted to general responses.

These parametric approaches are not generic, but are rather specialized towards specific parts of the speech signal.

Parametric Contour Generating Methods

In contour generating methods, one single contour is generated out of many. In (Gustafson et al., 2010) a generic contour generating method was introduced to visualize the prosody of "mhm" as spoken by the radio host Täppas Fogelberg. The procedure uses the same length invariant DCT (LI-DCT) as for TVCQCC, but it is applied to F0 or intensity from a pitch tracker, which makes the method generic. This procedure is schematically sketched in Figure 4. By assigning each "mhm" a variable *x* which is the relative position of the token in each call, the contours in Figure 5 was obtained. The variation of the contours as a function of *x* was interpreted as the degree of engagement.

As pointed out earlier, the LI-DCT has many useful properties: 1) The DCT basis functions are periodic which allows good interpolation of syllabic rhythm in speech. 2) The length-invariance gives normalization for duration or speaking rate. This allows for separation of this feature in the analysis. 3) The 0'Th coefficient is equal to the arithmetic average, which means if it is omitted, and then only the relative shape of a trajectory is parameterized. This property is useful for parameterizing features such as F0 (which has a speaker dependent additive bias), Intensity (which is dependent on the distance to the microphone). It should be pointed out that the author has not been able make polynomials to work for this type of visualization.

A completely different contour generating approach was introduced in (Neiberg, 2011). Based on the prosodic densities obtained through the special case of TVCQCC, the contour is generated by letting a particle move forward in time though the density. The path is chosen by maximising the voicing amplitude while minimizing the jerk, which is the derivate of acceleration. These criteria simulate the hand motor-movement used when a trained phonetician is sketching the F0 contour from left to right on a paper.

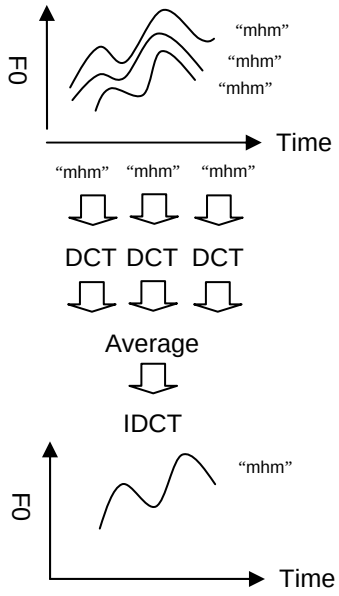


Figure 4. Estimation procedure for a generic and parametric contour generating method used in (Gustafson et al., 2010).

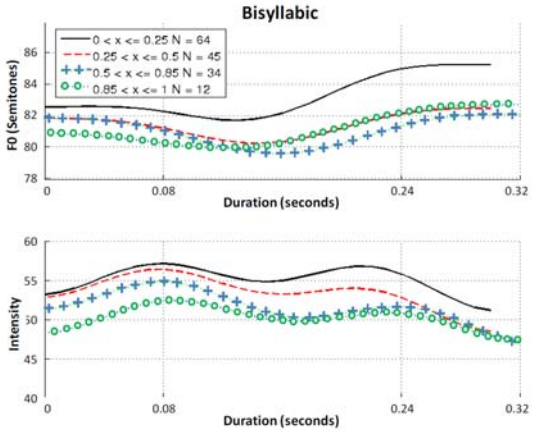


Figure 5. Täppas Fogelbergs "mhm" as a function of relative position "x" in each call. Two observations are 1) average F0 drops and curve becomes falter with respect to the second syllable 2) average intensity drops and the peak of the second syllable drops.

As an application, this model was applied to accent II in Swedish long compound words. By geographically selecting the speech material according to four dialect areas, the assignment and the number of speakers per type became

South: Type1A (N = 12); Gotland: Type1B (N = 4); East: Type2A (N = 24) and West: Type2B (N = 24). Figure 6 shows a comparison between the manually sketched pitch tracks for the four types suggested by (Gårding and Lindblad, 1973) and the automatically generated contours. One can see that the estimates from the proposed algorithm are close to the original sketched pitches. Finally, it should be pointed out that the resulting minimum jerk trajectory is not parametric, although the spectral density for which it travels through is parametric.

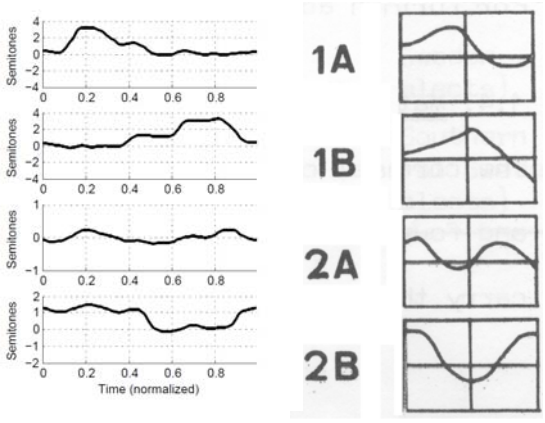


Figure 6. To the left: automatic F0 sketches and models for each dialect area. Speakers per type are N = 12 for Type1A, N =4 for Type1B, N = 24 for Type2A and N = 24 for Type2B. To the right, the corresponding manual sketches by (Gårding and Lindblad, 1973),reproduced with presmission.

Discussion

The survey shows that different types of methods are useful for different needs and applications. The non-parametric methods clearly have the advantage of using fewer assumptions than the parametric methods. On the other hand, the parametric approaches allows for the use of machine learning, clustering and distance calculations between discovered categories. The two types of generic approaches, the bitmap and the LI-DCT are least complicated to implement while still being generic and should be expected to gain some spread. It is not obvious to choose between density and contour generating approaches. Contour generation implies that the variance shown in the density generating methods is ignored. Thus, the choice must be up to proper judgment. Possibly, a hybrid method such as the "minJerk" approach might offer a solution if the resulting contour is overlaid on top of the density cloud.

Conclusions

This survey has categorized explorative visualization methods for prosody. Examples have been given, and strengths and drawbacks have been discussed. Based on the results these tools have produced, they seem to offer good opportunities for phonetic research, in particular for prosody.

References

Cohen S. Marc, "Aristotle's Metaphysics", Stanford Encyclopedia of Philosophy, accessed 3 Maj 2011.

Edlund, J., Heldner, M., & Pelcé, A. (2009). Prosodic features of very short utterances in dialogue. In Vainio, M., Aulanko, R., & Aaltonen, O. (Eds.), Nordic Prosody - Proceedings of the Xth Conference (pp. 57 - 68). Frankfurt am Main: Peter Lang.

Gårding, E. and Lindblad, P., "Constancy and variation in Swedish word accent patterns," in Working Papers 7, Lund, Lund University, 36–110, 1973.

Gustafson, J., & Neiberg, D. (2010). Prosodic cues to engagement in non-lexical response tokens in Swedish. In DiSS-LPSS.

Heldner, M., Edlund, J., Laskowski, K., & Pelcé, A. (2009). Prosodic features in the vicinity of pauses, gaps and overlaps. In Vainio, M., Aulanko, R., & Aaltonen, O. (Eds.), Nordic Prosody - Proceedings of the Xth Conference (pp. 95 - 106). Frankfurt am Main: Peter Lang

Laskowski, K., Edlund, J., & Heldner, M. (2008). Learning prosodic sequences using the fundamental frequency variation spectrum. In Proceedings of the Speech Prosody 2008 Conference (pp. 151-154). Campinas, Brazil: Editora RG/CNPq.

Neiberg, D., Laukka, P., & Ananthakrishnan, G. (2010a). Classification of Affective Speech using Normalized Time-Frequency Cepstra. In Prosody 2010.

Neiberg, D., & Gustafson, J. (2010b). The Prosody of Swedish Conversational Grunts. In Interspeech 2010, Special Session on Social Signals in Speech.

Neiberg, D., Ananthakrishnan, G., & Gustafson, J. (2011). Tracking pitch contours using minimum jerk trajectories. Submitted to Interspeech.

Schötz, S., Beskow, J., Bruce, G., Granström, B., & Gustafson, J. (2010). Simulating Intonation in Regional Varieties of Swedish. In Speech Prosody 2010. Chicago, USA.

Ward, N. G. and McCartney, J. L. Visualization to support the discovery of prosodic contours related to turn-taking. Technical Report UTEP-CS-10-24, University of Texas at El Paso, 2010.

Non-contrastive durational patterns in two quantity languages

Kari Suomi[a], Einar Meister[b] & Riikka Ylitalo[a]
[a]Phonetics, Faculty of Humanities, Oulu University
[b]Laboratory of Phonetics and Speech Technology, Tallinn University of Technology

Abstract

Experimental findings of a durational comparison of selected disyllabic word structures in the closely related fully-fledged quantity languages Estonian and Finnish are reported, findings that are not directly related to the durational realisation of the quantity contrasts. It was observed that the word-initial consonant, which is outside the quantity system in both languages, behaves differently in Estonian and Finnish. Despite the many differences in the quantity systems of the two languages, it was observed that the grand mean durations of phonetic segments other than the initial consonant (namely V_1, C_2 and V_2), pooled across the different quantities, were usually the same in the two languages. With respect to accentual lengthening, the two languages behaved identically in some respects and differently in some other respects, in a way that is predicted by a model of speech timing based on durational findings in English.

Introduction

Estonian and Finnish are closely related quantity languages. But the quantity systems of the two languages are in many respects different. Without presenting detailed arguments, we assume a syntagmatic interpretation of quantity in both languages (that is, we do not assume that qualitatively similar segments in different quantities are separate phonemes). In Finnish this means that contrastively long segments are interpreted as sequences of identical phonemes, and that diphthongs are interpreted as sequences of two different vowel phonemes. This interpretation is also reflected in the orthography, e.g. *tuli* [t̪uli] /t̪uli/ 'fire', *tuuli* [t̪uːli] /t̪uuli/ 'wind', *tulli* [t̪ulːi] /t̪ulli/ 'customs', *tuoli* [t̪uoli] /t̪uoli/ 'chair'. The quantity opposition is binary, and it is phonologically very clearly segmental. The vowel quantity contrast obtains everywhere in the word, irrespective of word stress (which invariably falls on the initial syllable). The consonant quantity contrast also obtains irrespective of stress, but it is not possible word initially and finally, and not in all consonant clusters. For more details see Suomi, Toivanen & Ylitalo (2008).

The Estonian quantity system is more complicated, and numerous phonological interpretations have been offered, see e.g. the collection of papers in Lehiste & Ross (1997). Phonetically, there is a ternary opposition associated with the primarily stressed syllable, realised by durational relationships in the stressed syllable and in the following unstressed syllable; in recent loanwords, primary stress may fall on a non-initial syllable. On the basis of this disyllabic sequence, disyllabic and longer words have one of three quantities, namely Q1, Q2 or Q3, traditionally called short, long and overlong. The duration of the vowel in the unstressed syllable following the stressed syllable is partly predictable: it is longer in Q1 words than in Q2 and Q3 words. In contrast to Finnish, a vowel quantity opposition is not possible in the syllable following the stressed syllable. Moreover, there are also tonal cues that distinguish among the three quantities, see e.g. Lippus, Pajusalu & Allik (2009) and the references therein. Perceptual experiments have shown that native speakers of Estonian cannot distinguish between Q2 and Q3 on the basis of the stressed syllable alone: information on the second syllable is also necessary to perceptually distinguish between Q2 and Q3 (Eek & Meister, 1997; 2003). For such reasons, we endorse the nowadays dominant view that the proper scope of interpreting the Estonian ternary quantity contrast is the disyllabic sequence consisting of the stressed syllable and the next, unstressed

one. In unstressed syllables there is no vowel quantity contrast, whereas a binary consonant contrast is possible, also in the word-final position. Monosyllabic content words are always in Q3. For more details see e.g. Lehiste (1997), Meister & Meister (2011).

In contrast to Finnish orthography, Estonian orthography does not always indicate the quantity oppositions. It does show the contrasts for plosive consonants as in *lugu* (Q1) 'story', *luku* (Q2) 'lock, gen. sg.', *lukku* (Q3) 'lock, part. sg.'; notice that in Estonian orthography, in fully native words, the letters <bdg> do not represent voiced plosives but voiceless plosives in Q1. But otherwise the opposition between Q2 and Q3 is not indicated, e.g. *lina* (Q1) 'linen', *linna* (Q2) 'town, gen sg.', *linna* (Q3) 'town, part.sg'.

In this paper we report durational results of an experiment in which selected word structures of the two languages were compared. However, in this paper we do not report on the durational realisation of the contrastive quantities in the two languages. Instead, we report on durational differences and similarities that are not directly related to the quantity oppositions. A full report of the results will hopefully appear elsewhere (Suomi, Meister & Ylitalo, submitted).

Methods

We looked at Estonian and Finnish disyllabic words consisting of two consecutive consonant-vowel sequences, words in which either the stressed-syllable vowel or the following consonant are mainly responsible for the phonetic realisation of quantity, together with the duration of the second-syllable vowel duration. The starting point of target word selection was the existence of triplets of segmentally identical disyllabic words in the quantities Q1, Q2 and Q3 in the Estonian lexicon, representing the structures CVCV, CVVCV and CVVVCV on the one hand and the structures CVCV, CVCCV and CVCCCV, on the other, i.e. altogether six Estonian word structures. The decision to represent the Q3 words as CVVVCV and CVCCCV is merely typographical, it does not imply any claim that these words contain three consecutive identical phonemes. Examples of such word triplets are *kilu* (Q1), *kiilu* (Q2), *kiilu* (Q3) and *kade* (Q1), *kate* (Q2), *katte* (Q3). Only such Estonian word triplets were used for which phonetically sufficiently close Finnish CVCV, CVVCV word pairs and CVCV, CVCCV word pairs exist. For the Estonian word triplets just mentioned the Finnish target words were *kela*, *kiila* and *katu*, *katto*, respectively Altogether there were 42 Estonian and 28 Finnish target words. In both languages, the target words occurred in meaningful carrier sentences, in three degrees of prominence, namely unaccented, thematically accented and contrastively accented.

Nine volunteer female speakers aged 20-50 years were recorded in both languages, the Estonian speakers in Tallinn and the Finnish speakers in Oulu, using high quality recording equipment and highly similar recording procedures. For each target word, the durations of all constituent segments (C_1, V_1, C_2 and V_2) were measured. Notice that the symbols C_1, V_1, C_2 and V_2 here represent the four phonetic segments of the target words in both languages, irrespective of their phonemic composition. The absolute segment durations were converted to proportional durations (durations of segments as proportions of total word durations, computed separately for each word in each degree of prominence). Proportional durations effectively normalise for any differences in mean speaking rate.

Results

In the grand mean durations of the phonetic segments C_1, V_1, C_2 and V_2 across the six Estonian and four Finnish word structures, there was no difference between the two languages in the mean durations of C_2 and V_2 in any of the three degrees of prominence [statistically it was usually the case that $F < 1$, but in one comparison $F(1, 88) = 2.24$, n.s.]. As concerns the grand mean duration of V_1, there was a difference in the unaccented versions: the mean proportional duration of V_1 was 30% for the Estonian speakers and 25% for the Finnish speakers [$F(1, 88) = 5.24$, $p < 0.05$]. But in the thematically accented versions there was no difference [$F(1, 88) = 2.30$, n.s.], nor in the contrastively accented versions [$F < 1$]. With a single exception, then, the mean proportional durations of the segments V_1, C_2 and V_2 were the same in Estonian and Finnish.

But C_1 behaved very differently in the two languages. Notice that in both languages, C_1 is outside the quantity system as there is no quantity opposition in word-initial consonants. Table 1 shows the proportional mean durations of C_1 in both languages in the three degrees of prominence; "***" means that $p < 0.001$.

Table 1. Mean proportional C_1 durations in the three degrees of prominence.

	una	acc	con	
Estonian	20.1%	20.9%	22.4%	***
Finnish	25.9%	25.4%	25.4%	n.s.
	***	***	***	

C_1 behaved differently in Estonian and Finnish in three ways. Firstly, as can be seen in Table 1, its proportional duration was always smaller in Estonian than in Finnish. Secondly, in Estonian the proportional duration of C_1 varied very minutely yet systematically as a function of degree of prominence; post-hoc tests indicated that the proportional duration of C_1 was different in each degree of prominence (p = 0.001 or smaller). But in Finnish, degree of prominence had no effect. Thirdly, in accentual lengthening (the lengthening observed in the contrastively accented words relative to the mean of the other two degrees of prominence) there was a difference between the languages in the extent of proportional lengthening of C_1 [$F(1, 88)$ = 7.06, $p < 0.01$]: in Estonian the mean was 58%, in Finnish 41%. What this means is that the proportional duration of C_1 did not vary as a function of prominence in Finnish because, in this language, C_1 was lengthened statistically as much as the other segments on average. But in Estonian, proportional accentual lengthening of C_1 was larger than that of the other segments because C_1 was lengthened more than the other segments on average. That is, contrastive accent increased the proportional duration of C_1 in Estonian, but not in Finnish. It has been shown in Estonian that C_1 duration act as a perceptual cue to local speaking rate (Eek & Meister, 2003; Meister & Meister, 2011).

Except for segment position C_2, there were differences between the two languages in how extensive proportional accentual lengthening was, as shown in Table 2.

Table 2. Proportional accentual lengthening according to segment position. "=" means that the cross-linguistic difference failed to reach statistical significance.

	C_1	V_1	C_2	V_2
Estonian	58	34	29 =	46
Finnish	41	59	27 =	28

Estonian speakers thus exhibited a larger proportional accentual lengthening than the Finnish speakers at segment positions C_1 and V_2, and a smaller proportional lengthening than the Finnish speakers at segmental position V_1. We wish to argue that these differences are connected to, and explainable by, differences in the respective quantity systems. Speakers of Estonian may be less reluctant than speakers of Finnish to lengthen V_1 extensively because such lengthening might interfere with durational signalling of the very complex quantity contrasts. In Estonian, the quantity relations among V_1 and C_2 yield nine (2^3) different possibilities, whereas in Finnish there are only four (2^2) possibilities (VC, VVC, VCC, VVCC), and the V - VV contrast is supported in the initial syllable by a very robust durational difference.

On the other hand, speakers of especially Northern Finnish must be careful not to lengthen V_2 too much when it constitutes a single vowel and the initial syllable is light (i.e., when V_2 constitutes the word's second mora, M_2), lest the second-syllable single V_2 be confused with a double vowel. In Northern Finnish, the difference between second-syllable single (V) and double vowels (VV) after a light initial syllable is rather precarious. For example, in Suomi & Ylitalo (2004; Table 2, p. 42), in which segment identities were fully controlled, the VV/V durational ratio measured in CV.CVV.CV and CV.CV.CV nonsense items was 1.5. And Nakai, Kunnari, Turk, Suomi & Ylitalo (2009) observed that utterance-final lengthening of V_2 *qua* M_2 was effectively blocked, obviously in order to prevent confusion with a double vowel in the same position. But speakers of Estonian may feel free to lengthen V_2 because there is no quantity opposition in the second unstressed syllable.

In many other respects the two languages turned out to be highly similar, if not identical. Firstly, in both languages accentual lengthening affected only contrastively accented words, but not non-contrastively accented words. In this respect both languages differ from e.g. Swedish, in which accentual lengthening is not limited to contrastively accented words. Secondly, the two languages did not differ from each other in terms of the amount of total accentual lengthening. Thirdly, in both languages the word structures did not differ among themselves with respect to the amount of accentual lengthening.

Discussion

In the domain-and-locus model of speech timing proposed by White (2002), domain refers to the prosodic constituent within which a timing process is operative, and locus refers to the particular segments that are affected by the process; processes may be distinguished by their distinct loci. According to the model, speech timing consists of localised effects: segments are produced with durations simply determined by intrinsic factors, modulated according to speech rate, until a locus of some timing process is reached. At this point, some extra, constant amount of duration is allocated to the locus, with no regard paid to the segmental composition of the locus. But this constant amount of lengthening is distributed within the locus according to the structure and the segmental composition of the locus. As a result of both of these factors, the lengthening will not be evenly distributed within a locus. In accentual lengthening the domain is the word, and the locus, in the present disyllabic word structures, clearly includes all segments in the word.

The present Estonian and Finnish findings on accentual lengthening are in perfect agreement with the predictions of White's model: firstly, as already mentioned, in both languages the word structures did not differ among themselves with respect to the amount of accentual lengthening. That is, a constant extra amount of duration was allocated to the locus, in both languages, with no regard paid to the segmental composition of the locus. Secondly, the lengthening was not evenly distributed within the loci in either language, and it was differently distributed in the two languages: C_1 and V_2 were lengthened more in Estonian than in Finnish, while V_1 was lengthened more in Finnish than in Estonian. These differences seem to be explainable by differences in the two quantity systems: the greater complexity, in Estonian, of the possible quantity relations between the stressed-syllable vowel and the following consonant, and the greater freedom, in Estonian, to lengthen V_2 because there is no vowel quantity contrast in the second syllable. In brief, in both languages accentual lengthening was unevenly distributed within the locus, according to language-specific rules that seem to be motivated on structural grounds relatable to quantity.

The agreements of the present findings in two different quantity languages with the predictions of White's (2002) model of speech timing, based on findings in English, suggest two mutually interrelated things, but let us spell them out. Firstly, the agreements indicate that White's model captures something cross-linguistically important in speech timing. Secondly, the agreements suggest that, in some respects at least, fully-fledged quantity languages and non-quantity languages (or only marginally quantity languages) like English (which does have a systematic durational difference among its vowels, but with accompanying large qualitative differences) are very similar if not identical.

References

Eek, A & Meister, E (1997). Simple perception experiments on Estonian word prosody: foot structure vs. segmental quantity. In I. Lehiste & J. Ross (Eds.), *Estonian Prosody: Papers from a Symposium* (pp. 71-99).

Eek, A & Meister, E (2003). Foneetilisi katseid ja arutlusi kvantiteedi alalt (I): häälikukestusi muutvad kontekstid ja välde. *Keel ja Kirjandus*, 46(11), 815 - 837.

Lehiste, I (1997). Search for phonetic correlates in Estonian prosody. In I. Lehiste & J. Ross, eds, *Estonian Prosody: Papers from a Symposium*. Estonia: 11-35.

Lehiste, I & Ross, J (1997). *Estonian Prosody: Papers from a Symposium*. Tallinn: Institute of Estonian Language.

Lippus, P, Pajusalu, K & Allik, J (2009). The tonal component of Estonian quantity in native and non-native perception. *Journal of Phonetics, 37*, 388-396.

Meister, L & Meister, E (2011). Perception of the short vs. long phonological category in Estonian by native and non-native listeners. *Journal of Phonetics, 39*, 212-224.

Nakai, S, Kunnari, S, Turk, A, Suomi, K & Ylitalo, R (2009). Utterance-final lengthening and quantity in Northern Finnish. *Journal of Phonetics*, *37*, 29-45.

Suomi K & Ylitalo R (2004). On durational correlates of word stress in Finnish. *Journal of Phonetics*, 32: 35-63.

Suomi, K, Toivanen, J & Ylitalo, R (2008). *Finnish sound structure*. Studia Humaniora Ouluensia 9. http://herkules.oulu.fi/isbn9789514289842.

Suomi K, Meister E & Ylitalo R (submitted). Durational patterns in Estonian and Northern Finnish.

White, L. (2002). English speech timing: a domain and locus approach. Ph.D. Dissertation, University of Edinburgh. http://www.cstr.ed.ac.uk./projects/eustace/dissertation.html.

An investigation of intra-turn pauses in spontaneous speech

Kristina Lundholm Fors

Department of Philosophy, Linguistics and Theory of Science and Graduate School of Language Technology, University of Gothenburg

Abstract

In this study, pauses within speakers' turns are described and analysed. Tentative results show that different pauses within a speaker's turn might differ in length. Pause length variations over time in dialogues were investigated, and in 5 out of 6 dialogues, a statistically significant correlation was found between the speakers' variations in pause length.

Introduction

Pauses are an essential part of human speech, and they fill many different functions. We need pauses to, for example, breath, plan what we are going to say and negotiate turntaking. In their now classic article, Sacks et al. (1974) categorized pauses in speech into *pauses*, *gaps* and *lapses*. They are defined as follows: a *pause* is a silence that occurs inside a speaker's turn. This includes the silence at a transition relevance place (TRP), when a speaker has been nominated but has not begun to speak. It also includes the silence at a TRP, when a speaker has stopped, but then continues to speak after the TRP. A *gap* is the silence that occurs at a TRP when the first speaker has not nominated another speaker, but another speaker self-nominates and there is a turn change. A *lapse* is the silence at a TRP, when the first speaker has stopped speaking, has not nominated a new speaker, and does not continue speaking. No other speaker takes the turn. A lapse is in part defined by the perceived length: thus, a lapse should be perceived as longer than a gap and as a discontinuity in the flow of conversation.

Heldner and Edlund (2010) provide an excellent review of pause research to date. It is evident in their review that researchers have used different methods and definitions, which means that results are difficult to compare. Even the definition of a pause differs between studies: some argue that a pause is that which a listener perceives as a pause, whereas others base their pause identification on the acoustic signal. This will undoubtedly lead to discrepancies in results, as it has been shown that perceived pauses do not necessarily coincide with pauses identified acoustically (Zellner, 1994).

The majority of pauses and gaps are shorter than 1000ms, and the most common gap length is 200ms (Heldner and Edlund, 2010), but there is a substantial amount of intra- and interspeaker variability. Pauses, that is intervals where the speaker is silent within her turn, can occur at different places in the turn, as described above. It might therefore be useful to divide pauses into sub-categories, as the context of the pause might affect its length; when a speaker pauses within her turn but not at a TRP, it is already clear that the speaker will continue after the pause. This can be compared to when a speaker pauses at a possible TRP: even when turn change does not take place, the speakers still have to negotiate who is going to speak after the pause. It is not unlikely that this will lead to the pauses at possible TRP:s being longer than the pauses that do not occur at a possible TRP. The pauses that occur at the beginning of a speaker's turn when the speaker has been nominated by a previous speaker should in that case also be shorter than the pauses at possible TRP:s.

Pauses that occur within a speaker's at a possible TRP will hereafter be referred to as pauses between syntactic units, whereas pauses that occur within a speaker's turn but not at a possible TRP will be referred to as pauses within syntactic units. A pause that exists at the beginning of a speaker's turn when she has been nominated but has not yet begun to speak can be referred to as initial pause. However, that pause type will not be further discussed in this paper. In the present study the pauses within and between syntactic units will be investigated and compared. If there is evidence of a difference in length between these pause types, that information could be used, in conjunction with other factors, in end-of-utterance detection in dialogue systems.

Edlund et al. (2009) have proposed a method

to investigate pause length variation in dialogues over time. Their method might make it possible to capture the dynamics of pause variation, and to visualize how the speakers influence each other. In this study I will apply the method proposed by Edlund et al. to pauses within turns.

Method and material

Five persons, all female, were recorded while speaking in pairs. Altogether, 6 dialogues were recorded, each lasting approximately 10 minutes. The subjects received a question to discuss but were informed that they were allowed to stray from the subject. The subjects will be referred to as a, b, c, d & e and the dialogues as D1-D6. The speakers were paired as follows:

- dialogue 1 (D1): speakers a + e
- dialogue 2 (D2): speakers d + b
- dialogue 3 (D3): speakers a + c
- dialogue 4 (D4): speakers d + e
- dialogue 5 (D5): speakers a + b
- dialogue 6 (D6): speakers d + c

The dialogues were transcribed ortographically in Praat, and pauses and gaps were identified manually based on the acoustic signal. No cut-off time for pause length was set. This is often done to exclude occlusion intervals, but they were instead excluded manually. The pauses were categorized into the three different categories described in the previous section: pauses within syntactic units, pauses between syntactic units and initial pauses.This coding schedule has not yet been tested for inter-rater-reliability tested, but this is planned in relation with future studies.

When pause length variation over time was analyzed, a moving, Gaussian-shaped window of 9 data points was used. The Gaussian shape of the window gives more weight to the central values in the window, and provides a smoother curve. Using interpolated curves, average pause lengths were calculated for both speakers in each dialogue; the pause lengths were measured at the data points of one speaker.

Results

Comparison between pause types

For each speaker, all pauses within and between syntactic units were identified. Pauses for each speaker were pooled across dialogues. The pause lengths were logarithmized (as pauses are not normally distributed) and mean values were calculated. This data is presented in table 1.

Table 1: Mean pause lengths and number of pauses per speaker

Speaker	Within units	Between units
a	0.41 (124)	0.65 (150)
b	0.58 (66)	0.70 (63)
c	0.49 (126)	0.38 (66)
d	0.47 (209)	0.47 (100)
e	0.40 (51)	0.55 (85)

For speakers a, b and e, pauses between syntactic units are on average longer than pauses within syntactic units. For speaker c, the opposite is true: the mean of pauses within syntactic units is 0.49 seconds, compared to 0.38 which is the mean of the pauses between syntactic units. The mean pause lengths of speaker d were the same both within and between syntactic units.

Pause length variation over time

Figures 1-6 show how the speakers' median pause lengths vary over time in each dialogue. The curves do not always cover the same amount of time: for example it can be seen in figure 2 that speaker d's curve continues for some time after speaker b's curve has ended. This is caused by the fact that speaker b did not make any intra-turn pauses after a certain time in the dialogue.

The pause lengths in each dialogue were correlated between speakers, using Pearson's product-moment correlation coefficient (Pearson's r). It is important to note that it is not the actual pause durations that are compared, but how they vary over the course of the dialogue. If there is a positive correlation, pause lengths tend to be higher and lower than average respectively at the same time for both speakers. On the other hand, if there is a negative correlation, pause lengths tend to longer than average for one speaker when they are shorter than average in the other speaker. This means that there can be a negative or positive correlation even if average pause times differ. An overview of the correlations is given in table 2.

In 1 the pause length variations of speaker a and e are presented. A significant positive correlation (p=0.01) between the speakers' pause length variations was found in this dialogue.

Figure 2 shows speakers b and d. In this dialogue there was a significant negative correlation (p=0.01) between the speakers' pause length vari-

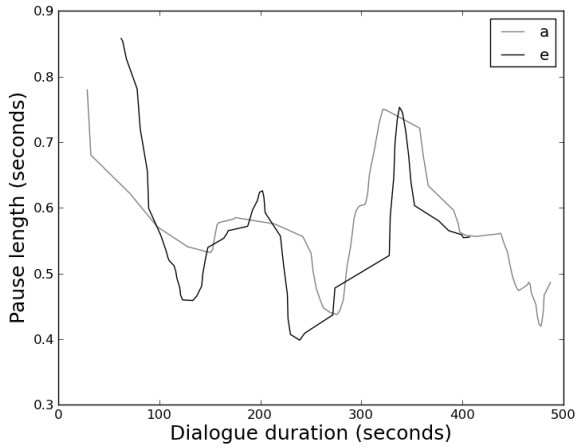


Figure 1: Dialogue 1

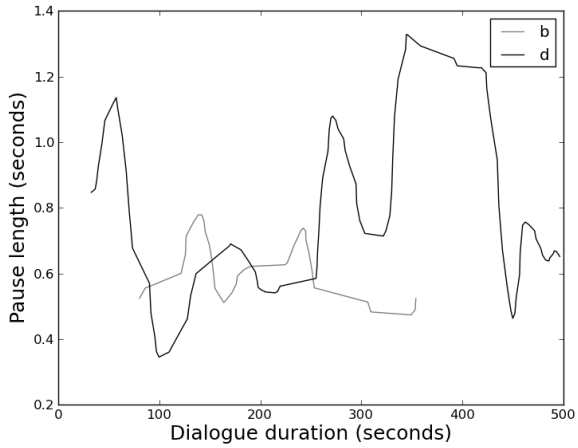


Figure 2: Dialogue 2

ations. This is the only negative correlation between pause length variations found in the study.

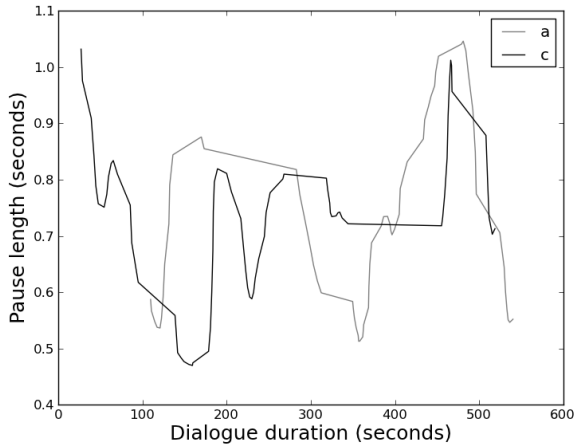


Figure 3: Dialogue 3

Speakers a and c took part in dialogue 3, which is presented in figure 3. A significant positive correlation (p=0.05) between the speakers' pause length variations was found.

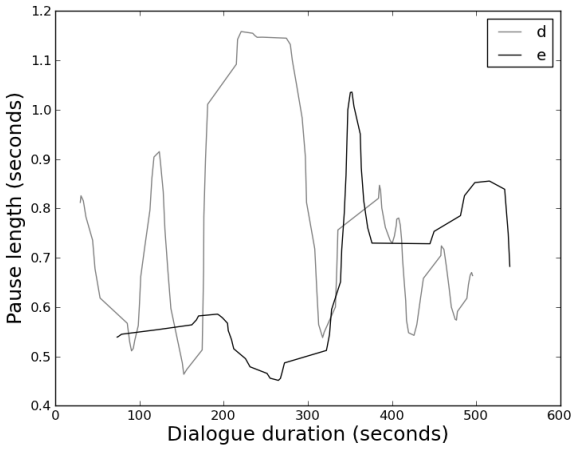


Figure 4: Dialogue 4

The pause length variations of speakers d and e is visible in figure 4. In this dialogue there was a significant positive correlation (p=0.01) between the variations in speakers' pause lengths.

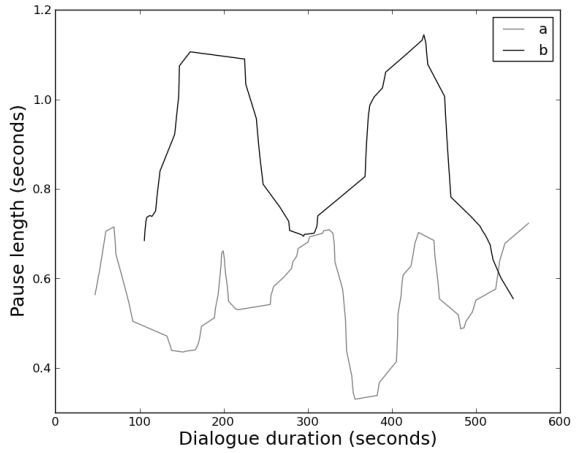


Figure 5: Dialogue 5

In dialogue 5, which can be seen in figure 5 speakers a and b participated. No significant correlation between their pause length variations in was found.

Finally, the pause length variations of speakers d and c are shown in 6. In this dialogue, dialogue 6, a positive correlation (p=0.05) was found.

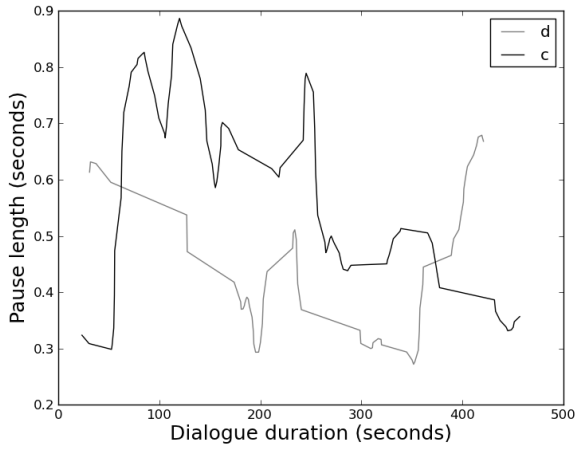


Figure 6: Dialogue 6

Table 2: Correlations between pause length variations in D1-D6

	Pearson's r
D1	.621**
D2	-.635**
D3	.327*
D4	.391**
D5	-.175 p=.132
D6	.405**

*Correlation is significant at the .05 level
**Correlation is significant at the .01 level

Summary and discussion

In the introduction, three subcategories of pauses were proposed, of which two are investigated in this study. The hypothesis is that pauses that occur between syntactic units (at a possible TRP) should be longer than pauses that occur within syntactic units. The reason for this would be that pauses that occur at a possible TRP include turn taking negotiation. Out of the five speakers investigated in this study, three present longer pauses on average between syntactic units compared than within syntactic units, which is in line with the hypothesis. However, one speaker showed no such difference, and one speaker made longer pauses within syntactic units. This suggests that the subcategorization of pauses may be useful, but more data is needed to back up this hypothesis.

When looking at the data presented in table 1, it is clear that average pause times vary markedly between speakers; for example, speaker e's mean value for longer pauses is shorter than speaker b's mean value for shorter pauses. This means that when using pause length data in for example end-of-utterance detection, a baseline for the speaker's pause lengths should be established, and subsequent pauses compared to this, rather than compared to a 'one-size-fits-all' general pause length value.

The method proposed by Edlund et al. (2009) was used in this study, but with some alterations. Instead of a rectangular moving window a Gaussian-shaped window was used, and fewer data points (9 in this study compared to 20). A Gaussian-shaped window did provide a efficient smoothing of the curve, but it is possible that a window that places more weight on the most recent pause would yield even better results. Regarding the number of data points, future research should investigate whether there is an optimal number of data points. That number could quite possibly depend on the length of the dialogue and the number of pauses identified.

Another distinct difference is that Edlund et al. used automatic pause identification, whereas in this study pauses where identified manually. Automatic pause identification has the advantage that a large amount of data can be used, while manual pause identification does not necessitate excluding pauses shorter than a certain length to not include occlusion intervals in stop consonants. In the future, the results of automatic and manual identification of pauses should be compared to see if, and then how, they differ.

When comparing pause length variations for speakers in dialogues, significant correlations were found in five out of six dialogues. This provides further evidence that the method outlined by Edlund et al. (2009) is an effective way of capturing how speakers are influenced by each other when it comes to pause length variation. With more data and more in-depth analysis, the results yielded by this method could be used to model pause behavior in dialogue systems.

References

Edlund J, Heldner M and Hirschberg J (2009). Pause and gap length in face-to-face interaction. In *Proc. of Interspeech*, 2779–2782.

Heldner M and Edlund J (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Sacks H, Schegloff E and Jefferson G (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Zellner B (1994). Pauses and the temporal structure of speech. In E Keller, ed., *Fundamentals of speech synthesis and speech recognition*, 41–62. Chichester: John Wiley.

Spoken Language Identification using Frame Based Entropy Measures

Giampiero Salvi and Samer Al Moubayed

KTH, School of Computer Science and Communication, Dept. of Speech, Music and Hearing, Stockholm, Sweden
{giampi,sameram}@kth.se

Abstract

This paper presents a real-time method for Spoken Language Identification based on the entropy of the posterior probabilities of language specific phoneme recognisers. Entropy based discriminant functions computed on short speech segments are used to compare the model fit to a specific set of observations and language identification is performed as a model selection task. The experiments, performed on a closed set of four Germanic languages on the SpeechDat telephone speech recordings, give 95% accuracy of the method for 10 seconds long speech utterances and 99% accuracy for 20 seconds long utterances.

Introduction

Spoken Language Identification (LID) has long been a topic of interest in speech technology. In a multilingual world, speech applications are in many cases required to automatically select the appropriate language when a new user interaction is initiated. Recently, the multilingual aspects of speech applications have gained focus, e.g., in speech translation and international call and information services. Furthermore, large multilingual speech databases are now available, allowing for more advanced LID systems to be developed and properly tested.

Some of the methods proposed to solve the language identification task make use of low level acoustic features. In Lamel and Gauvain (1994), e.g., the acoustic likelihoods of phoneme recognisers for each language are used.

Other methods are based on modeling longer context in speech by means of language models (LM). These may make use of a single set of phonetic acoustic models followed by language specific models (PRLM, e.g., Glembek et al. (2008)), or employ parallel phoneme recognisers for each language (PPRLM, e.g., Zissman (1996); Zhu and Adda-Decker (2006)). The language models may consist of large vocabulary recognisers, or, more commonly, of N-grams at the phonetic level as in Zissman (1996).

This paper proposes an alternative way of using the output of the language specific acoustic models. The frame-based class entropy of the posterior probabilities for each phoneme is used as a measure of uncertainty of each language-specific phoneme recogniser. This score is local in time and does not use any phonotactic or higher level linguistic information, making it more suitable to generalise to applications where the utterances may differ considerably from the data used during training. Another advantage over PPRLM methods, is that computing the entropy locally does not need a Viterbi like decoder, thus simplifying the LID task and allowing for faster detection of a language change. This is particularly interesting in applications such as lip-synchronisation systems Salvi et al. (2009), where the phonotactic information is not available or of any interest.

We test the method on a close set of four Germanic languages (Swedish, German, Flemish and English) and using recurrent neural networks as phoneme posterior probability estimators, and compare it to a PPRLM baseline using the same PPRs.

Method

The problem of language identification can be seen as a standard classification problem. Given an observation o, typically an acoustic feature vector in a d-dimensional space, we want to infer the true identity l^* of the language out of a set of possible languages $\mathcal{L}$. The solution to the problem can be formulated in general terms by defining a set of real valued discriminant functions g_l, one for each class (language), on the observation space $\mathbb{R}^d$. Depending on the definition of the g_l,

the classification problem is then casted into a minimisation or maximisation problem, e.g.:

$$\hat{l} = \arg\min_{l \in \mathcal{L}} g_l(o); \qquad g_l : \mathbb{R}^d \rightarrow \mathbb{R} \qquad (1)$$

Our method is based on the assumption that, for each language l, we can compute estimates of the posterior probabilities $p(c_i|o, M_l)$ of each phoneme $c_i \in \mathcal{C}_l$, given an observation o and a statistical model M_l. Typically, M_l is a phoneme recogniser used in the application for which we want to determine the identity of the language. The feature extraction procedure used to extract the observations o from the speech samples is assumed to be identical in each language, whereas the model M_l and the set of possible phonemes $\mathcal{C}_l$ vary with l (Parallel Phoneme Recognisers).

We propose to use the entropy of the posterior probabilities $p(c_i|o, M_l)$ of the phonemic classes as a measure of mismatch between the model M_l and the observation o. The higher the entropy, the higher the uncertainty of the model in describing the observation o. Comparing entropy measures obtained with different models M_l, we can select the model that best fits the observation o and therefore the most likely language o belongs to.

Formally, we define the entropy of the current observation given the model M_l as:

$$H(o, M_l) = -\sum_{i=1}^{N_l} p(c_i|o, M_l) \log_{N_l} p(c_i|o, M_l) \qquad (2)$$

Where N_l is the number of phonemes for language l. We choose to compute the logarithm in basis N_l to ensure that $H(o, M_l)$ ranges from 0 to 1 regardless of the size of the phonemic inventory of language l, thus simplifying the comparison between different models M_l.

The distribution of $H(o, M_l)$ varies, not only based on the mismatch between the input language and model language, but also depending on the performance of each model M_l. In case of matching input language and model language ($l = l^*$), more discriminative models will give more picky distributions of the posterior probabilities for each input and, therefore, lower entropy. Another effect is the mismatch between training and test data. Each model M_l is likely to give lower values for the entropy on the training data compared to data that is unseen during training. In order to compare the entropy across language models, it is necessary to normalise for the effect of intrinsic performance of each model. This can be done by using a global estimate of the mean and standard deviation of $H(o, M_l)$:

$$\bar{H}_G(o, M_l) = \frac{H(o, M_l) - \hat{\mu}_l}{\hat{\sigma}_l} \qquad (3)$$

Where $\hat{\mu}_l$ and $\hat{\sigma}_l$ should be computed on data that is unseen during training in order to normalise for the mismatch between training and test set as well as for the intrinsic performance of the model.

Another source of variation for the entropy is the phonetic content of the acoustic observation. This is because the classifiers M_l, used as posterior probability estimators, have an accuracy that is strongly dependent on the phonemic class. A more efficient normalisation method would, therefore, be phoneme dependent, i.e., by estimating the mean μ_{il} and standard deviation σ_{il} for each phoneme i in language l. A possible way to estimate these parameters is by referring to the true phoneme labels. A limitation of this method is that the identity of the true phoneme is not known during testing, therefore questioning the validity of this normalisation. Another possibility is to use the *winner phoneme* obtained by maximising the posterior probability $p(c_j|o, M_l)$ in the estimation of μ_{il} and σ_{il}. In this case the statistics are computed on the set of observations such as $\mathcal{O}_i = \{o : \arg\max_j p(c_j|o, M_l) = i\}$.

When computing the normalised entropy for a new observation, we perform a weighted average of the phoneme normalised entropy, using the posterior probability of each phoneme as weight.

$$\bar{H}_P(o, M_l) = \sum_{i=1}^{N_l} p(c_i|o, M_l) \frac{H(o, M_l) - \hat{\mu}_{il}}{\hat{\sigma}_{il}} \qquad (4)$$

This constitutes our best estimate of the belief a certain phoneme was uttered at a certain moment in time.

Summarising, the discriminant functions in Eq. 1 are defined as average over time of the entropy measures defined above:

$$g_l(o) = \frac{1}{T} \sum_{t=1}^{T} \bar{H}(o_t, M_l) \qquad (5)$$

The length T of this average can be varied and is one of the experimental factors. We tested four methods: the first is based on the raw entropy $H(o_t, M_l)$ defined in Eq. 2 This method will be referred to as RAWE. The second, uses the globally normalised entropy $\bar{H}_G(o, M_l)$ (Eq. 3) and will be referred to as GNE. The last two methods use phoneme based normalisations defined in Eq. 4. In the first case, named CPNE, the parameters in the formula are estimated by using the correct phoneme for each frame as given by the phonemic annotations. In the second case the parameters are estimated according to the winner phoneme (WPNE).

Experiments

Data

The experiments in this paper are based on the SpeechDat databases for Swedish, German, Flemish and English Elenius (2000). The databases contain recordings over the fixed telephone line sampled at 8 kHz. The content ranges from read sequences of digits, and phonetically rich sentences to spontaneously uttered names.

The test sets for each language included up to about 30 minutes of speech.

Baseline model

In order to compare our method with a standard implementation of PPRLM, we tested the same phoneme recognisers in combination with N-grams models as in Zissman (1996). The language dependent N-gram models were estimated by decoding the training data for each language with all the available phoneme recognisers. A total of 16 bi-grams were estimated for each combination of training language and language-specific phoneme recogniser. During testing, the log-likelihoods obtained with different phoneme recognisers and language models were averaged in order to select the best language for each speech chunk.

Experimental settings and Evaluation

The phoneme recognisers used in this paper to estimate the posterior probabilities are based on recurrent neural networks (RNNs) Salvi (2006). The input to the networks are Mel Frequency Cepstral Coefficients (MFCCs) extracted on 10 ms spaced frames of speech samples. The networks are trained using Back Propagation through time Werbos (1990) with a cross entropy error measure Bourlard and Morgan (1993). This ensures an approximately linear relation between the output activities of the RNN and the posterior probabilities of each phonetic class, given the input observation Ström (1992). The transcriptions used as targets to train the neural networks were obtained by force alignment using the orthographic transcriptions and the lexica in the SpeechDat databases.

Performance is computed for the four methods described in Section and for the baseline system on a close set of the four target languages. Results are computed as in Caseiro and Trancoso (1998) as Identification Rate, i.e., as percentage of the test chunks that are correctly classified.

Table 1: Identification rates (%) for SpeechDat tests

	window length (sec)				
Method	1	2	5	10	20
RAWE	45.0	41.0	42.5	39.0	43.0
GNE	47.5	47.2	51.5	54.2	55.5
CPNE	53.0	55.5	64.7	66.2	69.5
WPNE	70.5	78.5	89.5	94.5	99.0
baseline	59.3	69.7	80.0	86.9	90.7

Table 2: Language dependent identification rates (%) on the SpeechDat database for the WPNE condition

	window length (sec)				
Language	1	2	5	10	20
Swedish	68	80	88	92	98
German	52	64	88	92	100
Flemish	76	82	88	98	98
English	86	88	94	96	100

The test set consists, for each of the four target languages of 100 non-overlapping speech chunks for a total of 400 speech chunks. The length of the speech chunks is varied from 1 to 20 seconds. The silence segments were automatically removed in the entropy estimation. This is because silence is irrelevant to the problem and might introduce artifacts that depend on the characteristics of the channel or the performance of the phoneme recogniser, rather than on the identity of the test language.

Results

Table 1 summarises the results obtained on the SpeechDat data with the four methods defined in Section and for the baseline system. Results are shown in terms of average identification rate (IR) computed over the four test languages. The test are performed with speech chunks of 1, 2, 5, 10 and 20 seconds.

All methods perform largely above chance level (25%). The method based on the raw entropy RAWE has the lowest IR. Normalising the entropy globally (GNE) brings an improvement on the IR and so does the phoneme dependent normalisation with normalisation parameters based on the correct phoneme identity (CPNE). Finally, the WPNE method outperforms consistently the PPRLM based baseline system with a relative improvement of about 19% for the 1 second test and 9% for the 20 second test. The same information is plotted in Figure 1 for varying length of the input speech chunks (with 1 second steps). The results for the WPNE method are also detailed for

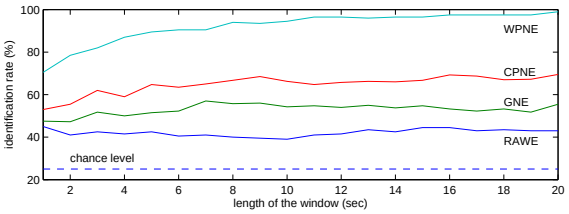


Figure 1: Identification rate versus length of the test speech chunks. The plot shows results for the four methods defined in Section , the PPRLM based baseline system and the chance level considering the four target languages.

each target language in Table 2. Although the performance varies for each language, this effect is mostly evident for short speech chunks, below 5 seconds.

Discussion and Conclusions

This paper describes a method for language identification based on the entropy of the posterior probabilities estimated by language specific phoneme recognisers. We showed that comparing the raw entropy of the model gives results above chance. We also showed that different methods of normalisation of the entropy give increasing identification rates.

The identification rate increases in general with the length of the speech chunks considered in the classification. The best method, based on phoneme dependent normalisation (WPNE) gives a 95% identification rate (IR) with 10 seconds speech chunks and 99% IR with 20 seconds speech chunks. This overperforms the baseline system based on bi-gram language models. The performance increase is larger for shorter speech chunks, making this method particularly appealing for applications where the language may be changing rapidly, or when the system should react in short time to language changes.

We experimented with Germanic languages in this paper because the phoneme recognisers are, at the moment, only available in those languages. Although this could be seen as a limitation to this study, we believe that showing that phonetically similar languages can be discriminated on the basis of acoustic properties alone, is a strong indication of the validity of our method. We therefore believe that the results will hold when other languages are added to the problem.

Future work will be dedicated to studying how this results are dependent on the intrinsic performance of the phoneme recognisers. Also we will investigate if the performance can be improved by feeding the entropy measures in a more complex classifier, e.g. based on Support Vector Machines.

Finally, we plan to adapt the model to open-set tests where we consider the possibility of rejecting an input language that does not correspond to any of the available recognisers.

Acknowledgements

This work was carried out at the Centre for Speech Technology, a competence centre at KTH. Author G.S. thanks the Swedish Research Council (Vetenskapsrådet grant 2009-4599).

References

Bourlard H and Morgan N (1993). Continuous speech recognition by connectionist statistical methods. *IEEE Trans Neural Netw*, 4(6):893–909.

Caseiro D and Trancoso I (1998). Spoken language identification using the SpeechDat corpus. In *Fifth International Conference on Spoken Language Processing*. ISCA.

Elenius K (2000). Experience from collecting two Swedish telephone speech databases. *International Journal of Speech Technology*, 3:119–127.

Glembek O, Matejka P, Burget L and Mikolov T (2008). Advances in phonotactic language recognition. In *Interspeech*, 743–746. Brisbane, Australia.

Lamel L and Gauvain J (1994). Language identification using phone-based acoustic likelihoods. In *IEEE ICASSP*, vol. 1.

Salvi G (2006). Dynamic behaviour of connectionist speech recognition with strong latency constraints. *Speech Communication*, 48(7):802–818.

Salvi G, Beskow J, Moubayed S A and Granström B (2009). SynFace — speech-driven facial animation for virtual speech-reading support. *EURASIP Journal on Audio, Speech, and Music Processing*.

Ström N (1992). Development of a recurrent time-delay neural net speech recognition system. *TMH-QPSR*, 1992(4):1–15.

Werbos P J (1990). Backpropagation through time: what it does and how to do it. *Proc of the IEEE*, 78(10):1550–1560.

Zhu D and Adda-Decker M (2006). Language identification using lattice-based phonotactic and syllabotactic approaches. In *Speaker and Language Recognition Workshop*, 1–4.

Zissman M A (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Tran Speech and Audio Proc*, 4(1):31–44.

Exploring phonetic realization in Danish by Transformation-Based Learning

Marcus Uneson, Ruben Schachtenhaufen***

Lund University, Copenhagen Business School***

Abstract

We align phonemic and semi-narrow phonetic transcriptions in the DanPASS corpus and extend the phonemic description with sound classes and with traditional phonetic features. From this representation, we induce rules for phonetic realization by Transformation-Based Learning (TBL). The rules thus learned are classified according to relevance and qualitatively evaluated.

Introduction

Language abounds with classification tasks – some we solve ourselves, some we hand over to machines. In the latter case, we may or may not be interested in what the machine actually learns. Stochastic classifiers such as HMMs and SVMs are useful for many purposes, but their target representation is usually inscrutable to humans. Rule learners, on the other hand, may or may not match stochastic classification performance, but what they learn may be interesting in itself – sometimes, it might even be the main point.

In the present paper, we explore one of those cases: the application of a well-known rule induction technique, Transformation-Based Learning (Brill, 1995), on phonetic string representations. The problem can be phrased thus: given a phonemic and a semi-narrow phonetic transcription of speech, can we extract transformation rules which will take the first to the second, or at least part of the way? If so, do these rules give us any new insights? Somewhat less abstractly, our aim is to automatically induce typical textbook rules for phonetic realization, from a transcribed, real-world corpus of spontaneous, connected speech. The language under study is Danish, where, arguably, the distance between these two representations is particularly noteworthy.

Background

Danish phonology

Grønnum (2005) analyzes Danish phonology into 11 vowel phonemes (/i e ɛ a y ø œ u o ɔ ə/) and 15 consonant phonemes (/m n p t k b d g f v s h l r j/), plus the prosodic elements *stød*, length, and stress. Briefly, most non-high vowels are realized more open before and/or after /r/, and some consonants are realized differently depending on syllable position: /p t k/ are aspirated in onset and unaspirated in coda; /d g v r/ are contoid in onset and vocoid or ∅ in coda.

The realization of /ə/ is quite complex. More often than not it is elided, leaving its syllabic trait and compensatory lengthening on adjacent sounds. The combination of consonant gradation in coda and a very fleeting /ə/ results in a highly unstable sound structure in current Danish.

In traditional descriptions, being based mainly on conservative, careful, read speech, Danish phonemes typically have one or two, rarely three, allophones, e.g. "/d/ > [ð] in coda, [d] elsewhere". In spontaneous speech phonemes have a much wider range of realization – for instance, in DanPASS (see below), /d/ is transcribed [d ð ɾ ɹ t z s], among others.

Transformation-based learning

Transformation-based learning (TBL) was proposed by Eric Brill (Brill, 1995). It is, in a one-sentence summary, a supervised machine learning method producing a compact, ordered, human-readable list of classification rules (or *transformations*), each chosen greedily from a set of candidates dynamically calculated from user-supplied patterns (the *templates*), so that it maximally reduces (a function of) the difference between the system's present idea of the classification (the *current corpus*) and a gold standard (the *truth*). One sentence is likely not enough; we refer to Brill (1995).

The task at hand reminds somewhat of letter-to-sound (LTS) conversion, to which TBL also has been applied (Bouma, 2000). Abstractly, both problems concern transforming one string representation of language into another.One major difference is that LTS aims at lexical pronunciation:

	der igen	er i	midten
phonemic:	deːʔr_iˈgɛn	ɛr_iːʔ	ˈmetən
phonetic:	daɪˈgɛn	ˈaɪ	ˈmedn̩

Figure 1: DanPASS phonemic and phonetic tiers for der igen er i midten *'that again is in the middle'*

it usually has a well-defined target. Phonetic realizations, by contrast, have several influencing factors but few truly functional dependencies. In this paper we will pay more attention to the rules themselves extracted than how close to the (partly arbitrary) target they will take us.

The present study

On transcriptions

Although historically much used, the method of taking transcriptions as point of departure for phonetic conclusions is not without its problems. Transcriptions imply a simplistic and much reduced 'beads-on-a-string' view on speech, often with weak support in data which have not been filtered through the perception of a native speaker. In the words of Grønnum (2009), "phonetic notation, specifically of the rather narrow kind, and prosodic labeling are both impressionistic exercises". For the purposes of this paper, however, we will accept this armchair view.

The DanPASS corpus

Our data is certainly not armchair; it was taken from the DanPASS corpus[1] (Danish Phonetically Annotated Spontaneous Speech) (Grønnum, 2009). In total, the corpus comprises about 10 hours (73kW) of annotated high-quality recordings of connected speech produced by 27 speakers, distributed among several tasks in non-scripted monologue and dialogue. DanPASS addresses no particular research need specifically, but is generally well suited for studying phenomena associated with connected, spontaneous speech. With the exception of a small fraction of non-spontaneous speech (elicited word lists), we used all of it.

The phonemic transcription in DanPASS is based on the analysis of Grønnum (2005) mentioned above. The annotations of DanPASS are available as Praat tiers. The ones of concern here are the the phonemic notation and and the semi-narrow phonetic notation. Figure 1 shows a small corpus sample for these.

[1] http://danpass.dk

Experimental setup

We employed the μ-TBL system (Lager, 1999), with the semi-narrow transcription tier taken as truth and the phonemic tier as the initial current corpus. TBL requires that the current corpus and the truth are containers of the same shape, which in the present case requires alignment of the transcriptions; for this task, we used the sound class alignment method proposed by List (2010).[2] Since our interest lies in rules which apply with few or no exceptions, all rules were required have a minimum accuracy of 0.95.

The problem encoding required more consideration. Rules should of course be conditioned on the immediate phonetic context. Importantly, however, the learner should also be capable of at least simple generalizations: if rule R applies in phonetic environment A, and phonetic environment B is "similar" to A, then maybe R applies in B as well? One way to operationalize the notion of similarity is to partition the phoneme set into predefined sound classes; another is to allow subphonemic descriptions. On a closer look, these are actually not very different: they both define characteristic functions and allow a rule learner to construct predicates on a given environment.

In the experiments described below, the sound classes follow a suggestion by Dolgopolsky, as adapted and extended by List (2010). In principle, the entire IPA space is partitioned into the classes in Table 1. The subphonemic description of a phoneme is simply its associated features in the traditional sense, treated as sets. A sample of the corpus thus encoded (which also exemplifies the alignment) is given in Table 2.[3]

The TBL templates chosen to operate on these features are given in condensed form below. An additional constraint was that an elided segment would not be subject to further changes.

Change segment A to segment B when …

- …(left/right) (segment/segment class) is X;
- …left (segment/segment class) is X and right (segment/segment class) is Y;
- …(left/right) (segment/segment class) is X and the next neighbour (segment/segment class) is Y;
- …(left/right) segment has feature F;

[2] http://lingulist.de/lingpy/

[3] It is worth noting that μ-TBL permits Prolog code as part of the template specifications, thus forming a little embedded language. Whether the class and phonetic features of Table 2 were prespecified or calculated dynamically only influences running time and memory use, not semantics. This is very useful in interactive experimentation.

Table 1: Dolgopolsky sound classes for phonological rules, as adapted by List (2010)

Code	Segment	Example
P	labial obstruents	p,b,f
T	dental obstruents	d,t,θ,ð
S	sibilants	s,z,ʃ,ʒ
K	velar obstr.; dent. & alv. affricates	k,g,ts,tʃ
M	labial nasal	m
N	remaining nasals	n,ɲ,ŋ
R	liquids	r,l
W	voiced labial fric.; init. rd. vowels	v,u
J	palatal approximant	j
H	laryngeals and initial velar nasal	h,ɦ,ŋ
A	all vowels	a,e,i

- …current segment shares feature F with (left/right/left and right) segment.

Results and discussion

From a training material of 210,000 phonemes and at score threshold of 10, the system learned 446 rules in about six hours. On unseen test data, the learned rules took the correspondence between truth and hypothesis from 41.2% to 73.2%. Neither of these numbers is very informative: the TBL evaluation function assumes a unique notion of truth, but for a given phonemic representation, there are many acceptable phonetic realizations. Even more ambiguously, for two given strings, one phonemic and one phonetic, there are many reasonable rule sequences that transforms the first into the second.

Indeed, quantitative evaluation of the learned rules is a challenge. One possibility is to arrange perception tests on the naturalness of the generated pronunciations; but this says nothing about the phonological validity of individual rules. For this paper, we opted for a less formal, manual evaluation. For coverage, we contented ourselves with noting that at a glance, the rule list appears to contain the majority of allophonic alternations in Danish. For rule accuracy, the main interest here, we took the first 108 rules (those with score > 100) and classified them as follows (ordered according to our intuitive idea of rule "quality"):

1. (3) False in Danish, unexplainable in data
2. (27) False in Danish, attributable to data
3. (74) Largely in agreement with current descriptions of Danish:
   (a) (49) inaccurate, could be more refined
   (b) (15) true, satisfyingly general
   (c) (10) true, interdependent with other rule/s found
4. (4) Interesting: not in agreement with current descriptions but possibly a new phonological development in progress

Table 2: The DanPASS sample of Figure 1, where phonemes are encoded with their identity (phm), class (cls), and features. Implicit time axis runs from top to bottom. The two left columns also show the resulting alignment of the phonetic (pht) and phonemic transcription.

Pht	Phm	Cls	Features
d	d	T	['voiced', 'alveolar', 'plosive']
aɪ	eːʔ	A	['length-mark', 'plosive', 'glottal', 'front', 'unrounded', 'close-mid']
-	r	R	['voiced', 'alveolar', 'trill']
-	i	A	['front', 'close', 'unrounded']
g	g	K	['voiced', 'velar', 'plosive']
ɛ	ɛ	A	['front', 'open-mid', 'unrounded']
n	n	N	['alveolar', 'nasal']
aɪ	ɛ	A	['front', 'open-mid', 'unrounded']
-	r	R	['voiced', 'alveolar', 'trill']
-	iːʔ	A	['length-mark', 'plosive', 'glottal', 'unrounded', 'front', 'close']
m	m	M	['nasal', 'bilabial']
e	e	A	['front', 'unrounded', 'close-mid']
d	t	T	['voiceless', 'alveolar', 'plosive']
-	ə	A	['schwa']
n̩	n	N	['alveolar', 'nasal']

Table 3 gives a few induced rules, chosen for illustration of the categories listed. In the following comments, "C#m" refers to the categories in the list above and "R#n" to the leftmost column of Table 3 (i.e., the position of the rule in the learned sequence).

Three of the learned rules make no sense, neither for Danish in general nor for DanPASS (C#1, R#93). These are as far as we can tell artefacts of the combined tokenization–alignment process.

More interestingly, several rules are found which are false for Danish but can be said to be true for the data (C#2). Such rules can be attributed to reductions which are uncommon in types but common in tokens (occurring, say, in a few, high-frequent function words). For instance, R#14 emanates from the modal verb /skal/ *skal* 'shall, must'. Usually, this is reduced to [sga].

The majority of the rules (C#3) can be described as reasonable, but not very interesting (outside verifying the validity of the procedure). Many of them are overly specific and would gain

Table 3: Some induced rules, in μ-TBL syntax. For instance, pht:A>B ← class:'C'@[-1] & feature:'F'@[1] *means that A transforms to B when the previous segment ([-1]) belongs to class C (Table 1) and the following ([1]) has feature F*

#	Score	Rule
1	11437	pht:r>∅ ← class:'A'@[-1]
5	1569	pht:ə>∅ ← class:'N'@[1]
14	752	pht:l>∅ ← phm:a@[-1] & phm:k@[-2]
18	605	pht:ə>ɐ ← class:'K'@[-1] & phm:r@[1]
20	561	pht:g>∅ ← phm:ə@[1]
24	458	pht:k>ɣ ← phm:ə@[1] & phm:r@[2]
29	384	pht:ə>ɐ ← class:'W'@[-1] & class:'R'@[1]
74	143	pht:ʁʔ>ʔ ← feat:open@[-1]
93	112	pht:ɒʔ>ɔʔ ← feat:approxim@[1]

from generalization (C#3a, Rs#18,29). However, some generalizations (C#3b, R#74) are indeed discovered. As is typical to phonology, many rules have a feeding order and can only be evaluated in conjunction with other rules found. In most cases the system finds such interdependent rules (but does not connect them) (C#3c, R#5).

Finally, some genuinely interesting rules are also discovered (C#4) that might for instance indicate ongoing phonological change. Thus, Rs#24,1,20) suggest progressive consonant lenition or elision, postvocalic or pre-schwa.

Conclusion

This paper presents an attempt to extract phonetic realization rules from transcribed spontaneous speech, by conditioning on local phonemic context only. Of course, we recognize that this is insufficient for real-world data, where phonetic variation can only partly be described by phonology. Other extra-phonological (information structure, word frequency, etc) and extralinguistic (speaker style, speaker mood, speech rate, acoustic environment, etc) factors are equally important. We also recognize the difficulties in evaluation, and the more general problems associated with doing phonetics on transcriptions. Nevertheless, for a first attempt, we find the results interesting, at least enough to pursue further.

One obvious source of potential improvement is additional features describing the phonetic and linguistic environment. Some of the relevant linguistic factors are readily available for featurization. DanPASS already has annotations of basic information structure and part-of-speech. At present, syllable boundaries are not part of the DanPASS phonemic annotation tier, but the second author is currently preparing their inclusion.

As mentioned, some inappropriate rules can be attributed to reductions appearing in a few, high-frequent words. Clearly, 10000 occurrences of a certain reduction in a single, high-frequent word carry much less phonological evidence than 100 occurrences in each of 100 different words. This observation could be exploited; e.g, by binning phonetic environments into lexical contexts and weighting those contexts sublinearly (e.g., logarithmically), much as sublinear term frequency scaling is used in information retrieval.

A more general problem is that of undiscovered rule generalizations. Although the current system can examine phonetic features of its surroundings, the rules work at phoneme level only. A more fine-grained representation might be beneficial, where rules are allowed to add or remove individual phonetic features. This would allow generalizations such as ”add voice to voiceless stop between two vowels”. Again, however, the more fine-grained the representation, the more fragile the beads-on-a-string assumption, and the higher the number of competing notions of truth.

Adding expressivity to the horizontal rather than the vertical direction, one might let the system simultaneously replace more than one segment. This is not very interesting for general TBL, as it comes with the cost of a much expanded search space and buys little or nothing in performance. However, in the present task the alphabet is small and the rules are the target, and it might be worth the effort.

References

Bouma G (2000). A finite state and data oriented method for grapheme to phoneme conversion. In *NAACL-2000*, 303–310. Seattle, WA.

Brill E (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565.

Grønnum N (2005). *Fonetik og Fonologi – Almen og Dansk*. Akademisk Forlag, Copenhagen.

Grønnum N (2009). A Danish phonetically annotated spontaneous speech corpus (danpass). *Speech Communication*, 51:594–603.

Lager T (1999). The μ-tbl system: Logic programming tools for transformation-based learning. In *Proceedings of CoNLL*, vol. 99.

List J M (2010). Phonetic alignment based on sound classes. In M Slavkovik, ed., *Proceedings of ESSLLI 2010, Student session*, 192–202.

Model space size scaling for speaker adaptation

Mats Blomberg
Speech, Music and Hearing, KTH/CSC, Stockholm

Abstract

In the current work, instantaneous adaptation in speech recognition is performed by estimating speaker properties, which modify the original trained acoustic models. We introduce a new property, the size of the model space, which is included to the previously used features, VTLN and spectral slope. These are jointly estimated for each test utterance. The new feature has shown to be effective for recognition of children's speech using adult-trained models in TIDIGITS. Adding the feature lowered the error rate by around 10% relative. The overall combination of VTLN, spectral slope and model space scaling represents a substantial 31% relative reduction compared with single VTLN. There was no improvement among adult speakers in TIDIGITS and in TIMIT. Improvement for this speaker category is expected when the training and test sets are recorded in different conditions, such as read and spontaneous speech.

Introduction

In this paper, initial work is presented on including a new speaker property for speaker adaptation. This property is the size of the space spanned by the set of trained models. It is well known that this type of acoustic property is closely related to articulation clarity, speech rate, and to difference between speech styles, such as read and spontaneous speech (e.g. Lindblom, 1963, Nord, 1986). There are also indications that the reduced spectral space of spontaneous speech in comparison with read speech is a major cause of the decrease of recognition accuracy in spontaneous speech (Nakamura, Iwano and Furui, 2008). These findings support the hypothesis that adaptation to this property would improve recognition performance in these types of mismatch.

In previous work, we have used vocal tract length and spectral slope for instantaneous speaker adaptation (Blomberg & Elenius, 2008, 2009). In the current paper, model space scaling is jointly estimated with these properties

Model space scaling

We apply the procedure, in a framework where the speaker properties are estimated by maximizing the likelihood output of the recognizer on the test utterance. In this search, the property values are implemented by property-specific transformations on the trained models and a recognition procedure is performed for each examined value. An alternative to transforming the models would be to perform the inverse transformation on the test utterance. We have chosen to operate on the models, since this facilitates phoneme-specific transformation.

The transformation implements a simple radial movement of the mean vector of each mixture component in a set of continuous-density HMMs towards/away-from a center-of-gravity point which is common to all models in the set. The new position of a component is derived by scaling its distance to the center-of-gravity by a scaling factor.

For a model space scaling factor α, $0 < \alpha$, the scaled mean feature vector of a mixture component will be

$$\begin{aligned}\tilde{\mu}_{ijk} &= CG_H + \alpha(\mu_{ijk} - CG_H) = \\ &= \alpha u_{ijk} + (1-\alpha) CG_H \ , \end{aligned} \qquad (1)$$

where u_{ijk} is the average feature vector of mixture component number k in state number j of model number i and CG_H is the center-of-gravity of the model set H. A scale factor value $0 < \alpha < 1$ corresponds to a compression of the model space. $\alpha > 1$ corresponds to an expansion. This linear equation is basically the same as one which was used to map formant frequencies of short vowels in mono-syllabic words to those spoken in sentences (Stålhammar, Karlsson and Fant, 1973).

It should be noted that linear scaling in the spectral or cepstral domains will not give the same result as in the formant frequency domain. Furthermore, studies have mainly been performed on vowels and the function for consonants is not as well known. For these reasons, it is uncertain if the simple linear interpolation formula in Eq. (1) will model the actual relations accurately enough to improve recognition performance.

It is possible to scale the static and the time differential elements differently. Even if both categories may be important for scaling, it is quite likely that the best scale factor value differs between them. It may therefore be necessary to estimate separate values for these feature categories.

The speech rate affects both static and dynamic features and is consequently expected to have impact on the model space. For this reason, it may be of interest to use speech corpora with mismatch in this respect for experiments. Children's speech has been found to be generally slower compared with adults (Lee, Potamianos and Narayanan, 1999) and is therefore a good candidate for evaluation. It would also be interesting to study read vs. spontaneous speech. This is planned for future work.

Three variance scaling functions have been considered. These are: (i) no change, (ii) the same scale factor as for the mean values and (iii) squared mean scale factor. In preliminary experiments, the best performance was achieved when the variance was not changed. This was used for the subsequent experiments in the paper. Further studies are required for a more decisive conclusion.

Experiments

In the experiments performed, model space size is evaluated in combination with frequency warping (Vocal Tract Length Normalization, VTLN) (Lee and Rose, 1996) and spectral slope. A low number of values of each property are examined in all combinations with the other properties. In these preliminary experiments, the model space scaling factors were tentatively set to 8 values from 0.8 through 1.5 with a linear step of 0.1. The frequency warping factor was quantized into 16 log-spaced values between 0.8 and 1.7 in TIDIGITS and between 0.79 and 1.24 in TIMIT. Spectral slope was implemented by two parameters, a spectral real pole and a spectral real zero. The pole and zero cut-off frequencies were varied in 8 logarithmically spaced steps between 100 and 4000 Hz (Blomberg and Elenius, 2009).

Corpora

Two American-English corpora, TIDIGITS and TIMIT, were chosen for initial evaluation. TIDIGITS consists of digit strings spoken by adults and children of both genders. The adult test set consists of 28583 digits. The adult male, the adult female and the children's test sets contain 14159, 14424, and 12637 digits, respectively. Models were trained on two sets: the adult (male + female) and the adult male training speakers. Evaluation was performed for the separate adult, male, female and children's test sets.

TIMIT contains read sentences of 630 adult speakers. The training set consists of 4620 utterances spoken by 462 subjects. The full test set of 1344 sentences from 168 speakers was used for evaluation.

System

The TIDIGITS experiments were performed using a connected-digit recognition system with triphone HMMs implemented in HTK. In TIMIT, monophones and a phoneme pair grammar (equal probabilities) were used.

In both cases, the acoustic models had 3 states with GMMs consisting of 32 mixture components and diagonal covariance matrices. Models were trained with a 57-dimensional acoustic feature vector, composed by 18 MFCCs and normalized log energy and their velocity and acceleration coefficients. Feature extraction was performed at a frame rate of 100 Hz with a 25 ms Hamming window and a mel-scaled filterbank of 38 filters in the range corresponding to 0 to 7.6 kHz.

Frequency warping was implemented as a piece-wise linear function using a linear transformation of models in the cepstral domain and truncation from 18 cepstral coefficients to 12 after transformation as in (Blomberg and Elenius, 2008). A standard 39-element feature vector was, thus, used in the decoder.

To reduce the computational load of searching the very large space of speaker property values, the estimation was performed by a tree-based joint search algorithm

(Blomberg & Elenius, 2009). In this procedure, an iterative recognition search starts at the root of the tree, which contains broad models representing all allowed values of the speaker properties. Child node models each represent a subset of the mother node property values. The maximum scoring child node for the test utterance is selected for further search until a leaf node is reached, whose corresponding models represent a single value of each property.

In the absence of separate development data, the insertion likelihood ("penalty") was adjusted to minimize the error rate on the baseline case of adult test data using the original adult model.

Results and Discussion

Results on TIDIGITS for varying sets of speaker properties and combinations of training and test speaker categories are presented in Table 1. Adding model space size adaptation to VTLN and spectral slope reduces the error rate by around 10% relative for children using adult or male models. The improvement when including this property indicates that there is a systematic difference between adult and child speech in the size of their spectral space and that the proposed technique can compensate for this. The spectral space difference agrees with (Lee, Potamianos and Narayanan, 1999).

There is no such error reduction visible between any of the adult speaker categories. A possible interpretation is that there is no space size mismatch between the two categories male and female speakers. Even though there may be differences between individual adult speakers in this respect, this variability is already included in the training data.

In order to have an indication of which elements of the acoustic feature vector that are mainly involved in the improvement with model space size adaptation, we ran two new experiments for children's speech against male adult models. The experiments differed from the previous ones in that only the static or the time differential elements were adapted. The results are presented in Table 2.

When model space size was performed only on the static elements of the feature vector, the error rate was not reduced compared with no size adaptation. When instead adapting only the time differential elements, the error rate decreased compared with adapting both static and dynamic elements. These results show clearly that it was the dynamic properties, which reduced the error rate by this kind of adaptation.

Even lower error rate was achieved by another selection criterion in the hierarchical search tree. When the model with the highest likelihood along the search path was chosen, the error rate was lowered further to 2.09%.

The distribution of the estimated size scale factor in the adult-male/child case and when only the time differential feature elements are adapted is displayed in Figure 1. For a majority of the utterances, the model space is compressed. Evidently, children's speech has in general slower and smoother transitions than that of adult males. This is in agreement with previous findings that children's speech is slower than that of adults (Lee, Potamianos and Narayanan, 1999). It is also obvious that the minimum allowed factor value has been set too high. Still better results are expected when this will be corrected in further experiments.

Table 1. WER for adaptation to different sets of speaker properties in TIDIGITS. Model space scaling is denoted "Size".

Train set	Adult	Adult	Adult	Adult	Male	Male	Male	Male
Test set	Adult	Male	Female	Child	Adult	Male	Female	Child
Original	0.55	0.76	0.35	3.17	6.44	0.58	12.19	46.73
Size	0.55	0.76	0.35	2.99	5.84	0.56	11.02	44.76
Slope	0.53	0.73	0.33	2.95	5.51	0.58	10.41	42.41
VTLN	0.54	0.73	0.35	1.23	0.64	0.52	0.76	3.56
Slope+Size	0.52	0.73	0.33	2.65	5.00	0.57	9.35	40.11
VTLN+Size	0.54	0.72	0.35	1.20	0.64	0.54	0.74	3.36
VTLN+Slope	0.55	0.74	0.36	1.07	0.62	0.52	0.66	2.83
VTLN+Slope+Size	0.54	0.74	0.35	0.96	0.63	0.54	0.71	2.58

Table 2. Word error rate with size estimation of different parts of the acoustic feature vector. Training and test speakers were male adults and children, respectively.

No size adaptation	2.83
All features adapted	2.58
Only static features adapted	2.85
Only Delta+Accel. features adapted	2.46

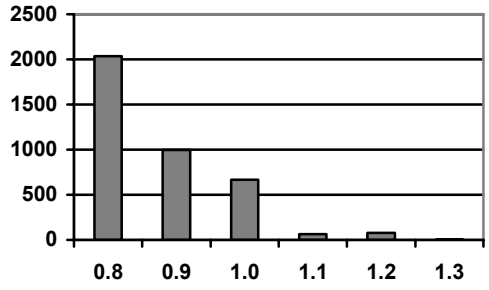


Figure 1. Histogram of number of utterances with estimated model space scale factor for time-differential acoustic features in children's speech using male adult models.

A few experiments have been performed on TIMIT. Due to computational load considerations, spectral slope was excluded from adaptation. The results are shown in Table 3.

Table 3. TIMIT results (Phoneme Error Rate). Both static and dynamic features are adapted.

Baseline	VTLN	VTLN+Size
37.03	36.66	36.64

In the adult/adult condition of TIMIT, there is a small improvement from VTLN but no further improvement from model space scaling, similarly to TIDIGITS. We tried scaling only static or differential features as well as estimating different scale factors for the two feature groups. These settings had only marginal influence on the result. A likely explanation to the lack of improvement in TIMIT is that there is no speaker mismatch between training and test speakers in the model space size respect.

Conclusions

Speaker adaptation by adjustment of model set size is efficient for the recognition of children's speech using adult or male adult models. Adding model space size to vocal tract length and spectral slope in a joint estimation framework lowered the word error rate by 10% and 9% relative, respectively, for the two training speaker categories. When also excluding static features from adaptation, the error rate using male adult models was further decreased by 5% relative. This overall combination of VTLN, spectral slope and model space scaling represents a substantial 31% relative reduction compared with single VTLN.

The method needs to be further developed for better scaling of the static features. For vowels, scaling in the formant frequency domain would be a natural choice, but the theoretical advantage is reduced by the unavoidable formant tracking errors.

Still another possibility would be to allow time varying model space size, as has been done for VTLN (Elenius and Blomberg, 2010).

Further experiments include testing on corpora with speech style mismatch between training and test, such as between read and spontaneous speech.

References

Blomberg, M, and Elenius, D (2008). Investigating explicit model transformations for speaker normalization. *Proc. of ISCA ITRW Speech Analysis and Processing for Knowledge Discovery*.

Blomberg, M, and Elenius, D (2009). Tree-based estimation of speaker characteristics for speech recognition. *Proc. of Interspeech 2009*, 580-583.

Elenius D and Blomberg, M (2010). Dynamic vocal tract length normalization in speech recognition. *Proc. of Fonetik 2010*. Centre for Languages and Literature, Lund University. 29-34.

Lee S, Potamianos A, and Narayanan S (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Amer.* 105: 1455–1468.

Lee L and Rose R C (1996). Speaker normalization using efficient frequency warping procedures. *Proc. of ICASSP*. 353-356.

Lindblom, B (1963) Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.* 35:1773-1781.

Nakamura M, Iwano K, and Furui S (2005). Analysis of Spectral space reduction in spontaneous speech and its effects on speech recognition performances. *Proc. of Interspeech 2005*, 3381-3384.

Nord L (1986). Acoustic studies of vowel reduction in Swedish. *STL-QPSR* 27/4: 19-36.

Stålhammar U, Karlsson I, Fant G (1973). Contextual effects on vowel nuclei. *STL-QPSR* 14/4: 1-18.

Gender differences in verbal behaviour in a call routing speech application

Håkan Jonsson & Robert Eklund*†‡*
** Voice Provider Sweden AB*
† Linköping University, Department of Computer Science
‡ Karolinska Institute, Department of Cognitive Neuroscience

Abstract

This paper reports results on verbal behavior in a live natural language call routing speech application. Differences between male and female callers in terms of verbosity are investigated, and put in relation to three variations of the system prompts. Findings show that in this particular application female callers are more verbose than male callers for open style prompts, while there is no difference for a directed style prompt.

Introduction

Almost all organisations and companies that handle large volumes of incoming telephony contacts from customers and end users have a need for some kind of call routing, that is, some means of assuring that calls reach their proper destinations promptly and with a high degree of service. As businesses and organisations grow in complexity, call routing has become a common application of speech technology. In many cases, such applications ask so-called open-ended questions, using natural language understanding technologies to allow callers to speak freely to describe their reason-for-calls.

This paper reports on findings in behavioural differences between genders in one such system, handling incoming calls to a major Swedish retail company.

Background

The study of open-ended call routing applications, oftentimes referred to as How May I Help You-systems, or HMIHYs, started with Gorin et al. (1997). They investigated data-driven methods for the development of such systems, and to date similar methods remain dominant, though in different flavours and with different means of representing meaning (cf. Kituno et al., 2003; Huang & Cox, 2006; Boye & Wirén, 2007; Lee et al., 2000, for a general discussion).

Behavioral aspects in call routing have been investigated since the late 1990s. For example, McInnes et al. (1999) found that using more 'open' style questions when prompting the caller for input elicited longer responses than when using a more direct, or 'closed' prompt style. Sheeder & Balogh (2003) investigated the impact on responses when the system presented examples of typical user responses before prompting the caller to speak, showing that examples increased routing precision in the application. Williams & Witt (2004) explored prompting strategies in relation to the domain of the application. They found that in a domain where callers had clear expectations on the structure of the task, open style prompting was more successful than in a domain where caller expectations were more vague. Eklund & Wirén (2010) investigated prompt style and its effects on the presence of filled pauses in user responses. They found that prompt style greatly influenced callers' verbosity in that open style prompts elicited longer answers.

None of these studies focused on caller characteristics such as gender, despite the fact that gender has most likely played a central role in human history. Detailed caller characteristics may often be inaccessible for analysis, but gender stands out in that it can usually be determined from just listening to recordings, with at least some degree of certainty.

The exact role, function and status the genders have been "given" has varied a lot throughout history, and recent Western political debate has highlighted gender aspects.

Method

The data we present in this study were collected from live callers using a prototype call routing application. The prototype application was designed solely for the purpose of data collection. It did not employ any real routing

logic, but posed as a live application in order to elicit genuine responses from callers.

The application was deployed and integrated with the main call center of a major Swedish retail company. Hence, all data are collected from real-world customers, calling in with real reason-for-calls. The original goal of this setup was to analyze and assess the viability of a full-functionality call routing application. The data utilized here is a subset of the original data set.

The prototype application, having no internal logic, presented each caller with two prompts; first an initial *main prompt*, and then one *follow-up prompt*.

For the main prompt, there were three alternative prompt wordings, or prompt styles, as follows:

Open question prompt:
"How may we help you?"
("Vad kan vi hjälpa dig med")

Basic question prompt:
"What is the reason for your call?"
("Vad gäller ditt ärende?")

Keyword prompt:
"Please state, with a word or two, the reason for your call!"
("Säg med ett eller två ord vad ditt ärende gäller")

The intention was to achieve variation in prompt style in terms of open-endedness. In any given call, the prompt was selected pseudo-randomly, even though the distribution is not equal – the basic question prompt was played in 50% of the calls, the other two in 25% each.

Follow-up prompts were also varied in a corresponding fashion, independently of the main prompts. However, effects of this variation are not detailed in this study.

The recorded utterances were transcribed and annotated according to the following classifications:

Gender – male or female. This was judged according to the annotators' perception only.

Number of words. The number of full lexical items uttered, not non-verbal sounds, etc.

Informativeness. Informative or non-informative. Whether the utterance contained any information pertaining to the subject matter of the call.

Finite verb. If there were was one (or more) finite verb uttered in the utterance.

In total, 856 calls, from 363 and 493 male and female callers, respectively, were transcribed and annotated.

Results

Table 1 shows the average number of words spoken in utterance 1 and utterance 2 for male and female callers, respectively. Note that no differentiation is made for prompt style.

Table 1. Average number of spoken words per utterance

Gender	Utt 1	Utt 2
Male	1.66	1.72
Female	2.52	2.26

As can be seen, females utter more words in both the first utterance (p = 0.000074, t-test, two-tailed) and in the second utterance (p = 0.0066, t-test, two-tailed).

Investigating whether this difference holds for all prompt styles, *Table 2* shows the average number of words for utterance 1, as a function of prompt style.

Table 2. Average number of spoken words in the first utterance, for each prompt style

Gender	Open	Basic	Keyw.
Male	2.03	1.65	1.29
Female	3.51	2.71	1.36

As is seen here, female callers utter more words than male callers in response to both the open prompt (p = 0.0029, t-test, two-tailed) and the direct prompt (p = 0.00096, t-test, two-tailed). However, for the keyword prompt, the difference is not significant (p = 0.64, t-test, two-tailed). This suggests that prompt style influences male and female callers' verbal behaviour to different degrees.

To investigate the nature of these differences further, the presence of finite verbs in responses was investigated (see Eklund & Wirén, 2010). The assumption was that finite verbs would be an indicator of more conversational responses. The results are shown in *Table 3*.

Table 3. Proportion of utterances containing a finite verb, per utterance

Gender	Utt 1	Utt 2
Male	12.9%	11.8%
Female	20.5%	15.4%

As is seen, females are more likely to include a finite verb form than are males (p = 0.0053, Z-test, two-tailed). There is a similar tendency for the second utterance, but the difference is not significant (p = 0.16, Z-test, two-tailed).

In *Table 4* it can be seen that female callers produce more utterances that contain at least one finite verb than male callers do for the open prompt (p = 0.016, Z-test, two-tailed) as well as for the basic prompt (p = 0.023, Z-test, two-tailed). For the keyword prompt, there is no significant difference.

Table 4. Proportion of utterances containing a finite verb, per utterance

Gender	Open	Basic	Keyw.
Male	14.1%	13.7%	9.0%
Female	28.8%	23.1%	9.7%

Still, as seen in *Table 5* below, there are no significant differences in terms of informative utterances between males or females, for any of the three prompt styles.

Table 5. Proportion of utterances rated as informative

Gender	Open	Basic	Keyw.
Male	87.7%	87.9%	87.2%
Female	88.2%	87.3%	84.2%

Discussion and future work

In short, the main finding here is that in this particular application, female and male callers respond differently to prompts that are more conversational in style. The female callers' responses are more verbose, and contain more finite verbs. These differences occur when the system uses a prompt style that to a higher degree encourages such behaviour, whereas for the keyword prompt, there are no significant differences between male and female callers. It would seem that in this particular case, female callers are more influenced by, or receptive to, the "speech style" of the system.

Looking for an explanation for the observed differences, the most obvious one would perhaps be that women just "speak more than men", which is also a claim that is often encountered. The origin of this claim is most likely Louann Brizendine's (2006) bestseller *The Female Brain* where it was claimed that women, on average, use 20,000 words per day, whereas men use only 7,000 words, a claim that received a lot of attention. However, closer scrutiny revealed that these figures had no scientific ground, and several subsequent, quantitative, studies either found no differences between men and women as to verbosity (Mehl et al., 2007; Cameron, 2007; Liberman, 2006), or even found that men are in fact more verbose than women (Leaper & Ayres, 2007).

However, several studies *do* point to gender differences in both cognitive abilities in general (Halpern & Tan, 2001; Mann et al., 1990) and in linguistic style/behaviour (e.g. Leaper & Ayres, 2007; Haas, 1979; Crosby & Nyquist, 1977).

While the reported general trend is that men, as a group, have superior spatial ability (Kimura, 1996), women, as a group, tend to exhibit superior verbal fluency (Halpern & Tan, 2001; Mann et al., 2001). One possible explanation for this that has been offered is that females have a thicker corpus callosum than men (De Lacoste-Utamsing & Holloway, 1982) which could perhaps also explain the often reported claim that women have less lateralized hemispheres for several cognitive functions, including language processing (Shaywitz et al., 1995).

However, although several studies point to neurological gender differences (which would also explain gender differences concerning e.g. dyslexia etc.), there are also several studies that have found no significant differences between the sexes (Brouwer, Gerritsen & De Haan, 2007; Sommer et al., 2004; Hyde, 1981). For reviews, see Cahill (2006) and Frost et al. (1999).

Another important factor that needs to be mentioned is the voice (gender) of the *system*, which in this case was male. One can speculate that females perceive this as more authoritative, leading them to speak more in the style of the prompts they respond to. It would be a natural extension of this study to also control for the variation of the gender of the system (see Nass & Brave, 2005).

These factors aside, considering other possible explanations for these behavioural differences, one possibility could be to consider how callers perceive the system. It may be argued that female callers, to a higher degree than male callers, view the system as a *conversational partner* with capabilities similar to a human. Male callers, on the other hand, would then perceive the system more as just another user *interface*. This is consistent with Edlund et al. (2008), who propose that users' conceptions of spoken dialog systems, and their interactional

style, can be explained in terms of what type of metaphor they use to conceive the system.

Finally, we can conclude that we have observed significant differences between the genders in the interaction with this particular human–machine dialog system. Whether or not this result is a fluke, or has an underlying basis in how the genders interact with systems needs to be investigated in future research.

However, irrespective of the underlying reasons for potential language use differences between the genders, the observed difference could (perhaps) conceivably constitute a parameter to consider when designing HMIHY systems with a clear gender bias in user profiles.

Acknowledgements

The authors would like to extend thanks to our colleagues at Voice Provider, in particular Carin Lindberg and Nils Hagberg.

References

Brizendine, Louann (2006). *The Female Brain*. New York: Morgan Road.

Boye, Johan & Mats Wirén (2007). Multi-slot semantics for natural-language call routing systems. In*: Bridging the Gap: Academic and Industrial Research in Dialog Technologies Workshop Proceedings*, 26 April 2007, Rochester, New York, 68–75.

Brouwer, Dédé, Marinel Gerritsen & Dorian De Haan (2007). Speech differences between women and men: on the wrong track? *Language in Society* 8:33–50.

Cahill, Larry (2006). Why sex matters for neuroscience. *Nature Reviews Neuroscience* 7:477–484.

Cameron, Deborah (2007). Applied linguistics and the perils of popularity. *International Journal of Applied Linguistics* 17(3):392–395.

Crosby, Faye & Linda Nyquist (1977). The female register: an empirical study of Lakoff's hypotheses. *Language in Society* 6:313–322.

De Lacoste-Utamsing, Christine & Ralph Holloway (1982). Sexual Dimorphism in the Human Corpus Callosum. *Science* 216:1431–1432.

Edlund, Jens, Joakim Gustafson, Mattias Heldner & Anna Hjalmarsson. (2008). Towards human-like spoken dialogue systems. *Speech Communication* 50(8–9):630–645.

Eklund, Robert & Mats Wirén (2010). Effects of open and directed prompts on filled pauses and utterance production. In: *Proceedings of Fonetik 2010*, Lund University, 2–4 June 2010, 23–28.

Frost, Julie A., Jeffrey R. Binder, Jane A. Springer, Thomas A. Hammeke, Patrick S. F. Bellgowan, Stephen M. Rao & Robert W. Cox (1999). Language processing is strongly left lateralized on both sexes. *Brain* 122:199–208.

Gorin, A. L., G. Riccardi & J. H. Wright (1997). How may I help you? *Speech Communication*, 23:113–127.

Haas, Adelaide (1979). Male and female spoken language differences: stereotypes and evidence. *Psychological Bulletin* 86:616–626.

Halpern, Diane F. & Uner Tan (2001). Stereotypes and Steroids: Using a Psychobiosocial Model to Understand Cognitive Sex Differences. *Brain and Cognition* 45:392–414.

Huang, Qiang & Stephen Cox (2006). Task-independent call-routing. *Speech Communication* 48:374–389.

Hyde, Janet Shibley (1981). How Large Are Cognitive Gender Differences? *American Psychologist* 36(8):892–901.

Kimura, Doreen (1996). Sex, sexual orientation and sex hormones influence human cognitive function. *Current Opinion in Neurobiology* 6(2):259–263.

Leaper, Campbell & Melanie M. Ayres (2007). A Meta-Analytic Review of Gender Variations in Adults' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech. *Personality and Social Psychology Reviews* 11:328–363.

Lee, Chin-Hui, Bob Carpenter, Wu Chou, Jennifer Chu-Carroll, Wolfgang Reichl, Antoine Saad & Qiru Zhou (2000). On natural language call routing. *Speech Communication* 31:309–320.

Liberman, Mark (2006). Sex on the brain. Women use 20,000 words a day, men only 7,000 – or so says a new bestseller. Fact-checking "The Female Brain". *Boston Globe*, 24 September 2006, D1.

Mann, Virginia A., Sumiko Sasanuma, Naoko Sakuma & Shinobu Masaki (1990). Sex differences in cognitive abilities. A cross-cultural perspective. *Neuropsychologia* 28(10):1063–1077.

McInnes, F. R., I. A. Nairn, D. J. Attwater & M. A. Jack (1999). Effects of prompt style on user responses to an automated banking service using word-spotting. *BT Technology Journal* 17:160–171.

Mehl, Matthias R., Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher & James W. Pennebaker (2007). Are Women Really More Talkative Than Men? *Science* 317:82.

Nass, C & S Brave (2005). *Wired for speech*. Cambridge. MA: The MIT Press.

Shaywitz, Bennet A., Sally E. Shaywitz, Kenneth R. Pugh, R. Todd Constable, Pawel Skudiarski, Robert K. Fulbright, Richard A. Bronen, Jack M. Fletcher, Donald P. Shankweiler, Leonard Katz & John C. Gore (1995). Sex differences in the functional organization of the brain for language. *Nature* 373:607–609.

Sheeder, Tony & Jennifer Balogh (2003). Say it Like You Mean it: Priming for Structure in Caller Responses to a Spoken Dialog System. *International Journal of Speech Technology* 6:103–111.

Sommer, Iris E. C., André Aleman, Anke Bouma & René S. Kahn (2004). Do women really have more bilateral language representation than men? A meta-analysis of functional imaging studies. *Brain* 127:1845–1852.

Teaching pronunciation in Swedish as a second language

Elisabeth Zetterholm & Mechtild Tronnier***
**School of Language and Literature, Linnaeus University, Sweden*
***Centre for Languages and Literature, Lund University, Sweden*

Abstract

As global migration affects Sweden in a similar way as many other countries, this contribution aims to trigger the discussion about how teaching of pronunciation for this group of students might be developed in consideration to new foreign accents in Swedish.

A new project is presented which aims to expand and update the course material for the education of future teachers in Swedish as a foreign language. Due to the new patterns of global migration in the last decades, the valuable material that exists so far needs to be expanded and needs therefore the addition of further language analyses and descriptions.

Introduction

New patterns in global migration affects Sweden in a similar way as many other countries concerning the education in second language learning. According to national statistics from 2010, 15% of the population in Sweden is born in another country (www.scb.se). More than 90000 persons were registered in the program Swedish for immigrants (SFI) during 2009. National statistics (www.scb.se) is available about the country of origin of immigrants in Sweden. This tells us something about the native languages, but gives not a clear picture about which type of students there are in the classroom. The most common first languages among immigrants in Sweden change over time and due to global migration. New analyses of Swedish foreign accents need to be expanded to some of these languages. This is the first step in a new project which aims to expand and update the course material for the education of future teachers in Swedish as a foreign language.

Everyone who has another first language than Swedish or speaks Swedish on a daily basis with one or both legal gardiens has the right to receive education in the subject *Swedish as a second language*, due to The Swedish national office of School and Education (Skolverket 2000a, b).

When learning a second language the speaker's first language plays a role, more or less (e.g. Abrahamsson 2009, Engstrand 2007). Depending on the grammar and the phonology of the first language there are different difficulties in learning a second language. Teaching Swedish as a second language is not only about phonetic variation in Swedish which is not only based on dialectal diversity, but comprises even characteristics of foreign accents of those residents in Sweden, which do not have Swedish as their first language. Therefore, teachers need more information about the differencies between the learners first and second languages and what kind of pronunciation problems there might be depending on the speakers native language. Deviant pronunciation might depend on the speaker's articulation habits as well as different perceptual cues between the L1 and the L2 speakers (Flege 1995). There is quite some agreement among the teachers of Swedish as a foreign language that intelligibility of L2-Swedish is most difficult for the L1-speakers of the South East Asian languages, which mainly seems to be related to the lack of complex consonant clusters in these languages, which are common in Swedish. Introduction of vowels to overcome consonant clustering also disturbs the language rhythm and perturbs the prosodic continuity.

Different studies show that prosodic features, such as intonation, stress and the quantity in Swedish, play an important role when learning to speak Swedish with a minimum of foreign accent (e.g. Bannert 2004, Engstrand 2007). In a comprehension-based learning program, Trofimovich et al (2009) found out that listening and reading English – but not speaking – as a second language improved the phonological

development and the pronounciation to the extend that the learners sounded just about as fluent to be easily understood.

Foreign accent and attitudes

Immigrants sometimes are judged by native speakers based on their foreign accent. There are studies indicating that pronunciation is important, not only for understanding, but for the listener's attitude to the second language speakers (e.g. Boyd 2004). Torstensson (2010) observed that listeners made differences when appraising speakers based on the type of foreign accented Swedish. Although his focus was on foreign accent in a legal setting, he also pointed out that people evaluate the speakers personal qualities depending on the foreign accent outside the court room.

Teachers should be aware of these results when teaching Swedish as a second language and therefore maybe even change focus of what is important in Swedish pronunciation and prosody. Pronunciation and accent is closely related to the speaker's identity and a slight foreign accent might be acceptable. Some types of foreign accents might even give a higher prestige to the speaker. One has therefore to ask whether it is desirable for the student to aim for the complete deletion of a foreign accent in Swedish. It is also important to know which components in the second language are more or less difficult to learn in order to understand how to teach these students. Priority must be given to everyday conversation in their new second language, for most of the students.

Immigrant languages

In the eighties of the last century, Bannert (2004, (first edition 1990)) and Garlén (1988) have done a good pursuit in collecting and describing immigrant languages and compared them to Swedish, partly with the implication to point out the potential difficulties that are likely to occur for the students of particular native languages. These studies contain descriptions of a varied number of languages, merely the phonological systems of the different languages, and also an analysis of observed pronunciation problems, based on a large amount of data collected from second language learners of Swedish. Bannert (2004) also ranks the different languages according to their distance in phonology from the Swedish language, based on the number of observed pronunciation errors and difficulties produced by the native speakers of the diverse languages collected in the material. Bannert also gives pedagogical advice on how to approach pronunciation difficulties. Besides the fact that this lecture book is out of print, the foreign accent analyses need to be expanded.

Thorén (2008) focuses on prosodic features of Swedish, mainly quantity aspects, and on how to approach them pedagogically in a second language learning situation of Swedish, without specification of the learners' first languages. McAllister et al. (1999) also are attracted by quantity aspects in Swedish and how they are produced and perceived by native speakers of English, Spanish and Estonian. However their work has more implication for the explanation of a foreign accent and does not include any pedagogical dimension.

Most frequent languages in Sweden today

The first step towards the expansion and update of the teaching material requires an analysis of which languages are more common as native languages among the learners of Swedish nowadays compared to the analyses introduced above (Bannert 2004, Garlén 1988). Swedish for immigrants (SFI) is a program for education in Swedish as a second language. All immigrants, who live in Sweden, are welcome to participate, irrespective their first language. The Swedish national office of School and Education (Skolverket) provides statistics about the native languages of the students who are enrolled in the program Swedish for immigrants (SFI). The statistics is calculated for different groups of learners, according to their enrollment for adult classes or their status in being part of the obligatory school education, i.e. children and teenagers. Table 1 gives an overview of the ten most frequent L1 of the adult learners who were registered for the years 2005-2009. It has been refrained from presenting all available statistics, as a comparison shows that there is a great amount of overlap between the L1s of the adult learners and the L1s of younger students.

It is obvious that Arabic, regardless dialect, is the most common language among immigrants learning Swedish as a second language. The increase of adult students speaking Arabic is shown in Table 1. As seen in Table 1, there has been a shift in ranking among the most frequent languages during 2005-2009. Languages such as Spanish, English, Kurdish and Bosnian/Kroatian/Serbian were more frequent a few years ago in the national

program. Somali on the other hand has clearly become more frequent. Therefore, new course material, including updated information about the most frequent languages nowadays, is needed for the students at the universities in the program of "Teaching Swedish as a Foreign Language".

Table 1. L1 of the adult students registered in the national program for Swedish for immigrants (SFI) for the years 2005-2009, in percentage [%] (taken from www.scb.se).

Rank	Language	2009	2008	2007	2006	2005
1	Arabic	23,9	24,8	22,9	18,1	17,2
2	Somali	6,8	5,7	5,6	5,2	4,1
3	Thai	6,0	6,0	5,9	6,2	6,9
4	Polish	5,4	5,7	5,6	4,5	4,0
5	Spanish	4,5	4,5	4,7	5,2	6,0
6	English	4,1	3,8	3,6	4,1	5,1
7	Kurdish/ North Kurdish	3,6	3,9	4,5	4,7	4,7
8	Persian	3,1	2,9	3,3	3,6	3,7
9	Bosnian/ Kroatian/ Serbian	2,9	3,2	4,0	4,7	4,8
10	Turkish	2,9	2,9	3,0	3,3	3,8
	others	36,7	36,6	37,0	40,3	39,7

An enquiry addressed to teachers currently engaged in teaching Swedish as a foreign language (scholarly year 2010-11) was undertaken, by the authors. Most of the teachers work in the southern part of Sweden. The ranking of the occurrence of the different L1s were assembled according to different components: a numbered list was provided where the encountered L1s could be ranked in the questionnaire, according to their apprehended frequency. Furthermore, the number of students was asked for representing the different L1s. However not all teachers filled in the figures so that measure could not count alone for the frequency of occurrence. Table 2 shows the ranking of L1s gathered from the authors' questionnaire filled in by the teachers. It was chosen to group together the languages Persian, Dari and Pashto, as some teachers did not made the distinction in the questionnaire, coherent with not making a difference between different Arabic or Chinese variants.

Table 2. Most frequent languages acquired by the authors' enquiry

Rank	Language
1	Arabic
2	Somali
3	Bosnian/Kroatian/Serbian
4	Albanian
5	Chinese
6	Turkish
7	Kurdish
7	Vietnamese
9	Russian
10	Persian/Dari/Pashto
11	Thai

When comparing Table 1 and 2 it is clear that there are some similarities concerning languages in the top of the rankings, namely Arabic and Somali. There is also an interesting difference between languages such as Thai and Bosnian/Kroatian/Serbian. Some languages occur only in one of the tables and it is unclear what languages 'others' include in Table 1. The differences between the two tables might be explained by the amount of answers, time and place differences.

Further statistics about the frequency of occurrence of various languages in Sweden is also presented in Dahl (2010). His data however accommodates languages that are rather unlikely to be represented in the classrooms of learners of Swedish as a second language, like other Scandinavian languages.

Spelling and pronunciation

One interesting question is if the speaker's pronunciation of the second language has any impact of the spelling in the new language. In a study by Andersson (1981) he argue that since the phonological system of the learner's first language has an impact of the second language that might influence the spelling in Swedish. Andersson found that it is mostly the Swedish vowels å, ä, ö, the length of the consonants and the different spelling of sounds like [ɔ], [ʂ] and [ɕ] that gives the second language learner problems with their spelling in Swedish.

Except from the study by Andersson there is only a few student papers about this subject

(Hökbring 2008, Vitikka 2009). In this ongoing project the relationship between spelling and pronunciation will be studied using recordings of children and papers written by the same children at a school in Rosengård, Malmö.

Further analysis procedure

Several steps have to be taken for the compilation of updated course material. This includes generation and revision of phonological descriptions and the making and analysis of recordings of Swedish speech of speakers with different L1s. The analysis of foreign accent features in Swedish due to the L1 of the speaker is aimed to be based on recordings of the L2-Swedish learners. These recordings will contain read words and phrases and spontaneous speech, like the description of a picture. One extra idea is to let the L2-Swedish learners try to imitate native Swedish speech. These imitations will also be recorded and taken into account for foreign accent analysis, basically in respect to prosodic features.

The phonological systems of the most common immigrant languages and the typical foreign accent features need to be made clearer, including more listening examples to the future teachers of Swedish as to what extent they were presented in the current material, so that they can help the L2-learners of Swedish to improve their communication skills.

References

Abrahamsson (2009). *Andraspråksinlärning*. Sweden: Studentlitteratur.

Andersson A-B (1981). Diktamensövningen. In: Tingbjörn G & Andersson A-B, *Invandrarbarnen och tvåspråkigheten*. Sweden: Liber Utbildningsförlaget, 58-95.

Bannert R (2004). *På väg mot svenskt uttal*. Sweden: Studentlitteratur.

Boyd S (2004). Utländska lärare i Sverige – attityder till brytning. In: Hyltenstam K & Lindberg I, eds. *Svenska som andraspråk – forskning, undervisning och samhälle*. Sweden: Studentlitteratur.

Dahl Ö (2010). *Språken i Sverige*. Sveriges Nationalatlas, Norstedts Förlagsgrupp AB, Stockholm, Sweden.

Engstrand O (2009). *Fonetik light*. Sweden: Studentlitteratur.

Flege J, Munro M, MacKay I (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America* 97:3125-34.

Garlén C (1988). *Svenskans fonologi*. Studentlitteratur, Lund, Sweden.

Hökbring A (2008). *Påverkar uttalet stavningen? En undersökning av svenska som andraspråksinlärares avvikande uttal och hur det påverkar stavningen*. Examensarbete. Göteborgs universitet: Institutionen för svenska språket.

McAllister R, Flege J, Piske T (1999). The Acqusition of Swedish Long vs. Short Vowel Contrast by Native Speakers of English, Estonian and Spanish. In: *Proceedings of the XIVth International Congress of Phonetic Sciences*, 751-754.

Praat, http://www.praat.org

Skolverket, http://www.skolverket.se/sb/d/1745/a/20537, (requisition 6.3.2011)

Skolverket (2000a). Svenska och svenska som andraspråk. Stockholm: Skolverket. http://skolverket.se/sb/d/1478

Skolverket (2000b). Kursplan för svenska som andraspråk i grundskolan. Stockholm: Skolverket. http://www3.skolverket.se/ki03/front/aspx

Statistiska Centralbyrån, http://www.scb.se

Thorén B (2008). *The priority of temporal aspects in L2-Swedish prosody*. Doctoral Thesis in Phonetics at Stockholm University, Sweden.

Torstensson N (2010). *Judging the Immigrant: Accent and Attitudes*. Doctoral Thesis, Department of Lanuage Studies, Umeå University, Sweden.

Trofimovich P, Lightbrown P M, Randall H Halter, Song H (2009). Comprehension-based practice. The development of L2 Pronunciation in a Listening and Reading Program. In: *SSLA* 31:609-639. Cambridge University Press.

Vitikka J (2009). *Finns det något samband mellan andraspråksinlärares uttal och stavning i svenska? En undersökning om hur uttal och stavning påverkar varandra*. Examensarbete. Göteborgs universitet: Institutionen för svenska språket.

Detecting confusable phoneme pairs for Swedish language learners depending on their first language

G. Ananthakrishnan, Preben Wik, Olov Engwall

Centre for Speech Technology, KTH

Abstract

This paper proposes a paradigm where commonly made segmental pronunciation errors are modeled as pair-wise confusions between two or more phonemes in the language that is being learnt. The method uses an ensemble of support vector machine classifiers with time varying Mel frequency cepstral features to distinguish between several pairs of phonemes. These classifiers are then applied to classify the phonemes uttered by second language learners. Using this method, an assessment is made regarding the typical pronunciation problems that students learning Swedish would encounter, depending on their first language.

Introduction

Computer Assisted Pronunciation Training (CAPT) is a fast growing and an important aspect of Computer Assisted Language Learning (CALL) systems. However, problems faced by student with different first language (L1) backgrounds are often very different. At the same time some of the problems are common to almost all language backgrounds. In the context of Computer Assisted Pronunciation Training (CAPT), this aspect is very relevant. Bannert (1980) pointed out what sounds and phonemes in Swedish may be confusing to students depending on their L1 and found a large variation. However, this study was made based on expert knowledge of a trained phonetician, based on students with classical pronunciation problems. In this paper, we describe an automated method that can extract such knowledge from data collected from several second language (L2) students. Such explicit knowledge can be used to increase the accuracy of detecting specific types of pronunciation errors, as well as developing customized training methods for students with a particular L1 background.

We follow the approach of Truong (2004) who used a set of binary classifiers, to help classify often confused phonemes. The above study required careful selection and construction of the acoustic parameters in order to make reliable classifications and claimed detection accuracies of somewhere between 70 and 90 %. They also tried to train their classifiers on native as well as non-native speech, and found that the performance, as expected, was better on native speech, which in general showed lower variance.

In our approach, we extend this methodology to include a very large number of classifiers in order to assess what kind of pronunciation errors and confusions are most probable, given the L1 of the student. This requires a method in which the same classification system should in principle hold for classifying several classes of pairs of phonemes. Since different kinds of acoustic features are useful for classifying different types of phonemes, including static as well as dynamic sounds, our approach requires a common platform to select the suitable features automatically. Secondly, the accuracy when classifying different types of phonemes would also be largely different. To side-step this problem, we do not endeavor to make assessments on every incorrect utterance, but instead make a judgement on entire sessions of utterances (around 80) of several speakers (2 to 24). We compare the performance of the classifiers on native speech (assuming the native speech to be correct). We assess only on those phoneme pairs, on which the classifiers report significantly higher error rates for L2 learners than on native speech, to be problematic.

Ensemble of Classifiers

The block diagram of the ensemble classification framework we used in this study is illustrated in Figure 1, previously described in (Picard et al., 2010). Given the acoustic signal and the text of what the subject is supposed to have uttered, the acoustic signal is segmented into the sequence of phonemes using an Hidden Markov Model (HMM) based alignment (Sjölander and Heldner, 2004). We use the native speech for training our models and test them on non-native speech ut-

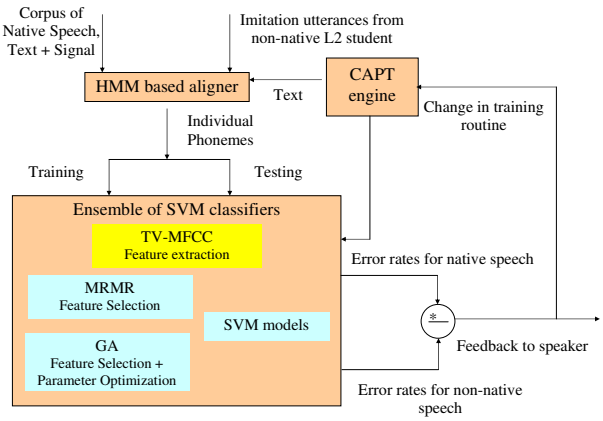


Figure 1: Block Diagram of our system using an ensemble of binary classifiers, to detect specific mispronunciation errors, by finding significant performance differences in the performance on native speech and non-native speech.

tered by the L2 language learners. The input to the classification framework are the acoustic segments of individual phonemes. The classifier system consists of 4 components, described below.

Acoustic Features

In order to model static as well as transient sounds, we used dynamic features in the form of ime-Varying Mel Frequency Cepstral Coefficients (TV-MFCC). If $A(f,t)$ is the time varying log spectrum of the audio signal, with F mel frequency sub-bands and T time samples, then $\tilde{A}(p,q)$ are the 2D-DCT coefficients obtained by performing DCT along the dimensions f (frequency) and t (time). The dimensions $p : \{1 \leq p \leq P\}$ and $q : \{1 \leq Q\}$ are called 'quefrency' and 'meti' respectively (Picard et al., 2010). The duration is added at the end of the vector. Thus, the total number of features are $P * Q + 1$.

Minimum Redundancy Maximum Relevance

Since we use many different kinds of binary classifiers, the most relevant and optimum acoustic features are not always the same. We therefore use two feature selection algorithms. Minimum Redundancy Maximum Relevance (MRMR) (Peng et al., 2005) relies on estimating feature redundancy (selecting features that are dissimilar to each other) and relevance (maximizing the contribution of the features towards classification) using mutual entropy, through a greedy search. Processing time varies linearly with the number of features to be retained. This method reduces the search space by a large amount, and thus the time taken for the Genetic Algorithms (GAs) to converge.

Genetic Algorithms

In order to ensure optimal performance for the binary classifiers, the poor set of features needs to be discarded. Besides, the optimum parameters for each classifier would also be different. We, therefore, employ an implementation of GA (Goldberg, 1989) with a K-fold cross validation scheme in order to select both the feature indices and to optimize the parameters of the classifier.

Support Vector Machines

Support Vector Machine (SVM) try to find the best possible hyper-plane separating two classes, by maximizing the distance between the elements of the two classes. SVMs are known to perform well for binary classification problems even on sparse and high dimensional data. We also use the Gaussian kernel in order to allow non-linear hyper-planes. The SVM models are trained on native speech, using a K-fold cross-validation different folds, applying MRMR to each fold, then applying the genetic algorithm to find the optimal features and the classifier parameters over all the folds. The optimum features and parameters are then used to train the SVM binary classifiers over each fold. The error rates over each fold is calculated using the optimal set of selected features and parameters.

Data and Experiments

The corpus consisted of 78 phonetically rich short sentences and words uttered by 11 native Swedish speaking subjects who recorded the utterances reading a text displayed on the screen. This was done using a desktop microphone and a sampling frequency of 16 KHz. The material was used for training 95 binary classifiers using the process described in the above section. The non-native speech data consisted of 2 to 24 students each from 11 different L1 backgrounds, learning to speak Swedish in a Swedish course. The students used the VILLE Swedish virtual language tutor (Wik et al., 2009; Wik and Hjalmarsson, 2009) and produced the utterances while trying to repeat the words or sentences uttered by the virtual tutor. The data was cleaned up to remove instances of hesitations or completely incorrect utterances in the data. In this experiment, the native and non-native speakers recorded the same set of utterances, but in principle, they can be completely different.

In total, 95 pairwise classifiers under 6 categories were created. The 6 categories were

1. Plosive *vs.* Fricative (PF) (6 pairs)
2. Voiced *vs.* Unvoiced consonants (VU) (5 pairs)
3. Front *vs.* Back vowels (FB) (23 pairs)
4. Short *vs.* Long vowels (SL) (11 pairs)
5. Unrounded *vs.* Rounded Lips (UR) (23 pairs)
6. Open *vs.* Closed vowels (OC) (27 pairs)

Under each category, all possible confusable pairs of phonemes were considered. For each phoneme, TV-MFCCs, with the number of quefrency coefficients $P = 18$, and meti coefficients $Q = 3$, were extracted. The total number of features were 55 initially, including the duration of the phoneme. At the first stage, MRMR was performed to select the best 20 features with respect to the particular binary classification task. Optimization was then performed using GA with a 4-fold cross-validation, and the best features and classifier parameters were chosen. The time taken to build each classifiers ranged from less than a second to 26 minutes, depending on the number of samples available for the respective phonemes, in a MATLAB^{TM} implementation of the algorithms. The error rate of the classifiers on native speech ranged from 0 to 34%, assuming that the natives had a perfect pronunciation. The worst performing category was the Short *vs.* Long vowel category. For every L2 learner (student), all the phoneme boundaries were extracted using the HMM based alignment using the phonemic transcription of what they were supposed to have uttered. All the relevant phonemes were chosen from the entire session of each students from a particular L1 background and classified using the ensemble of classifiers.

Classifiers performing with an accuracy of around 70% on native speech would normally not be useful for providing pronunciation feedback on non-native speech. Therefore, we adopted a method to side-step the problem. A binomial significance test was conducted to see if the error rate estimated by each classifier on the utterances by students with a said L1 background was significantly higher than on the native speech. Thus, even if the classifier error rate is quite high on native speech, it could still be useful for providing suitable assessment. In this study we illustrate only three examples per L1 background, with the highest probability of the error rate being significantly higher.

Results and Discussion

Table 1 displays the three most likely pairs of Swedish phonemes that could cause confusing pronunciations from students of each L1 background. It should be noted that this table does not provide an exhaustive list of confusions, but only the three phonemes pairs with largest significant classifier error rates compared to native speech. Again, since the sample set of students was rather small in some cases, the examples may not be indicative of the entire language group.

From a cursory glance at the Table 1, it is clear that most of the errors are made in the Swedish vowels. The distinction between /ɛ/ *vs.* /ə/ is most confusing to students from almost all language backgrounds. Some other recurrent confusable pairs are /ɑː/ *vs.* /oː/ and /ɛː/ *vs.* /e/. Most of the errors made on consonants were in voicing. Several language groups like Polish, Greek and Persian had significantly larger error rates from the classifiers than native Swedish pronunciation. Polish stop consonants can be hard or soft depending on the vocalic context. This difference in pronunciation may be the source of errors while pronouncing Swedish sounds. For Greek, the unvoiced stop consonant disappear in the context of nasal sounds. This may be the source of errors for Greek pronunciations. The high error rates for Persian voiced and unvoiced distinctions are, however, puzzling. This may, however, be due to different ranges of voicing onset for Persian and Swedish stop consonants. Students with a Spanish L1 background are known to make errors when producing the sound /oː/ and /ɔ/, which is reflected in the confusions, /øː/ *vs.* /ɔ/ and /ɑː/ *vs.* /oː/. Some language backgrounds indicate significantly larger classifier errors for lip rounding distinctions, such as ə → œː for Arabic and Polish L1 backgrounds.

This method, thus, automatically lists problem areas for students with different L1 backgrounds, which will help the CAPT system to provide a customized training material according to the apriori knowledge. The power of this paradigm is the ability to circumvent the low accuracies of certain classifiers as well as to locate specific problems. The paradigm can be extended to detecting specific cases of mispronunciations. However, not all the classifiers may be accurate enough to be used in specific detections.

Acknowledgements

We would like to thank the Swedish Research Council projects 80449001, Computer-Animated Language Teachers (CALaTea) for financial support. We would also like to acknowledge the help of Chris Koniaris for processing the data which we used in this study.

Table 1: Illustration of the three most common confusions that students with different L1 backgrounds, learning Swedish as an L2, as estimated by our algorithm. The arrow indicates in which direction the confusion is detected to occur.

L1 Background (No. of students)	Error Category	Confusable phonemes (IPA)	Examples	Diff. in Error (%) between native and non-native speech
American English (4)	VU	ɡ → k	**g**ås → **k**al	12%
	OC	æː ← ə	h**ä**r ← pojk**e**n	11%
	OC	oː → ʊ	h**å**l → b**o**tt	10%
Arabic (2)	OC	ɛ ← ə	r**ä**tt ← pojk**e**n	18%
	PF	b → v	**b**il → **v**år	16%
	UR	ə → œː	pojk**e**n → f**ö**r	15%
Mandarin Chinese (10)	VU	ɡ → k	**g**ås → **k**al	17%
	OC	ɑː ← oː	h**a**l ← h**å**l	13%
	OC	æː ← ə	h**ä**r ← pojk**e**n	12%
French (4)	OC	ɛ → ə	r**ä**tt → pojk**e**n	10%
	UR	ə → ø	pojk**e**n → f**ö**ll	9%
	OC	ɔ ← ʊ	h**å**ll ← b**o**tt	8%
German (3)	VU	ɡ → k	**g**ås → **k**al	11%
	OC	ɔ → ʊ	h**å**ll → b**o**tt	9%
	OC	ɛː ← e	r**ä**tt: ← v**e**tt	8%
Greek (5)	OC	ɛ → ə	r**ä**tt → pojk**e**n	8%
	OC	ɛ ← ə	r**ä**tt ← pojk**e**n	8%
	OC	ɔ ← ʊ	h**å**ll ← b**o**tt	7%
Persian (24)	VU	d → t	**d**al → **t**al	14%
	OC	ɑː ← oː	h**a**l ← h**å**l	11%
	VU	ɡ → k	**g**ås → **k**al	11%
Polish (4)	VU	ɡ → k	**g**ås → **k**al	30%
	VU	d → t	**d**al → **t**al	24%
	UR	ə → œː	pojk**e**n → f**ö**r	20%
Russian (7)	OC	ɑː ← oː	h**a**l ← h**å**l	11%
	OC	ɔ ← ʊ	h**å**ll ← b**o**tt	11%
	OC	ɛ ← ə	r**ä**tt ← pojk**e**n	9%
Spanish (11)	FB	øː ← ɔ	f**ö**ll: ← h**å**ll	14%
	OC	ɛ ← ə	r**ä**tt ← pojk**e**n	14%
	OC	ɑː ← oː	h**a**l ← h**å**l	12%
Turkish (7)	OC	ɑː ← oː	h**a**l ← h**å**l	20%
	FB	œː ← ɑː	f**ö**r ← h**a**l	10%
	OC	ɛ ← ə	r**ä**tt ← pojk**e**n	10%

References

Bannert R (1980). *Svårigheter med svenskt uttal: inventering och prioritering*. Inst. f ”or lingvistik, Lunds univ.

Goldberg D (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-wesley.

Peng H, Long F and Ding C (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 1226–1238.

Picard S, Ananthakrishnan G, Wik P, Engwall O and Abdou S (2010). Detection of Specific Mispronunciations using Audiovisual Features. In *Proc. Int. Conf. on Auditory-Visual Speech Processing*. Kanagawa, Japan.

Sjölander K and Heldner M (2004). Word level precision of the NALIGN automatic segmentation algorithm. In *Proc. of Fonetik*, 116–119.

Truong K (2004). *Automatic pronunciation error detection in Dutch as a second language, an acoustic-phonetic approach*. Master’s thesis, Utrecht University, The Netherlands.

Wik P, Hincks R and Hirschberg J (2009). Responses to Ville: A virtual language teacher for Swedish. In *Proc. SLaTE*. Wroxall Abbey Estates, UK: Citeseer.

Wik P and Hjalmarsson A (2009). Embodied conversational agents in computer assisted language learning. *Speech communication*, 51(10):1024–1037.

Do Germans produce and perceive the Swedish word accent contrast? A cross-language analysis

Regina Kaiser
Institute of Phonetics and Speech Processing, University of Munich, Germany

Abstract

The present paper presents the results of a cross-language analysis with 10 L1 German speakers who have learned Swedish for at least 18 months. In an imitation and a perception experiment they had to produce and identify the Swedish word accents in the first post-focal position. 15 L1 Swedish speakers took part in the same perception experiment and also evaluated the imitations of the L1 German speakers. Results were analysed according to the type of the experiment and to the native language of the subjects. As expected the L1 German speakers could not identify the word accents. Surprisingly 5 of the L1 Swedish speakers were not able to identify them either. The imitations of the L1 German speakers were arbitrary and did not show a preference towards the accent I as it was assumed from a comparison of the tonal inventory of Standard German and Standard Swedish.

Introduction & Background

Although German and Swedish are both Germanic languages, they developed differently concerning the use of the fundamental frequency (F0). German as an intonation language uses pitch only at the utterance level. However, Swedish as a pitch accent language uses pitch to mark one of the stressed syllables of a prosodic word by associating this syllable with a word accent. A comparison of the tonal contours of the two Swedish word accents with the tonal inventory of Standard German shows that there is at least one contour tone that can be found in both languages according to the Lund model for Swedish (Bruce, 1977; Bruce and Gårding 1978; Bruce, 2005) and GToBI (German Tones and Break Indices, Grice and Baumann, 2002; Grice et al., 2005) for German. In the notation of these models, the two Swedish word accents are represented by H+L* for accent I and H*+L for accent II. The former is similar to the H+L* in GToBI in that in both cases the F0 falls from a peak to a valley that is in the stressed syllable: for this reason, the H+L* accent is sometimes called an early peak in other notational systems (e.g. Kohler, 1987). On the other hand, there is no direct equivalent in GToBI to the H*+L accent in Swedish.

The fundamental frequency in Swedish is influenced not only by word accents but also on the position of the word accent in relation to the focus accent in the phrase. The focus accent marks the prosodic word that is prominent in the phrase by associating it additionally with the focus tone H-. All accents, focus accent as well as word accents, are produced with a single F0-peak. The sole exception occurs in the first post-focally accented word. In this position the F0-peak of the word accent is the final part of a F0-plateau and has no directly preceding rise. (Bruce, 1977) Hence, the only difference between the two word accents is one of timing. The timing contrast of the word accents can be observed in the first post-focal position.

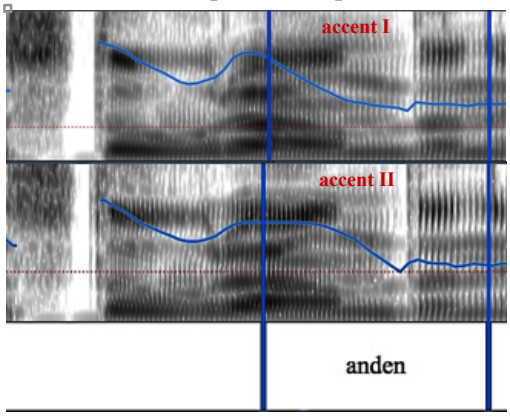


Figure 1. Accent I word anden *(the duck) and accent II word* anden *(the spirit). The vertical lines mark the beginning and the end of* anden *in the sonagram.*

For example, Figure 1 shows the timing difference of the F0-fall in the phrase *Det var den STORA anden jag menade* ('It was the the great duck/spirit I meant') in which *stora* is associated with the focus accent. The minimal pair *anden* ('the duck'/ 'the spirit') is in the first post-focal position. As it is evident from Figure 1, the timing of the F0-fall differentiates the word accents (Bruce, 1977). In fact, the gradient of the fall has no influence on the perception of the word accents.

In this study, we are interested in how L1 German speakers, who are learning Swedish, produce and perceive the timing contrast of the word accents. Taking into account that there is no comparable tonal contour for accent II in Standard German, the following hypothesis can be made:

H1: L1 German speakers only produce accent I in Swedish.

H2: L1 German speakers cannot distinguish between the Swedish word accents.

Hypothesis 1 postulates that L1 German speakers only produce the accent I when speaking Swedish. The second hypothesis is a further qualification of the first: according to H2, L1 German speakers cannot produce the accent II because they do not perceive any difference between the two word accents.

Method

A male L1 Standard Swedish speaker produced the sentence *Det var den STORA anden jag menade*, once with the accent I word *anden* ('the duck') and once with the accent II word ('the spirit'). In both sentences, *stora* was focused because the L1 Swedish speaker was instructed to produce the sentence as an answer to the question "*Var det den lilla anden du menade?*" ('Was it the small duck/spirit you meant? '). His productions were manipulated with the PSOLA resynthesis in Praat in order to create a 10-step continuum with the first stimulus of the continuum being the non-manipulated sentence with the accent I word and the last stimulus a sentence with the accent II word. The non-manipulated sentence with the accent II word was not used in the experiment, but the time of the occurrence of H and L in relation to the vowel segment was used as a reference point for the manipulation of this tenth stimulus.

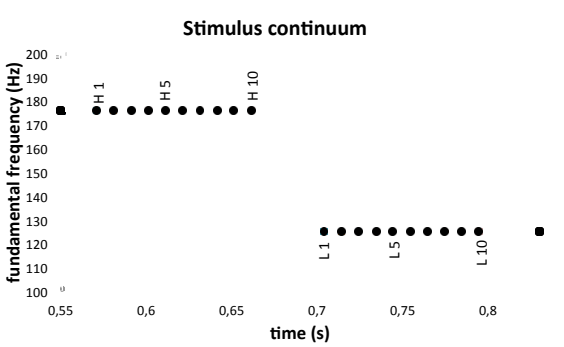


Figure 2. The schematic F0-gradients of the stimuli of the continuum based on the non-manipulated accent I sentence. H1-L1 represents an accent I fall, H10-L10 an accent II fall.

The F0-fall of the word accent was shifted parallel in equal 10 ms steps thereby deriving 9 further stimuli from the first original one. The slope of the F0-fall was constant in all cases (Fig. 2).

10 L1 German speakers (5 male, 5 female; age from 20 to 28 years) who had learned Swedish for at least 18 months participated in an imitation experiment in which they were asked to imitate the first (accent I) and the last (accent II) stimulus of the continuum. For the experiment the software SpeechRecorder (Draxler and Jänsch, 2010) was used so that the subjects could read the sentence on a screen 1 s before they heard the production of the L1 Swedish speaker over Beyerdynamic DT 770 headphones. The recordings were made in a semi-anechoic room with a Beyerdynamic Opus 54 microphone. Each L1 German speaker produced 20 sentences, 10 as imitations of the sentence with the accent I word and 10 with the accent II word in the sentence. In addition, each subject had to imitate *anden* 10 times as an isolated accent I word and 10 times as an isolated accent II word. The stimuli to be imitated by L1 German speakers were presented in randomized order.

The 10 subjects then took part in a forced-choice perception experiment in which they had to judge whether each stimulus from the continuum meant 'the duck' or 'the spirit'. For this purpose, the stimuli were presented 10 times (thus 100 stimuli in total) in random order. 15 L1 Swedish speakers (3 male, 12 female; average age 32) also took part in the same forced-choice perception experiment. In addition, the Swedish listeners were presented with the German speakers' imitations and also

had to decide whether the imitation were corresponding to 'duck' or 'spirit'. The stimuli were presented in an online experiment to the Swedish listeners who made forced-choice judgements to a total of 300 stimuli including 100 stimuli as described above plus a total of 200 imitations produced by the L1 German subjects. Since a subsequent analysis showed that five L1 Swedish speakers were unable to distinguish perceptually between the two accents, we excluded their data from the further analysis.

Results

The results from the perception experiments with the L1 Swedish speakers and the L1 German speakers were divided into two parts. First the results from the experiments, in which the stimulus continuum was used, were evaluated. For each listener separately a generalized linear mixed model (GLMM) in R was used to obtain a psychometric response curve that was fitted to the raw data using a logistic function (Kleber et al., in press) from which the 50% cross-over boundary was obtained.

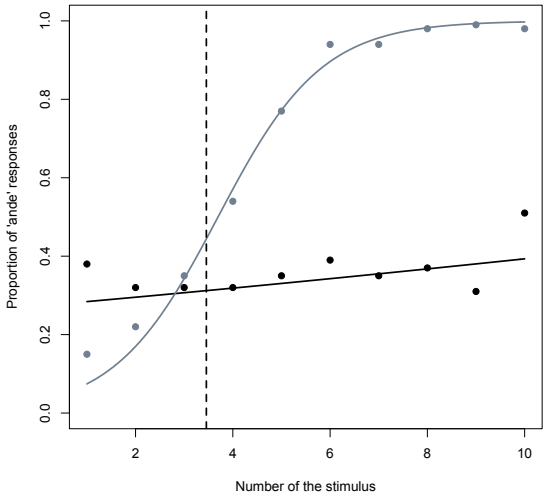


Figure 3. Psychometric response curves averaged across the Swedish (grey) and German (black) listeners. The points show the mean response values for each of these two groups to which the curves were fitted. The 50% cross-over boundary for the Swedish listeners is marked by a vertical dashed line.

One of the most striking aspects of the data in Figure 3 is the difference between the two groups: whereas there was evidently a clear S-shaped response curve and an identifiable cross-over boundary for the Swedish listeners, neither of these were in evidence for the German listeners for whom the response curve was flat. These data show that the Swedish, but not the German, listeners were able to distinguish perceptually between the two categories. These between group differences that are evident in Figure 3 were quantified by comparing the gradient of the logistic function (one per listener): a t-test showed that the gradients were significantly different ($t[18] = 8.2$, $p < 0.001$) for the two groups, thereby confirming that the Swedish listeners were better able to distinguish between the two accents than were the German listeners. Moreover, no cross-over boundary could be found for the German listeners (that is, the theoretically calculated cross-over boundary was beyond the range of the stimuli) showing once again that the German listeners were not able to distinguish perceptually between the two meanings.

In the second part of the analysis the results from the perception experiment with the imitations of the L1 German speakers were evaluated. These stimuli had only been presented to L1 Swedish speakers.

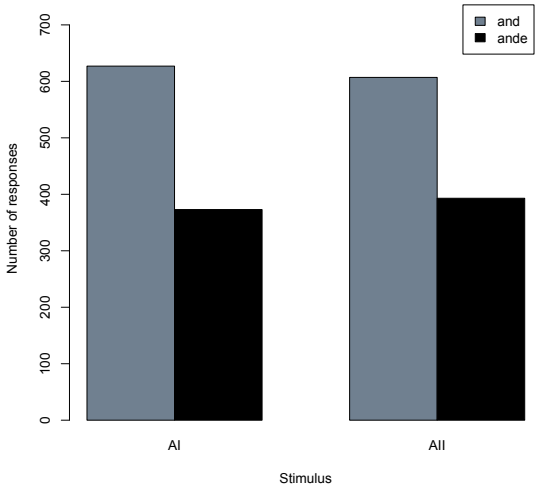


Figure 4. The number of 'duck' (grey) and 'spirit' (black) responses pooled across the Swedish listeners to the AI-endpoint (left) and AII-endpoint (right) imitations produced by the L1 German speakers.

As Figure 4 shows, the L1 Swedish listeners perceived the majority of the L1 German speakers' productions as accent I. This was so, even when the German speakers imitated an

accent II. Compatibly, a GLMM with the L1 Swedish listeners' response as the dependent variable, imitated accent type (two levels: accent I, accent II) as fixed factor and with listener as a random factor showed no significant influence of imitated accent type on the Swedish listeners' responses.

Summary and Discussion

Based on a comparison of the tonal inventory of Standard Swedish and Standard German, we hypothesized that L1 German speakers are not able to produce the Swedish word accent II when it appears in the first post-focal position and that they produce the accent I instead. As the analysis of the imitations of the L1 German speakers showed (Fig. 3), they did not only produce the accent I, but also the accent II. Nevertheless, the L1 German speakers did not differentiate the two word accents. Five L1 German speakers produced the accents more or less arbitrarily. Four L1 German speakers produced overall more accent I words. And one subject produced more accent II words for both stimulustypes. In general, these results show that the L1 German speakers were not able to associate the word accents with different meanings.

The perception experiment with the stimulus continuum showed that the L1 German speakers could neither identify the word accents nor distinguish between them in production. Surprisingly, neither were 5 of the 15 L1 Swedish speakers able to identify the stimuli in the continuum, although 3 of them were from the Stockholm region that is known as an area where Standard Swedish is spoken. The reason why some of the L1 Swedish speakers had such difficulties in identifying the word accents may be that in an everyday conversation their meaning is inferable from context. Thus, the present experimental condition, in which they had to identify the meaning of a minimal pair word without any context, was quite unusual even for a native speaker.

The L1 German speakers had no difficulty imitating the word accents in isolated words in which the two word accents were differentiated by the number of F0-peaks (accent I: 1 F0-peak; accent II: 2 F0-peaks). But they were not able to neither perceive nor produce the timing contrast of the word accents in the post-focal position. This might be related to the point mentioned above that the L1 German speakers did not link the word accents to different meanings. In any case, the influence of the focus accent should not be underestimated. For an L1 German speaker, the prominent focus accent could mask the slight timing difference between the F0-falls of the word accents in post-focal position: this might lead to an additional complication in the identification of the word accents by the L1 German speakers.

References

Boersma P, Weenink D (2010). Praat: doing phonetics by computer (version 5.1.41). Webpage http://www.fon.hum.uva.nl/praat/

Bruce G (1977). Swedish Word Accents in Sentence Perspective. Lund: CWK Gleerup.

Bruce G, Gårding E (1978). A Prosodic Typology for Swedish Dialects. In: E Gårding, G Bruce, R Bannert, eds, *Nordic Prosody*. Lund: CWK Gleerup, 219-229.

Bruce G (2005). Intonational Prominence in Varietis of Swedish Revisited. In: S-A Jun, ed, *Prosodic Typology. The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 410-429.

Draxler C, Jänsch K (2010). SpeechRecorder (version 2.2.8). Webpage http://www.phonetik.uni-muenchen.de/Bas/software/speechrecorder/

Grice M, Baumann S (2002). Deutsche Intonation und GToBI. (Online: www.coli.uni-saarland.de/publikationen/softcopies/Grice:2002:DIG.pdf)

Grice M, Baumann S, Benzmüller R (2005). German Intonation in Autosegmental-Metrical Phonology. In: S-A Jun, ed, *Prosodic Typology. The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 55-83.

Kleber F, Harrington J, Reubold U (in press). The relationship between the perception and production of coarticulation during a sound change in progress. In: *Language & Speech*, in press.

Kohler K (1987). Categorical Pitch Perception. In: *Proceedings of the 11th ICPhS*, 331-333.

Chinese perception coaching

Guohua Hu

Department of Languages and Literatures, University of Gothenburg

Abstract

The article first presents the nowadays reality of Chinese as Second Language (CSL) in Sweden. It is empersized that Chinese Perception Coaching (CPC) should start before the Chinese speech production.

It also introduces how CPC should be planed both at group/class and individual level. The analysis of Perceptual Assimilation Model (PAM) is used in CPC. In particular, it suggests how to build the language awareness of the students. Finally, it shows a possible way for research in this subject in order to improve the practice.

Introduction

The main difficulties of Chinese as a second language (CSL) for learners having Swedish as L1 in Sweden lie in the following aspects: the teachers as resources, the students, and lack of suitable teaching material.

First of all, teachers with efficient knowledge in the phonetics and the phonology of both languages (Swedish/Chinese) are very rare. How do the teachers pedagogically apply the new findings of research in their pratice? Does it really help students to acquire the Chinese tone if the teacher stands in the front of the classroom with stretched hands drawing the contours of the tones with gestruers? Is the only way to grasp the Chinese speech to learn Chinese characters first? Even the most intelligent sinologist, speaking fluent Chinese, can not give good suggestions on CSL if they lack linguistic schooling.

It can be assumed that the background of the students' mother tongue (skånska vs. götamål, having different tone accent system) and experiences of foreign languages (mostly at least two, English and another one) might interfere with their acquisition of Chinese sound and tone system. Experience from our music-trained students shows that acquire the tones more easily than others (see also, C.J., 1985). Students who are bilinguals (one dialect of Chinese and Swedish) have an *even worse* situation since the sound systems of Chinese dialects are quiet different from Mandarin and furthermore they may have more tones than the four ones of Mandarin, for instance Cantonese, Hakka, Min, and the Shanghai dialect. It is also important to know the goals of their studies; their motivation will influence their results.

Earlier Chinese teaching material in Swedish (Ahlgren and Löfstedt, 1973; Björkstén and Erlandsson, 2007; Garlén, 1988; Malmqvist, 1979) does not describe Chinese sounds nor the contrastive sound system in the aspects of perception, acoustics, and production.

The education of Chinese language at University of Gothenburg starts with a course of listening skill called Chinese Perception Coaching (CPC). Adult L2 learners can be characterized as having an "accented" perception as well as their accented production (Strange, 1995 p. 22). The present article will discuss the question how teachers could, in their practice, help the students so they get aware what to focus on in L2 using Perceptual Assimilation Model (PAM) (Strange, 1995 p. 193-199).

Goals

The primary aim of CPC is to help students to understand the L2 prosody by comparing with their own mother tongue with as few interpretations from the teacher as possible in order to build greater language awareness. The students get help both individually and in groups before they start the speech production of their new L2. The difficulty is to balance guidance with the happiness of discovery (compare with a child's acquisition of its own native language).

Pedagogics

For groups or whole class

In this stage of CPC contains, in contrast to traditional didactics, lectures in groups/class about

contrastive prosody avoiding sophisticated terminology in order to establish a language consciousness. Each lesson starts with different short actual speech examples collected from internet or from Chinese dialogue parners including different genders and ages, listening to the melody of the new langague, sometimes also to music (even revolutionary songs). They listen carefully without time stress and then report their feelings, experiences, and interpretations. The data discover that the students have capacities to catch some prosodic features for instance basic emotions like anger, fear, sadness, and joy.

A good way to avoid techers' interpretations is to let the students observe Swedish with Chinese accent. They observe that it is easy for Chinese to produce Swedish CV/CVCV words without distinguished short and long vowels. It is still hard for Swedes to understand this kind of words without the typical Swedish tone accent. They also notice that if the Chinese (wo)men use prominently Chinese tones when trying to imitate Swedish prosody it does improve their Swedish pronuanciation. They furthermore perceive that Chinese *ladda* sounds like *latte*. They remark the difference between tones in citation form and actual speech, respectively. They discuss then how citation tonemes relize in intonation. The students notice that it is difficult to perceive the Chinese final nasal [n] since Chinese usually nasalize an ultima like *boken→bokẽ*. The student complain that it is very hard to perceive the vowels before the Chinese retroflex since they are not aware of how Swedish retroflex affects the vowels like m**ödd**a [ø] m**örr**da [œ˞].

For pedagogic sake, on this level the only thing for teachers to do is to emphersize what the students have detected themselves, especially distinguishing phonetic phenomen like *ˊhälsa på*/ *hälsa ˊpå* and *tomten* (accent I)/*tomten* (accent II).

For individuals

Chinese scholars have never separated syllables into consonants (initials), vowels (cores), and tones; these three elements have always been treated as one whole unit (Zhang et al., 1982). It is, however, possible that for some Swedish students, at the beginning of their studies, it is easier to identify sound categories (vowels and consonants, respectively) since that is in accordance with the traditions of Western phonetic/phonologic scholars. Kiriloff (1969) found that tone percption results were much better if the participants concentrated only on the parameter on tones and he therefore advocates this method to be *the* correct one for beginners in Chinese learning, concentrating on tone perception.

Earlier experiences teach us that CPC has to be flexible. Both tone and sound perception training start at the same time; it is free for the students to prioritize. Those who listen to the tones do not write down *pīnyīn* transcription, nor *Level Tone/T1 (ˉ), Raising Tone/T2 (ˊ), Dip Tone/T3 (ˇ), Falling Tone/T4 (ˋ), and Natural Tone/T0 (˙)* but only the numbers 1, 2, 3, 4, and 0. Other students prefer to start with sound analysis. They are asked to write down only the *pīnyīn* without any tone marking. After they have finished listening to the material the students are asked to voluntarily hand in their perception documents.

The method of sound recording is also adapted in CPC. At the start of the semester the teacher got the students' permission to record those who wanted to take part in the coaching. Sometimes the teacher, for nature's sake, recorded without telling the students and asked them to watch what the spectrograms looked like, explaining the differences and letting them experience the differences between the sounds of the two languages.

A portion of the students' data in CPC during the academic years 2008-2011 was selected for analysis of sound categories and discussion about tone perception. The corpus is the actual natural speech and the tests were conducted in the classroom. It appears that some of results are consistent with PAM.

Difficulties with phonemes and tonemes

Generally, the Chinese affricates are the most difficult sounds to acquire, then follow the fricatives and finally come the stops. The stops are the contrastive pairs [p]/[p^h], [t]/[t^h], and [k]/[k^h] respectively and, that is, the two sounds [p]/[p^h] are interpreted as a Swedish [p^h], a phenomenon called Single-Category Assimilations (SC Type). The Chinese fricatives [ɕ], [s], and [ʂ] are close to the Swedish [ç], but not exact, Category-Goodness Difference (CG Type). Individual perception discrepancies are different, some students have a problem only with [ɕ] and [s]. Some misinterpretations occur only in one direction, for instance [ɕ]→[ʂ] but not the opposite way (other examples are [ɕ]→[s]→[ʂ], [s]→[ʂ],

[ɕ]→[ʂ] vice versa or in all possible directions). The Chinese affricates are [tɕ]/[t^hɕ], [ts]/[t^hs], and [tʂ]/[t^hʂ]. The data show that the confusions occur not only within these affricates, there is also a clear tendency for them to be perceived as plain fricatives. PAM can, to some extent, categorize these confusions as different patterns like CG Type, SC Type ect, but where to find the cause of the misidentificaitons? A plausible explanation is that in the Swedish phonological system there is redundance for the Chinese feature [-aspiration] and [-voice] for the stops and also for the Chinese [+apical] and [+apical-velar]. In short, features that L2 utilizes as different phonemes are perceived in the students' as allphones. The student's individual perception capacity can not be the only reason for misinterpretations, also the syllable structures, the context (CV(C), where the C can be either an [n] or an [ŋ] *and nothing else*), lexical/non-lexical status, and their positions in the sentence have great influence.

PAM can be also utilized for tone perception. The situation of tones is more sophisticated than only sound category. Information only upon the [+High], [+Middle], and [+Low] features does not give practical help to the students. It might be suggested that they should compare Chinese tones and a segment of Swedish inotation instead of comparing in first stage only at CV level for instance Chinese **bù** [^{51}pu(:)] with Swedish **bo** [ˊbu(:)]. Let them then discriminate only Chinese tones in citation form. It is impossible for Chinese to image how difficult Swedes perceive these tones and why are they so difficult for them.

Individual tone confusions are very different, like sound categories. Some of them make confusions between T1 and T2, some have problem with T2 and T4. Some have a stable confusion model, or T2 is misidentified as T3, whereas, on the contrary, some have different confusion types. Some change the confusion patterns after a period, some do not. However, they get better result if tone contours (Level, Rising, Dip, and Falling) are introduced.

The data show that both perception and production of the four citation tones on individual level are easily acquired. More than 85% modern Chinese words are disyllabic ones so the situation will be more complicated than only to perceive tones in citation forms. Some of students, however, have the same confusion model even though they listen to the disyllabic words. It appears that tones are perceived stably in actul speech even they are misidentified, which shows that L2 listeners could even perceive the citation tones perform stably on the level of intonation.

The teacher's task does not stop at collecting the data, the important phase is to analyze them scientifically and explain a pedagogic way for the students. But not even this is enough. The teacher should have the capacity to hypothesize not only the expected difficulties of next academic year's students but also build models, for example for confusions of affricatives and tones. The students will not like to come to the lab only for listening to the synthetic sounds, they need to listen to an actual speech.

Research and Practice

Earlier tone research were concentrated on both acoustics and perception with different L1 listeners. Gandour (1978) summarize that according to the common results (a) T2 is the tone that is most difficult identify, (b) the mix-up of T2 and T3 is frequent also for the L1 Chinese, and (c) T1 and T4 are relatively more easily identified. In terms of perception, Yip (2002) states that F0 is not the one and only perceptual cue even though it plays a crucial role. The timing of turning point of F_0 constitutes a salient perceptual cue for discriminating T2 from T3 (Shen and Lin, 1991) and T3 from T4 (Gårding Eva et al., 1986). Gandour (1981, 1984) builds different perceptual dimensions of tones contour, direction, high and so on, which was examed by Lether (1987). It is worth mentioning that all material above consists of sythesitic monosyllabic sounds in citation form.

The data from CPC show that it is evident that first of all tone confusions do not occur randomly, secondly that certain types of tone confusions occur only within the environment of certain consonants and vowels, respectively (Hu & Lindh, 2010). The question for nowadays in CPC at least in contrast with Swedish lies in how a hypothesis should be established.

First of all, CPC model shoul not be only based on tone reseach nor should not like synthetic in ramdom order. The model should be a kind of normal distribution or a regression, under the conditions where different students and various stimulus in diverse linguistic status (lexical, non-lexical, stimuli in different places like narration, interrogation) are presented during a certain study processes. We know that F_0, duration, amplitude, tone contours in both the Chinese and the Swedish prosodic systems are

important, but what more might play roles is unknown today. It is also important to know how perception occurs within a syllable and over a syllabic boarder (in particular, how the vowel and final sound of the previous syllable may affect a Zero initial (initial consonant missing), for instance V/VC, [n]/VC, and [ŋ]/VC, therefore, how and to which extent consonant initials and vowels interfere with tone perception.

References

Ahlgren, I.; Löfstedt, J.-I.: Lärobok i folkkinesika Ahlgren I & Löfstedt J.-I (1973): Lärobok i folkkinesiska.

Björkstén J & Erlandsson A (2007): Kinesiska: Språket i mittens rike

C.J L.W. (1985): Teaching mandarin tones to adult English speakers: Analysis of difficulties with suggested remedies. RELC Journal *16*, 31-47.

Fox & James (1985): The effect of lexical status on the perception of tone Journal of Chinese *13-1* 69-90

Fox & Qi (1990): Context effects in the perception of lexical tome Journal of Chinese Linguistics *18-2* 261-284

Gandour J. (1978). Fromkin: *The perception of tone; in Tone - in A Linguistic Survey* 41-76 Academic Press

Gandour J. (1981): Perceptual dimensions of tone: Evidence from Cantonese Journal of Chinese Linguistics *9-1* 20-36.

Gandour J. (1984): Tone dissimilarity judgments by Chinese listeners Journal of Chinese Linguistics *12-2* 235-261.

Gårding E; Kratochvil P; Jan-Olof, S.; Zhang, J. (1986). Tone 4 and tone 3 discrimination in modern standard Chinese. Language and Speech *29,* 281-293.

Garlén C. (1988): Svenskans fonologi

Ho AT (1976): Mandarin tones in relationship to sentence intonation and grammatical structure Journal of Chinese Linguistics *4-1* 1-13

Hu & Lindh (2010): Perceptual mistakes of Chinese tones in 2-syllable words by Swedish listeners The 4th European Tone and Intonation Conference, Stockholm.

Kiriloff C. (1969): On the auditory perception of tones in mandarin Phonetica *20,* 63-67.

Lether J (1987): F0 pattern inference in the perceptual acquisitions of second language tone. In Leather 59-80

Liang & Heuven (2007): Chinese tone and intonation perceived by L1 and L2 listeners. In: Gussenhoven; Riad 27-61

Malmqvist G (1979): Nykinesisk fonetik

Shen X.S & Lin M. (1991). A perceptual study of Mandarin tones 2 and 3. Language and Speech *1991, 34* 145-156.

Strange, W.(1995): Speech perception and linguistic experience: Issues in cross-language research (York Press, Baltimore).

Yip M.J.W. (2000): Tone Cambridge University Press, Cambridge.

Zhang J., Lü S., & Qi S (1982): A cluster analysis of the perceptual features of Chinese speech sounds Journal of Chinese Linguistics *10-2*, 189-206.

Parent-Child Interaction: Relationship Between Pause Duration and Infant Vocabulary at 18 Months

Dahlby Malin, Irmalm Ludvig, Kytöharju Satu, Wallander Linnea, Zachariassen Helena
Karolinska Institutet, Stockholm, Sweden

Ericsson Anna, Marklund Ulrika
Department of Linguistics, Stockholm University, Stockholm, Sweden

Abstract

Studies of child language development have shown that children from an early age are aware of turn-taking patterns in interaction. The aim of this study is to investigate if there is a relationship between turn-taking pauses in parent-child interaction and child vocabulary at 18 months of age. Analysis of pause duration is conducted on recordings from the SPRINT language intervention project and pause duration is found to correlate with child vocabulary size. Different possible reasons for this correlation are discussed.

Introduction

When speaking to children all adults, regardless of language, use a certain type of modified speech referred to as infant-directed speech (IDS). IDS differs from adult-directed speech (ADS) in several features, such as higher fundamental frequency, greater pitch variation, longer pauses, slower speech rate, shorter utterances (Fernald, Taeschner, Dunn, Papousek, De Boysson-Bardies & Ikuko, 1989) as well as longer vowel duration (Albin & Echols, 1996). This type of speech seems to serve three main functions: to modulate infants' attention and arousal level, to communicate affect to the infant, and to facilitate language acquisition (Fernald et al., 1989).

Acoustic correlates of syntactic boundaries in spoken language may assist infants in perceptually segmenting speech into linguistically relevant units (Seidl, 2007). These acoustic correlates are exaggerated in IDS (Fernald et al., 1989) providing the infant with more information than ADS. Infants rely on prosodic cues when parsing the continuous speech stream first into larger units (clauses), then intermediate units (phrases) and later to smaller units (words) (Jusczyk, 1997). Pauses provide further information in clause segmentation together with prosodic cues, but not by itself (Seidl, 2007). Dutch children have been proven to rely more on pauses than on prosodic cues compared to English children (Johnson & Seidl, 2008). This indicates a potential difference between languages.

For a conversation to run smoothly speakers have to have a mutual understanding of when a speaking turn shift is possible. The basic patterns and rules of turn-taking are perceived and learned by infants at an early prelinguistic stage of language acquisition. Bloom, Russell and Wassenberg (1986) saw that three-month-old infants produced a higher rate of speech-like sounds when interacting with an adult using a clear turn-taking pattern. In conversation speakers tend to match the length of pauses both in and between speaking turns, and when this matching occurs the speaker is perceived as warm and affective. This phenomenon occurs in mother-infant conversations with prelinguistic infants as young as four months (Bebee, Alson, Jaffe, Feldstein & Crown, 1988) and is also seen in gaze behavior of six-week-olds in interaction with adults (Crown, Feldstein, Jasnow, Beebe & Jaffe, 2002) providing further evidence for early understanding of turn-taking.

A productive vocabulary of 50 words is considered an important milestone in language development (Tamis-Lemonda, Bornstein, Kahana-Kalman, Baumwell & Cyphers, 1998), usually achieved during the second year of life. Several things have been shown to predict infant vocabulary development, including parental communicative behavior. A responsive parenting style has been argued to contribute more to lexical development than a directive

style (Masur, Flynn & Eichorst, 2005). It is possible that different parenting styles manifests in different pause behavior.

In light of previous studies, it seems lexical development is facilitated by certain type of parental communicative behavior and that infants have an early awareness of turn-taking pauses. This study aims to investigate if there is a relationship between turn-taking pauses and infant lexical development. The ongoing SPRINT project[1] makes this possible by providing audio recordings of parents interacting with their 18 months old infants.

Method

In the SPRINT project families are given intervention films to promote parent-child interaction and stimulate child language acquisition. The children's language skill is classified at the age of 18 months using SECDI, a Swedish adaptation of the CDI (Child Development Inventory). In this questionnaire which has been classed as a good predictor of child language development (Berglund & Eriksson, 2000), parents rate their children's vocabulary. To analyze turn-taking in parent-child interaction, this study used audio recordings from 10 families in the SPRINT project.

Participants

Based on the SECDI results five low performing and five typical performing children were selected. The low performers (LP) had SECDI results in the 0-25 percentile and the typical performers (TP) in the 50-65 percentile. Both groups consisted of two girls and three boys. The children were 18 months old when recorded. At least one parent in each family had Swedish as their first language.

Audio recordings

The material consisted of recordings made by the parents, before the SPRINT intervention started. They were asked to record four common situations: clothing, playing, reading and cooking/eating. No other specific instructions were given. Of the 10 children, 6 had one recording of each requested situation while the other 4 had two recordings of the same or of other situations (e.g. bathing). In the recordings the child as well as at least one parent participated and some included more participants (siblings and other adults).

Procedure

The files were tagged using the transcription function in the speech analysis program Wavesurfer 18, version 8.4.2.9. The beginning and ending of every communicative utterance produced was tagged, using a separate transcription pane for each participant. Five minutes per recording were tagged, starting from the first parental child directed utterance. One of the files was less than five minutes long but was nevertheless included. Whether an utterance was communicative or not was subjectively judged by the researchers based on situation and content of the utterance. Utterances also included cooing, babbling, screaming and other non-lexical communicative behaviour. Beginning and ending of each utterance was judged both auditory and visually using the spectrogram function in Wavesurfer with an estimated error margin of approximately 50ms. The resulting files were processed using Mathematica 7.0 to extract the pauses between utterances. A pause was only registered by the program if no one spoke in the silent gap left between utterances. The content and directions of the utterances were not considered. In recordings with more than one parent and/or child, intrusions by other participants may have resulted in no pause being registered. The pauses were then statistically analyzed in PASW Statistics 18. The pauses used for analysis were parent-parent, parent-child and child-parent. All other pauses (e.g. mother-father, sibling-parent, child-child) were not considered relevant for this study. Pauses with a duration greater than 3 seconds were excluded since they were too long to be considered as turn-taking pauses. The same limit was used in Kondaurova and Bergeson (2010).

Results

The duration of pauses between child and parent (cp), parent and child (pc) and the pauses after a parental utterance until the next parental utterance from the same parent (pp) were computed within the two groups (low performers, LP, and typical performers, TP). In

[1] SPRINT is an ongoing project funded by the Swedish Research Council (2008-5094) "Effects of enhanced parental input on young children's vocabulary development and subsequent literacy development" – a collaboration between the department of Linguistics and department of Special Education at Stockholm University.

figure 1 the mean difference in pause duration is shown with a 95 % confidence interval.

Univariate Analysis of Variance (ANOVA) and LSD Post hoc-test was conducted on the six groups (LPcp, TPcp, LPpc, TPpc, LPpp, TPpp) to see whether there was significant differences between LP and TP.

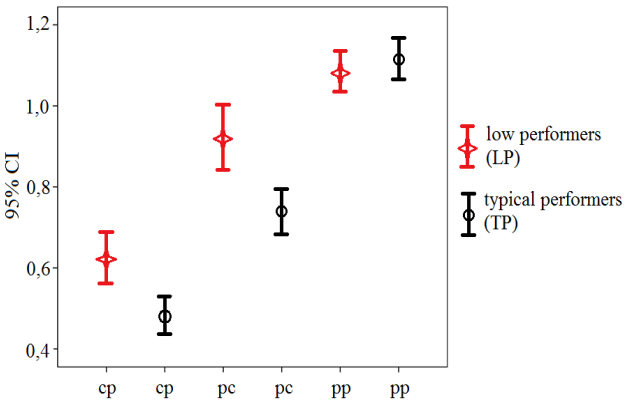


Figure 1. Pause duration: 95 % confidence interval of mean values. The y-axis shows duration in seconds.

Child-parent pauses (cp)

The LPcp mean was 0.63 s (sd: 0.59) for the whole group, within the group the individual cp mean varied from 0.47 to 0.76 s. The TPcp mean was 0.48 s (sd: 0.53) with individual cp means between 0.29 and 0.70 s. Thus there is a big within-group difference in both groups. One LP-child had a lower individual cp mean than the total mean for TPcp, and one child from the TP group had a higher cp mean than the total mean for LPcp. The total amount of LPcp-pauses shorter than three second was 330, and 480 for TPcp. The lengths of LPcp and TPcp differ from each other significantly ($p<0.001$). The amount of excluded pauses, longer than 3 seconds, was 27 for LPcp and 13 for TPcp.

Parent-child pauses (pc)

The LPpc mean was 0.92 s (sd: 0.73), lowest individual pc mean in the LP group 0.85 s and highest 1.00 s. The TPpc mean was 0.74 s (sd: 0.62) with individual pc means between 0.66 and 1.09 s. Thus the total mean was highest in the LP group but the individual highest pc mean was in the TP group. The total amount of LPpc-pauses shorter than 3 seconds was 320, for TPpc it was 488. The difference in duration between LPpc and TPpc is significant ($p=0.003$). The amount of excluded pauses, longer than 3 seconds, was 56 for LPpc and 34 for TPpc.

Parent-parent pauses (pp)

The LPpp mean was 1.09 s (sd: 0.69) with 746 pauses shorter than 3 seconds, and the TPpp mean was 1.12 s (sd: 0.71) with totally 765 pauses shorter than 3 seconds. The within-group-variation of individual means was low. LPpp and TPpp do not significantly differ. The amount of excluded pauses, longer than 3 seconds, was 153 for LPpp and 91 for TPpp.

Discussion

The results show that the children's SECDI results and the duration of the turn-taking pauses correlate. There is a significant difference between LP and TP regarding parent-to-child and child-to-parent pause duration. The LP turn- taking-pattern is slower in both cases, longer and fewer pauses, which implies fewer utterances as well as less time spent speaking. This is also shown in the number of excluded pauses (longer than 3 seconds): more in LP and fewer in TP. If the parents' do not give feedback fast enough and spend too little time interacting verbally with their children it might affect the vocabulary development of the children. On the other hand it is possible that the parents of LP answer slower because they adapt to the child's slow turn-taking behavior. This is consistent with the observations that speakers match pause and utterance length to those of their conversation partner (Bebee et al., 1988). The fact that intrapersonal parent-to-parent pause length does not differ between the two groups, and the mean pause length is significantly longer for pp than for pc pauses in both groups shows that the parents in both groups give their young ones enough time to speak, i.e. time to take over the turn, before starting to talk again. This indicates that the fact that the pauses are longer in the LP group is a parental adaptation to their children's behavior. This could also explain the shorter TP pause duration, since those children have advanced further in the speech development and according to Soderstrom (2007) IDS is adjusted to the child's language competence. These interpretations do not necessarily contradict each other since it is likely that heredity and environment interact in child language development.

The results might have been affected by methodological aspects such as the decision not to count silences longer than 3 seconds as turn-taking pauses. This was based on the method used in Kondaurova and Bergeson (2010). The

limit of three seconds is fairly high set, and acoustic analysis of the material in the current study verified that a turn after 3 seconds rarely belonged to the previous turn-taking sequence.

Since the recordings were made in the homes of the participating families, the lack of control over the recording situations may also have affected the results. There was a broad diversity of occurrences that could not be controlled in the recordings. For example one child slept through the first two minutes of the recording session, and another child was ill. This of course limited the number of child utterances. In some recording with more than one child it was hard to discriminate both between the children's sounds and the direction of the parental utterances. Greater control of situations would have been gained by only using recordings with one child and one parent making it easier to identify turn-taking. Another way to analyze the material would have been to only use child-parent interaction pauses. By looking at the content of each utterance it would have been possible to exclude pauses that were not part of a turn-taking sequence (e. g. parent talking to sibling and child making noises outside the conversation). However to do this a lot of subjective judgments of the utterances' content would have been necessary, which would decrease the objectivity.

The fact that the SECDI scores were used for dividing children into groups may also affect the results because basing the groups on parental assessment alone may open up for disparity between reported performance and actual ability. With such a small sample as 5 individuals per group one single misjudgment can affect the results even though SECDI has been proven to be a reliable predictor of child language development (Berglund & Eriksson, 2000).

The results of this study show an evident relation between pause behavior and child language development. To achieve a higher grade of generalization of the results it might be necessary to study a greater sample with regards to parent child pause behavior. It would also be of interest to further investigate cause and consequence in the observed relation. Longer pauses is only one of many things that characterizes IDS, and to fully understand the impact pause duration has on child language acquisition one would have to look at the other aspects of IDS in combination with pauses. This could be an area for further research.

Acknowledgements

We would like to thank Ellen Marklund at Stockholm University for mathematical and statistical assistance.

References

Albin D D and Echols C H, (1996). Stressed and word-final syllables in infant-directed speech. *Infant Behavior and Development*, 19, 401-418.

Beebe B, Alson D, Jaffe J, Feldstein S and Crown C (1988). Vocal congruence in mother-infant play. *Journal of psycholinguistics*, 17, 245-259.

Berglund E and Eriksson M (2000). Reliability and content validity of a new instrument for assessment of communicative skills and language abilities in young Swedish children. *Logoped Phoniatr Vocol*, 25, 176-85

Bloom K, Russell A and Wassenberg K, (1987). Turn taking affects the quality of infant vocalizations. *Journal of Child Language*, 14, 211-227.

Crown C L, Feldstein S, Jasnow M D., Beebe B and Jaffe J, (2002). The cross-modal coordination of interpersonal timing: Six-week-olds infants' gaze with adults' vocal behaviour. *Journal of Psycholinguistic Research*, 31, 1-23.

Fernald A, Taeschner T, Dunn J, Papousek M, De Boysson-Bardies B and Ikuko F (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16, 477-501.

Johnson E K and Seidl A, (2008). Clause Segmentation by 6-Month-Old Infants: A Crosslinguistic Perspective. *Infancy*, 13, 440-455.

Jusczyk P W (1997). *The discovery of spoken language.* Cambridge, MA, USA: MIT press.

Kondaurova M V and Bergeson T R (2010). The effects of age and infant hearing status on maternal use of prosodic cues for clause boundaries in speech. *JSLHR Papers in Press*, Published October 21.

Masur E F, Flynn V and Eichorst D L (2005). Maternal responsive and directive behaviours and utterances as predictors of children's lexical development. *Journal of Child Language*, 32, 63-91.

Seidl A (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language*, 57, 24-48.

Soderstrom M (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27, 501-532.

Tamis-Lemonda C S, Bornstein M H, Kahana-Kalman R, Baumwell L and Cyphers L (1998). Predicting variation in the timing of language milestones in the second year: an events history approach. *Journal of Child Language*, 25, 675-700.

Effects of a film-based parental intervention on vocabulary development in toddlers aged 18-21 months

Donya Afsun, Erika Forsman, Cecilia Halvarsson, Emma Jonsson,
Linda Malmgren and Juliana Neves
Speech and Language Pathology Program, Karolinska Institute
Ulrika Marklund
Stockholm University

Abstract

SPRINT is a language intervention project aimed to study the outcome of a parental home training program on children's language development and future reading and writing skills. This study, which derives data from the SPRINT project, intended to examine the possible effects of a parental-based film intervention. It was conducted on toddlers aged 18-21 months from the Stockholm area with at least one parent who has Swedish as a first language. Parents of 78 children participated in the study and filled in 3 SECDI-w&s questionnaires rating their children's productive vocabulary. Children were randomized to either the intervention or the control group. Results indicated that the intervention group demonstrated significantly higher scores over time, $F(2,78) = 5{,}192$, $p < .007$. In the light of previous research it is concluded that this intervention contributes to an increase in productive vocabulary. However, the scores of the intervention group did not exceed the average range for Swedish children in the same age span. Furthermore the possible impact of parental education and the presence of siblings on productive vocabulary was discussed.

Key words: film intervention, vocabulary development, SECDI-w&s, early language intervention, toddler language, SPRINT

Earlier studies have shown that the development of productive vocabulary size may differ with respect to rate and content. Some children might experience a so called "vocabulary spurt" whereas other children experience a more gradual development of lexical growth (Goldfield & Reznick, 1990). First-born children are more likely to experience the "naming explosion". Later-born children may not receive the same clear, focused language input as their older siblings, as they receive language input to a greater extent from various sources such as older brothers and sisters (Goldfield & Reznick, 1990). It is known that parental education level has been associated with faster vocabulary development. In fact, a study by D'Odorico et al. (2001) shows that mother's education explains 20% of the variance in vocabulary size.

A wide vocabulary is essential for comprehension with regard to both oral language and reading. A study by Marulis and Neuman (2010) has shown that a broad vocabulary provides a better prospect to an early reading development. Furthermore, research has demonstrated that an increase in productive vocabulary also improves structural complexity and the use of multiword expressions (Girolametto & Pearce, 1996). A study from Oller (2010) looks into the subject of what role directed input of language may present on vocabulary learning as opposed to indirect language input such as overheard conversations. The author's hypothesis was verified when the results indicated that direct language input had a significantly greater positive effect on vocabulary learning than indirect.

According to the same study by Marulis and Neuman (2010), it is known that even small amounts of intervention have positive effects on vocabulary. The study enlightens the importance of promoting vocabulary in the early childhood years. In addition, earlier studies of different types of interventions have led to conclusions of a positive outcome both in language in general and in vocabulary specifically (van Balkom et al., 2010, Mendelsohn et al., 2001). Intervention seems to produce positive effects on language whether infants present developmental language delay (van Balkom et al., 2010) or not (Mendelsohn et al., 2001). On the other hand,

one study (Pile et al., 2010) showed that shared book reading intervention did not affect the vocabulary in participating children, who were aged between 4 and 5 years and had language impairment.

The current study

The current study focuses on productive vocabulary in toddlers aged 18 to 21 months who participated in an intervention program with their parents. Typical development of vocabulary in this age span will be discussed so as to investigate whether vocabulary scores may or may not be attributed to parental intervention. Due to previous research it is expected that the intervention group will present better results in terms of productive vocabulary. Since normative data for typical productive vocabulary for Swedish children aged 16 to 28 months is available using the same measurement tools ([1]Berglund & Eriksson, 2000), a comparison will be possible. If a positive correlation between the intervention group and larger productive vocabulary is attained, the intervention films could serve as important tools in the promotion of skills to improve vocabulary for all children in the future.

In the light of previous research, the present study aims to investigate in which way productive vocabulary size develops over time. In specific, the study focuses on whether or not a parental-based intervention program causes an effect on infant's productive vocabulary.

Method

The present study uses data from SPRINT, which is a language intervention project that aims to study the effects of a parental home training program on children's language development and future reading and writing skills.

Participants

The families in the SPRINT-project were selected from population registers from Stockholm. Data from 78 families was analyzed. Participants were randomly assigned to each of the groups, 40 to the intervention group and 38 to the control group. The criterion for eligibility in this study was that at least one of the parents considered their first language to be Swedish. The children were born between August and November 2008. The parental education level in the control group was 76% for higher education, 16% high school, 1% elementary school, 4% others and 3% omitted answers. As for the intervention group the parental education level was 78% for higher education, 8% high school, 1% elementary school, 12% others and 1% omitted answers.

Materials

Vocabulary size was measured with the Swedish-adapted version of the MacArthur Communicative Development Inventories (CDI), namely the Swedish Early Communicative Development Inventories words and sentences (SECDI2). The SECDI2 questionnaire consists of two components: "words children use" and "sentences and grammar". In this study only data from "words children use" was included, since it is the only part that assesses single word production.

Outline of intervention

The intervention families were able to access educational films online. The films contain samples of parent-child interaction judged to stimulate language acquisition recorded in family home environments. They consist of examples of how parental behavior in interacting with a child could stimulate the child's language acquisition. The family sequences are then discussed and assessed by two SPRINT researchers. Lectures by the researchers were also available concerning scientific findings on language acquisition and the importance of interacting with, and reading to the child.

The film material raises the concept of a pedagogic model by SPRINT, containing three areas presumed to be important for children's language development: focusing on the child, turn-taking and adding language.

Procedure

The intervention started when the children were 18 months old. The parents were instructed to fill in which words in the SECDI questionnaire they had actually heard their child produce, regardless of deviating pronunciation. SECDI was filled in three times during the course of this study, at the ages of 17-18, 18-19 and 20-21 months.

Processing of data

The data was analyzed in the SPSS software. The included variables in the analysis were total scores of the three SECDI questionnaires and intervention status.

Results

The SPSS analysis showed a significant interaction between the linear trend along the three different occasions and groups (Fig. 2). The general linear model showed a significant effect of occasion, $F(2,78) = 68{,}671$, $p < .0005$, and a significant difference between groups and occasions, $F(2,78) = 5{,}192$, $p < .007$, sphericity assumed.

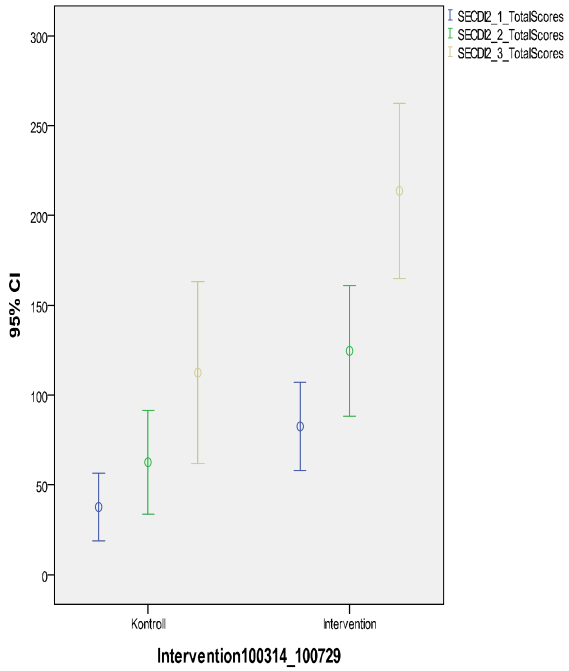


Figure 1: Productive vocabulary, range and means for intervention group and control group before, under and after intervention (CI 95%).

An increase in productive vocabulary mean value was attained by both groups with every measuring occasion (Tab. 1). The intervention group achieved significantly higher scores in every occasion (17-18 months $t(76) = 2{,}990$, $p < .004$; 18-19 months $t(76) = 2{,}681$, $p < .009$; 20-21 months $t(76) = 3{,}034$, $p < .003$).

Table 1: Mean values of productive vocabulary for the control and intervention groups during the three measuring occasions. Standard deviation for every group and occasion in parenthesis.

Mean number words	17-18 months	18-19 months	20-21 of months
Intervention Group (N40)	82 (sd:77)	125 (sd:113)	214 (sd:152)
Control Group (N38)	37 (sd:56)	63 (sd:86)	109 (sd:150)

Discussion

The present study aimed to investigate the effects of a parent-based intervention on children's productive vocabulary as well as in which way productive vocabulary develops over time. According to van Balkom et al., (2010), parent-targeted video intervention produces positive effects on children's productive vocabulary. The results of the current study support these findings. Figure 1 shows that the intervention group demonstrated significantly higher scores over time which indicates a higher development.

It is interesting to note that this increase was attained over a period of three months, which is a relatively short period of time. The results of the intervention group support the conclusions from Marulis and Neuman (2010) which state that even small amounts of language intervention have positive effects on vocabulary. A probable reason for this outcome is that the children likely received greater amounts of directed language input than indirect, which goes along the lines with Oller's study (2010) on the effect of direct versus indirect language input. The validity of the theory relies on the assumption that the parents applied the pedagogic model and general advice from the film material.

However, mean values of productive vocabulary in this study for both groups are within the normal range for Swedish children in all three age groups, as established by [1]Berglund and Eriksson (2000). There is a clear tendency towards an increase in productive vocabulary over time regardless of intervention status. A possible explanation for the high rates of productive vocabulary can be retrieved from Goldfield and Reznick (1990) in that there may have been children who were experiencing a "vocabulary spurt".

Potential confounding variables to be taken in consideration are the presence of older siblings as well as parental education. According to Goldfield and Reznick (1990) children with older siblings receive a less clear and focused

language input since it is derived to a large extent from various sources. However, in the present study this variable was not analyzed.

When it comes to parental education, our sample is biased towards higher education levels. According to Statistics Sweden (SCB) 32% of the Swedish population has a higher education level, whereas in this study's sample the same education level was in average 77%. Since parental education has been shown to have an impact on vocabulary development (D'Odorico et al., 2001) conclusions can be drawn that the children in this sample were subjected to a more favorable language environment which may have influenced the positive outcome in the current study. The data collection method used in this study, parent reports, can lead to biased answers which in turn may contribute to misleading results.

A possible area of research could be to prolong the intervention period in order to decrease the impact of "vocabulary spurt" on the increasing rate of productive vocabulary. If the results of the current study prove to be consistent with further research, the intervention material could be used to enhance vocabulary development. Even though SECDI is not convenient for clinical purposes and is mainly used for research, it can be used as an important tool to determine normative data that later can be applied to clinical practice ([2]Berglund & Eriksson, 2000).

Although the scores of the intervention did not surpass the average range for Swedish toddlers aged 18-21 months, in this case the intervention contributes to an increase in productive vocabulary. Further research on this topic might benefit from a follow-up on the early reading and writing development of the intervention group.

References

[1]Berglund, E, & Eriksson, M (2000) Communicative development in Swedish children 16-28 months old: The Swedish early communicative development inventory- words and sentences. *Scandinavian Journal of Psychology*, 41, 133-144.

[2]Berglund, E & Eriksson, M (2000) Reliability and content validity of a new instrument for assessment of communicative skills and language abilities in young Swedish children. *Logopedics Phoniatrics Vocology*, 25, 176-185.

D´Odorico, L, Carubbi, S, Salerni, N & Calvo, V (2001) Vocabulary development in Italian children: a longitudinal evaluation of quantitative and qualitative aspects. *Journal of Child Language*, 28, 351-372.

Girolametto, L, & Pearce, P S (1996) Interactive focused stimulation for toddlers with expressive vocabulary delays. *Journal of Speech & Hearing Research,* 39 (6), 1274.

Goldfield, B A, & Reznick, J S (1990) Early lexical acquisition: rate, content, and the vocabulary spurt. *Journal of Child Language,* 17, 171-183.

Marulis, L M, & Neuman, S B (2010) The effects of vocabulary intervention on young children's word learning: a meta-analysis. *Review of Educational Research*, 80, (3), 300-335.

Mendelsohn, A L, Mogilner, L N, Dreyer, B P, Forman, J A, Weinstein, S C, Broderick, M, et al. (2001) The impact of a clinic-based intervention on language development in inner-city preschool children. *Pediatrics*, 107, 130-134.

Oller, K (2010) All-day recordings to investigate vocabulary development: a case study of a trilingual toddler. *Communication Disorders Quarterly*, 2010 31:213.

Pile, E J S, Girolametto, L, Johnson, C J, Chen, X, & Cleave, P L (2010) Shared book reading intervention for children with language impairment: using parents-as-aides in language intervention. *Canadian Journal of Speech-Language Pathology and Audiology,* 34, (2), 96-109.

Van Balkom, H, Verhoeven, L, van Weerdenburg, M, & Stoep, J (2010) Effects of parents-based video home training in children with developmental language delay. *Child Language Teaching and Therapy*, 26 (3), 221-237.

Productive Vocabulary Size Development in Children Aged 18-24 Months – Gender Differences

Ida Andersson[1], Jenny Gauding[1], Anna Graca[1], Katarina Holm[1], Linda Öhlin [1]
Ulrika Marklund[2], Anna Ericsson[2]
[1]Karolinska Institutet, The names of the authors are in alphabetical order.
[2]Department of Linguistics, Stockholm University

Abstract

Several studies have shown slight differences in language skills between genders, favouring females. In order to investigate gender differences in speech production for Swedish children, the productive vocabulary size of 295 children, aged 18-24 months, was measured by the validated instrument SECDI-2. The size of the productive vocabulary was found to grow rapidly during this age. Significant gender differences were found at 21 and 24 months, but not at 18 months. The girls' mean scores were higher.

Introduction

Early vocabulary development

Most children with normal development follow a similar pattern in language acquisition. At the end of the first year the child has phonetic recognition and a memory of concepts for objects, people and actions of the surroundings, which are necessary conditions for developing words. Closely before or after the first birthday the first word appears (e.g. Strömqvist, 2008).

The growth rate of the expressive vocabulary changes from slow to more rapid on majority of children at the end of the second year. This period of rapid word-learning, often referred to as the vocabulary spurt, could be characterised as a "naming explosion" (Goldfield & Reznick, 1990). The largest category of words in this period of time is nouns. Children with small vocabularies show a high percentage of social terms, such as onomatopoetic words, routines and naming of people (Stolt, Haataja, Lapinleimu & Lehtonen, 2008). Some children, especially those with older siblings, develop their expressive vocabulary at a more gradual pace instead of going through a vocabulary spurt. These children maintain a steady balance of nouns and other kind of words throughout their early lexical development (Goldfield & Reznick, 1990).

Expressive vocabulary size seems to be related to general language skills at the end of the second year (Stolt et al., 2008). There is also a strong correlation between expressive vocabulary size and grammar skills during the end of the second year and forward (Berglund & Eriksson, 2000b).

Gender differences

Several studies have shown slight differences in language skills between genders. A meta-analysis by Hyde and Linn (1988), including 165 studies on children aged 3 to adult, reported data on gender differences in various verbal abilities. Their result indicated only a slight female superiority in performance. Large differences across multiple specific and general measures of language (including expressive vocabulary) were found between genders in the second through fifth year, but not before or after (Bornstein, Hahn & Haynes, 2004). The same study found that individual differences in children showed moderate to strong stability; girls and boys alike were stable. Gender-typed differences have been found to be apparent in communicative actions at 13 and 18 months by Stennes, Burch, Sen and Bauer (2005). They interpreted their results as some of the earliest evidence of gender-based schematic processing. Westerlund and Lagerberg (2007) found that girls had a more developed vocabulary and were more involved in reading than boys at 18 months.

A study of over 3000 2-year-old twin pairs that examined genetic and environmental origins showed, like other studies, that girls scored

higher on verbal ability, measured by productive vocabulary (Galsworthy, Dionne, Dale & Plomin, 2000). More importantly, boys showed greater heritability than girls. This study concluded that the influence of factors correlate differently for male and female verbal development. Henrichs et al. (2010) found at both 18 and 30 months that boys were more likely to be delayed in expressive vocabulary skills than girls, but the gender difference's contribution to the overall variance was small. These findings are consistent with the knowledge that early language delay is highly heritable and that boys are more prone to language problems (Dale et al., 1998).

In a study on birth order Bornstein, Leach and Haynes (2004) found that at 20 months first born girls outperformed boys on all vocabulary competence measures, and second born girls outperformed boys on vocabulary comprehension and vocabulary production. Berglund, Eriksson and Westerlund (2005) also found that the effect of gender was larger than the effect of birth order. Though, they see no need for different praxis for girls and boys, since the difference is corresponding to only about one word. In Hyde and Linn's meta-analysis (1988) the conclusion was that "the magnitude of the gender differences in verbal ability is currently so small that it effectively can be considered to be zero". The one possible exception, according to Hyde and Linn, was measures of speech production which favoured females.

Aim of the study

The aim of this study was to investigate gender differences in speech production of Swedish children. The Swedish Early Communicative Development Inventory (Berglund & Eriksson, 2000a) was used to measure productive vocabulary. The research questions were: How does productive vocabulary size develop from 18 to 24 months of age? Are there any gender differences?

Method

Participants

This study drew its data from the SPRINT project, which is an ongoing Swedish prospective language study in which parents are invited to take part in an intervention program intended to actively support children's communicative development. The children are followed from the age of 12 months. We used data collected at approximately 18, 21 and 24 months (±1 month) from the randomly selected control group, as well as the two groups that had not yet started their intervention. The sample consisted of children who were born during a 12-week period in 2008 (August-November). They all lived in the Stockholm area. We excluded children who were adopted or did not have Swedish as their first language, to avoid confounding variables. The final sample consisted of 295 children. However, all children did not participate at each assessment. The number of children and the proportion of boys and girls at every assessment are shown in table 1.

Table 1. Participants: age and gender

	Age (months)			
Gender	18	21	24	Total
Girls (n)	123	112	92	137
Boys (n)	129	136	101	158
Total (n)	252	248	193	295

Material and procedure

A Swedish online version of the MacArthur Communicative Development Inventory, *The Swedish Early Communicative Development Inventory – Words and Sentences* (SECDI-2), was used to collect data on productive vocabulary. Good reliability for the Swedish version has been demonstrated (Berglund & Eriksson, 2000a). SECDI-2 consists of two parts. Part I includes A - a vocabulary checklist that consists of 710 items presented in 21 semantic categories, B - feedback morphemes, 10 items, C - pragmatic skills. Part II is a section for sentences and grammar. Only part I A and B was used here (max score 720). The parents filled in the SECDI-2 when the children were approximately 18, 21 and 24 months old.

PASW Statistics 18 was used for statistical analysis: descriptive statistics (M, SD, range, and percentiles) and one-way ANOVAs of productive vocabulary size and gender.

Table 2. Number of words: means, SDs and range for boys, girls and total sample at different ages

	Boys			Girls			Total	
Age (months)	M	SD	Range	M	SD	Range	M	SD
18	85	98	3 - 448	95	82	2 - 423	90	91
21	191	158	4 - 651	231	140	9 - 540	209	152
24	317	177	6 - 681	417	160	20 - 658	365	176

Result

The mean number of spoken words showed a steadily rising curve by increasing age. The general range is larger in boys. Means, SDs, and range are shown in table 2.

The percentile scores for the total of all words are shown for boys in figure 1 and girls in figure 2. The boys in the lowest percentiles had a small increase between 18 and 21 months compared to the girls. Between 21 and 24 months both boys and girls had a larger increase in the lowest percentiles, but the change was more apparent for the girls.

Table 3 shows a one-way ANOVA of gender differences in vocabulary size which showed significance at 21 and 24 months, explaining 1.7% respectively 8.1% of the total variance. The girls' mean scores were higher.

Table 3.Results of the one-way ANOVAs made to measure gender differences at different ages.

	F	P	Df	Eta^2
18 months	0.776	n.s.	251	(0.003)
21 months	4.328	0.039	247	0.017
24 months	16.781	<0.001	192	0.081

Discussion

This study has focused on the productive vocabulary size development and gender differences in Swedish children aged 18-24 months. Productive vocabulary size was found to grow rapidly during this age. Significant gender differences were found at 21 and 24 months, but not at 18 months.

Due to a server error a paper version of the SECDI-2 was sent out to some families at 21 months. This complication could have contributed to a loss of data seen at 24 months. Another methodological consideration is the use of parental reports. Self-report bias cannot be excluded; however, the SECDI is a validated instrument.

The curves of the lowest percentiles (figure 1 and 2) become steeper at around 21 months, and could indicate the start of a vocabulary spurt, which tends to occur around the end of the second year according to Goldfield and Reznick (1990). According to Strömqvist (2008), the vocabulary spurt starts somewhere between 1 and 3 years. The higher percentiles show a steeper curve already by 18 months. These children might already have started their vocabulary spurt. It is worth noting that 49 % of the chil-

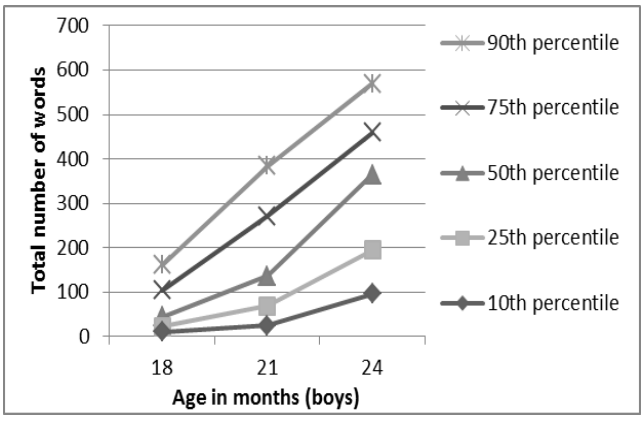


Figure 1. Productive vocabulary size for boys aged 18, 21, and 24 months. Median values, and 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile.

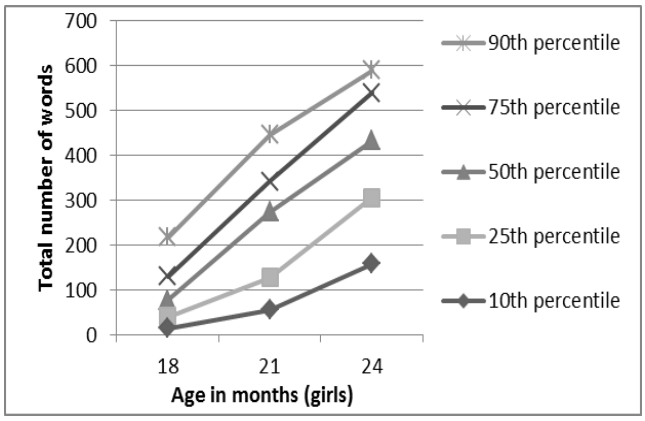


Figure 2. Productive vocabulary size for girls aged 18, 21, and 24 months. Median values, and 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentile.

dren in our sample had at least one older sibling, which can have an effect on the pace of the early vocabulary development (Goldfield & Reznick, 1990). Another possible confounding factor could be that 21% had more than one language in their home environment. Since the distribution of girls and boys was even in both cases it has probably not resulted in further gender differences.

We got no significant result from the ANOVA concerning gender differences at 18 months. Several other studies have found gender differences at this age (e.g. Berglund et al., 2005; Stolt et al., 2008; Westerlund & Lagerberg, 2008). However, Berglund et al., argue that their result is so small that it is negligible. We argue the same thing for our findings at 21 months, where we found a significant difference explaining 1.7% of the variance. Other studies have found larger explained variance for similarly aged children (20 months), e.g. Bornstein, Hahn and Haynes (2004). At 24 months, though, the result explains 8.1% of the variance which we argue can no longer be considered negligible. A difference at this age was also found by Galsworthy et al. (2000) and Bornstein, Hahn and Haynes (2004). Stolt et al. (2008) did not find a significant difference, however they question the effect of their small sample.

A part of our aim was to investigate gender differences in the vocabulary of young Swedish children. We conclude that there are small differences appearing at around the end of the second year. Girls have a more even distribution and boys have a wider range. The highest and lowest scores are generally held by boys, while girls have the highest mean score. It could depend on how genetic and environmental factors correlate differently for girls' and boys' early verbal development (Galsworthy et al., 2000). The lowest score for boys could be a consequence of them having a greater heredity for language delay (Dale et al., 1998). We cannot exclude though, that different expectations on boys and girls, resulting in different treatment, has a potential effect on verbal ability. It could be interesting in further studies to see what happens with the difference at later ages. Another possible further study could be to compare high performing girls with high performing boys, as well as compare low performing girls and boys. One could also study the gender composition in the highest and lowest percentiles.

References

Berglund E & Eriksson M (2000a). Reliability and content validity of a new instrument for assessment of communicative skills and language abilities in young Swedish children. *Logopedics Phoniatrics Vocology, 25,* 176-185.

Berglund E & Eriksson M (2000b). Communicative development in Swedish children 16-28 months old: The Swedish early communicative development inventory – words and sentences. *Scandinavian Journal of Psychology, 41,* 133-144.

Berglund E, Eriksson M & Westerlund M (2005). Development and Aging: Communicative skills in relation to gender, birth order, childcare and socioeconomic status in 18-month-old children. *Scandinavian Journal of Psychology, 46,* 485-491.

Bornstein M H, Hahn C-S & Haynes O M (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language, 24,* 276.

Bornstein M H, Leach D B & Haynes O M (2004). Vocabulary competence in first- and secondborn siblings of the same chronological age. *Journal of Child Language, 31,* 855-873.

Dale P S, Simonoff E, Bishop D V M, Eley T C, Oliver B, Price T S, Purcell S, Stevenson J & Plomin R (1998). Genetic influence on language delay in 2-year-olds. *Nature Neuroscience, 1,* 324-328.

Galsworthy M J, Dionne G, Dale P S & Plomin R (2000). Sex differences in early verbal and nonverbal cognitive development. *Developmental Science, 3,* (2), 206-215.

Goldfield B A & Reznick J S (1990). Early lexical acquisition: rate, content, and the vocabulary spurt. *Journal of Child Language, 17,* 171-183.

Heinrichs J, Rescorla L, Schenk J J, Schmidt H G, Jaddoe V W V, Hofman A, Raat H, Verhulst F C & Tiemeier H (2010). Examining continuity of early expressive vocabulary development: The generation R Study. /Electronic version/. *Journal of Speech, Language, and Hearing Research,* doi:10.1044/1092-4388(2010/09-0255)

Hyde J & Linn M (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104,* 53-69.

Stennes L, Burch M, Sen M & Bauer P (2005). A longitudinal study of gendered vocabulary and communicative action in young children. *Developmental Psychology, 41,* (1), 75-88.

Stolt S, Haataja L, Lapinleimu H & Lehtonen L (2008). Early lexical development of Finnish children: A longitudinal study. *First language 28,* (3), 259-279.

Strömqvist S (2008). Barns språkutveckling. In: L. Hartelius, U. Nettelbladt & B. Hammarberg, eds, *Logopedi*. Sweden: Studentlitteratur, 69-83.

Westerlund, M. & Lagerberg, D. (2008). Expressive vocabulary in 18-month-old children in relation to demographic factors, mother and child characteristics, communication style and shared reading. *Child: care, health and development, 34*, (2), 257–266.

Phonetic markedness, turning points, and anticipatory attention

Mikael Roll, Pelle Söderström, and Merle Horne
Department of Linguistics and Phonetics, Lund University

Abstract

Phonetic markedness regarding linguistically relevant tonal patterns (Accent 2, boundary tones) in Central Swedish is discussed. Both tonal markedness and F0 turning points are assumed to be important cues for anticipatory attention to grammatical structure during speech processing. Empirical evidence from neurolinguistic and psycholinguistic experiments for the assumed relation between anticipatory attention and marked tonal patterns' association with Swedish word and clause structures is presented.

Introduction

The present article reviews results from neurolinguistic and psycholinguistic studies on the perception of Central Swedish grammatically relevant tonal patterns and relates these results to the notions of 'phonetic markedness' and 'passive attention'. We suggest that marked tonal patterns tend to be associated with relatively more kinds of grammatical structure as compared to their unmarked counterparts due to their intrinsic tendency to allocate attention. For example, Central Swedish Accent 2 is associated with more lexical information than Accent 1. On the one hand, it is the word accent associated with compounds, and on the other hand, it is thought to be lexically specified on a certain class of suffixes (Riad, 2009). Both neurophysiological (Roll et al., 2010) and response time studies (Söderström et al., submitted) suggest that Accent 2 is also 'phonetically' marked as compared to Accent 1. A similar co-occurrence of marked tonal patterns and increased association to grammatical structure is further illustrated with reference to other prosodic phenomena, such as Central Swedish boundary tones. We will begin by discussing the notion of 'phonetic markedness' before moving on to discuss 'passive attention' and finally to describe neurophysiological and behavioural results supporting our proposal.

Phonetic markedness

According to Trubetzkoy (1969), the *unmarked* member of a sound opposition should be the one that deviates the least from normal breathing and the *marked* member should consequently be the one that deviates more from normal breathing. Although the definition was originally formulated with respect to segmental obstructions, it might be extended to also include suprasegmental structures. Furthermore, elements deviating more from normal breathing could also be expected to be more perceptually salient, and involve more effort in their production. Thus, similar to Trubetzkoy's notion of markedness, Gussenhoven (2004) formulates an "effort code", according to which a tonal pattern that is perceived as involving more production effort tends to be associated with an emphatic interpretation. Gussenhoven assumes this to be what underlies the common grammaticalisation of focus in terms of relatively wide F0-excursions. Consolidating Trubetzkoy's and Gussenhoven's ideas, it could be assumed that 'marked' tonal patterns should tend to be perceived as more emphatic.

Tonal markedness and passive attention

We further propose an extension of this idea of tonal markedness, namely that there is a natural tendency for marked – or perceptually salient – F0 contours to increase attention at different levels of speech processing. Sudden changes in the physical environment tend to draw our involuntary attention. In the auditory domain, such "passive attention" is triggered by onset and offset of sounds as well as significant change in pitch or loudness (James, 1860; Näätänen, 1992). Especially relevant for language is "an-

ticipatory attention", i.e. attention directed to upcoming events rather than to the triggering event itself (Näätänen, 1992; Posner & Petersen, 1990). We suggest that marked tonal changes tend to function as triggers for passive anticipatory attention to associated grammatical structures.

Markedness and word accents

An example of the relation between markedness and need for increased attention to certain features of the speech signal is constituted by Swedish word accents. Stressed Accent 2 syllables have a high tone, whereas stressed Accent 1 syllables have a low tone (HL*) (Bruce, 1977). High tones would generally be thought to be more marked than low tones, since they involve faster vocal fold movement. Moreover, since the Accent 2 H*L tone falls through the stable spectral state of the vowel of the stressed syllable, it can be expected to be perceived more saliently than the Accent 1 fall, which occurs earlier, in the pretonic syllable. Because of its association with the pretonic syllable, the Accent 1 HL* is likely to be realized phonetically by a fall through segments with more rapid spectral change and therefore not be perceived as saliently as the Accent 2 movement (House, 1990). In other words, Accent 2 can be considered to be the *phonetically* marked accent.

According to the reasoning above, the phonetic markedness of Accent 2 could be expected to draw listeners' anticipatory attention to certain upcoming grammatically relevant aspects of the speech signal. Riad (1998; 2009), following Rischel (1963), has suggested that Accent 2 is lexically associated with a certain class of suffixes. Moreover, in Central Swedish, Accent 2 is the prosodic marker of compound words. Hence, it would seem likely that the marked Accent 2 tone could activate rapid anticipatory attention to possible associated suffixes or compound words.

Markedness and boundary tones

Similar reasoning can be applied to other marked tonal patterns. Left-edge boundary tones in Central Swedish are realized by a high tone in the last syllable of the first prosodic word in main clauses (Roll, 2006; Roll et al., 2009; 2011; Roll & Horne, submitted; Myrberg, 2010). The high left-edge boundary tone can be seen as a marked pitch excursion as compared to the alternative of not having a left-edge boundary tone. This marked tonal pattern can be expected to allocate anticipatory attention to the main clause structure it is associated with. As for right-edge boundary tones, these low (usually HL) patterns can, due to their relatively sudden decrease in pitch, be seen as a pitch excursion deviating from a 'default' slow downward F0 drift. Accordingly, these right-edge tones are associated with additional structural information as compared to their absence, i.e. the closure of a clause.

Tonal turning points as attentional cues

Since changes in pitch draw involuntary attention to them, the initiation of rises or falls, i.e. the phonetic *turning points* in F0-patterns (Bruce & Gårding, 1978) are likely to call for anticipatory attention. According to this approach, Central Swedish word accents have been analysed phonologically as HL* or H*L contours which differ only in the timing of the fall in relation to the stressed syllable (Bruce, 1977, 1987). However, in accordance with Bruce (2005), we assume that not only is the word accent fall relevant for word accent distinction, but also the rise to the H* in the Accent 2 LH*L pattern. We suggest that phonetic turning points in marked tonal patterns activate early anticipatory attention to the linguistic structures they are associated with. In what follows, we will present neurophysiological as well as behavioural data supporting this view.

Neurophysiological evidence

Using Event-Related Potentials (ERPs), Roll, Horne, and Lindgren (2010) investigated the neurophysiological correlates of perceiving the rise in the LH*L pattern in words with Accent 2 suffixes, e.g. minkar 'mink+pl', as compared to the fall in the corresponding singular forms with LHL* Accent 1 pattern e.g. minken 'mink+sg'. An early positive deflection, interpreted as an effect on the P200 component, was seen in the ERPs for the Accent 2 rise (Fig. 1). P200 is an electrophysiological component thought to reflect early allocation of attention following detection of behaviourally – or, in the case of language, linguistically – relevant acoustic changes (Näätänen, 1992). Further, the effects of mismatch between stem-tone pattern and suffix were investigated. Accent 2-associated suffixes yielded a late positivity, a 'P600' effect, when preceded by an Accent 1 tonal pattern on the stem, whereas Accent 1-associated suffixes were

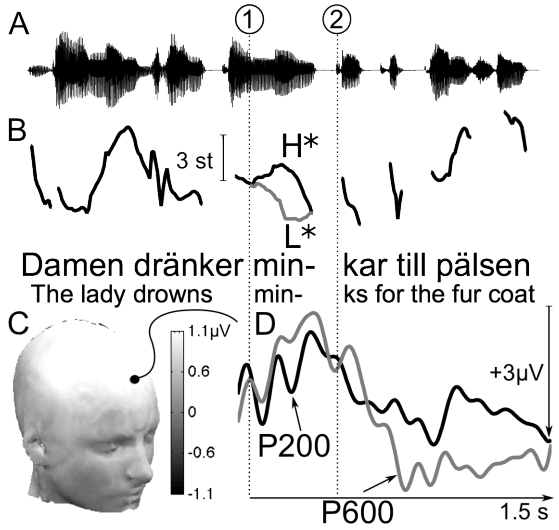


*Figure 1. Waveform (A) and F0 (B) for a sentence containing an unfocused Accent 2 word (*minkar *'minks') with matching (H*, black line) and mismatching (L*, gray line) word accent. Average Event-Related Potentials (ERPs) for 40 similar sentences and 22 participants are shown for the frontal electrode FZ (D). Vertical dotted lines indicate critical word/Accent 2 rise onset (1) and suffix onset (2). The Accent 2 rise yielded a P200 effect. Following a mismatching L* stem tone, the Accent 2 suffix –*ar *gave rise to a P600 effect. The central to anterior distribution of the P200 effect is also presented (C). Data from Roll et al. (2010).*

unaffected by the preceding stem tone. The P600 reflects reprocessing of unexpected or incorrect structures. The results suggest a process where the turning point identifying the onset of the marked high tone triggered allocation of early anticipatory attention to its associated suffixes (P200), leading to facilitated processing of the suffix. In the absence of a marked stem tone, the suffix was not activated, which resulted in reprocessing of the word form (P600).

High left-edge boundary tones also give rise to a P200 effect as compared to their absence (Roll et al., 2009; 2011; Roll & Horne, submitted). Thus, similar to Accent 2, the rise to the marked tone in the last syllable of a word is perceived as linguistically pertinent, and allocates attention to its associated main clause structure. Therefore, it reduces the P600 of the word *inte* 'not' in embedded *att* 'that' clauses such as *…att vandalerna intog inte Spanien* 'that the Vandals not conquered Spain', where *inte* shows main clause word order following the verb *intog* 'conquered'. The reduced P600 indicates that the left-edge boundary tone in the last syllable of *vandalerna* 'the Vandals' makes the main clause word order more expected in spite of the embedded context.

If the P200 reflects early allocation of attention cued by grammatically relevant tonal events, it should also be seen for marked falls in the appropriate context. Effectively, Roll and Horne (submitted) found a P200-like positive effect even for HL right-edge boundary tones in the second noun (*vännen* 'the friend') in sentences like *Lingvisten spöa' vännen och fonetikern stöp/stort i ett uppgjort lopp* 'The linguist beat the friend and the phonetician fell/greatly in a set-up race'. The P200 was followed by a later positivity, interpreted as a Closure Positive Shift (CPS), showing closure of the first clause (*Lingvisten spöa' vännen* 'the linguist beat the friend'). It thus seems that in this case, the marked fall constituted by the right-edge boundary HL pattern oriented attention to upcoming clause closure instead of clause continuation.

Psycholinguistic evidence

Support for the assumed phonetic markedness of Accent 2 has also been found in a response time experiment originally carried out to investigate the effects of tonal mismatch on the perception and interpretation of Central Swedish word accents (Söderström et al., submitted). Test stimuli consisted of short pronoun+verb utterances, with narrow focus placed on the pronoun so as to avoid the focal rise on the verb. The task was to decide as quickly as possible whether the verb was in the present or the past tense, i.e. either HAN röker ('HE smokes', Accent 1) or HAN rökte ('HE smoked', Accent 2) (see Fig. 2). Half of the stimuli had a stem tone/suffix mismatch. All stimuli were created by splicing together stems with suffixes. The hypothesis, based on the ERP research discussed above, was that words with an Accent 1 stem tone and Accent 2-inducing suffix would be the most difficult to process, because there was no H* tone to cue the upcoming Accent 2 lexicalized suffix. This finding was replicated in the response time study: mismatched Accent 1 stems followed by Accent 2-inducing past tense suffixes were indeed the most difficult to process (Fig. 2). Another interesting finding was that non-mismatched ("correct") Accent 2 words were more difficult to process as compared with non-mismatched ("correct") Accent 1 words in general. This was taken as further support for

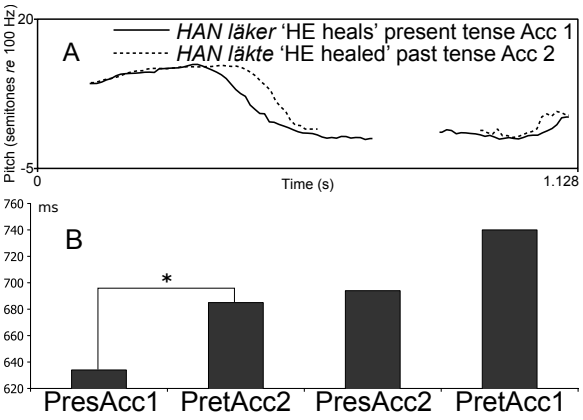


Figure 2. F0 curves for two Swedish utterances HAN läker *'HE heals' and* HAN läkte *'HE healed', Accent 1 and 2, respectively (A). Mean response times for four conditions: 'Pres' = 'present tense', 'Pret' = 'past tense', so that e.g. "PresAcc2" = present tense word with Accent 2 associated with the stem (B).*

the idea that Accent 2 is the marked member of the opposition Aand that Accent 1 is unmarked. As already stated above, this could be explained by the fact that the H* Accent 2 tone is associated with more word forms as compared to Accent 1. A further finding in the response time study was that words with an Accent 2 stem tone mismatched with an Accent 1 suffix were also difficult to process, although less difficult than Accent 1 stem tones mismatched with Accent 2 suffixes. An explanation for this is that many forms with Accent 2 including compound words are activated when an Accent 2 tone is heard, and must be deactivated when the Accent 1 suffix is heard.

Conclusion

Results from the empirical studies discussed above support the idea that phonetically marked tonal patterns in Central Swedish (Accent 2, boundary tones) constitute important cues for anticipatory attention in speech processing. Both tonal markedness and F0-turning points related to Accent 2 and boundary tones are associated with frequent, well-defined linguistic structures (Accent 2 suffixes, compounds, initial and final clause boundaries) in Swedish.

References

Bruce G (1977). *Swedish word accents in sentence perspective*. Lund: Gleerups.

Bruce G (1987). How floating is focal accent? In: Gregersen K, Basbøll H, eds, *Nordic prosody IV*. Odense: Odense Univ. Press, 41-49.

Bruce G (2005). Intonational prominence in varieties of Swedish revisited. In Jun, S-A, ed, *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford Univ. Press, 410-430.

Bruce G, Gårding E (1978). A prosodic typology for Swedish dialects. In: Gårding E, Bruce G, Bannert R, eds, *Nordic prosody*. Lund: Gleerups, 219-228.

Gussenhoven C (2004). *The phonology of tone and intonation*. Cambridge: Cambridge Univ. Press.

House D (1990). *Tonal perception in speech.* Lund: Lund Univ. Press.

James W (1890). *The principles of psychology*. New York: Holt.

Myrberg S (2010). *The intonational phonology of Stockholm Swedish*. Ph.D. dissertation, Nordic languages, Stockholm University.

Näätänen R (1992). *Attention and brain function*. Hillsdale, NJ: Erlbaum.

Posner MI, Petersen SE (1990). The attention systems of the human brain. *Annual Review of Neuroscience,* 13: 25-42.

Riad T (1998). The origin of Scandinavian tone accents. *Diachronica,* 15: 63–98.

Riad T (2009). The morphological status of accent 2 in North Germanic simplex forms. In: Vainio M, Aulanko R, Aaltonen O, eds, *Nordic prosody X*. Frankfurt am Main: Peter Lang, 205–216.

Rischel J (1963). Morphemic tone and word tone in Eastern Norwegian. *Phonetica,* 10: 154–164.

Roll M (2006). Prosodic cues to the syntactic structure of subordinate clauses in Swedish. In: Bruce G and Horne M, *Nordic prosody IX*. Frankfurt am Main: Peter Lang, 195-204.

Roll M, Horne M (submitted). Interaction of right- and left-edge prosodic boundaries in syntactic parsing.

Roll M, Horne M, Lindgren M (2009). Left-edge boundary tone and main clause verb effects on syntactic processing in embedded clauses—An ERP study. *Journal of Neurolinguistics,* 22: 55-73.

Roll M, Horne M, Lindgren M (2010). Word accents and morphology—ERPs of Swedish word processing. *Brain Research*, 1330: 114-123.

Roll M, Horne M, Lindgren M (2011). Activating without inhibiting: Left-edge boundary tones and syntactic processing. *Journal of Cognitive Neuroscience*, 23: 1170-1179.

Söderström P, Roll M, Horne M (Submitted). Processing morphologically conditioned word accents.

Trubetzkoy N (1969). *Principles of phonology*. Berkeley: University of California Press.

Acknowledgements

This research has been supported by grants 2007-1759 and 2009-1773 from the Swedish Research Council.

Children's perception of their modified speech – preliminary findings

Sofia Strömbergsson
Department of Speech, Music and Hearing, KTH

Abstract
This report describes an ongoing investigation of 4-6 year-old children's perception of synthetically modified versions of their own recorded speech. Recordings of the children's speech production are automatically modified so that the initial consonant is replaced by a different consonant. The task for the children is to judge the phonological accuracy (correct vs. incorrect) and the speaker identity (me vs. someone else) of each stimulus. Preliminariy results indicate that children with typical speech generally judge phonological accuracy correctly, of both original recordings and synthetically modified recordings. As a first evaluation of the re-synthesis technique with child users, the results are promising, as the children generally accept the intended phonological form, seemingly without detecting the modification.

Background

In children with speech sound disorders, the relation between speech perception and speech production is still not fully understood. Some children who produce a sound in error also have problems with the perception of the distinction between this sound and the sound that they substitute it for (Locke, 1980). Other children with the same speech deviance do perceive this error when others produce it, but still fail to perceive the same error in their own speech production (Aungst & Frick, 1964). This discrepancy between external perception (i.e. perception of other people's speech) and internal perception (i.e. perception of one's own speech) is common in children with speech sound disorders, but also in typical speech development (Kuczaj, 1983).

Presenting children with recordings of their own deviant speech has been suggested as a way of increasing the children's awareness that their speech is deviant and to stimulate self-monitoring (Hoffman & Norris, 2005). Shuster (1998) presented children and adolescents not only with recordings of their own deviant /r/ productions, but also with corrected versions of the same recordings. The author hypothesized that the children would perceive the corrected versions as incorrect, and that this could explain why they persisted in producing /r/ in error. However, the results showed that the children judged both the deviant (original) /r/-recordings and the corrected (edited) /r/-recordings as correct, a finding that suggests that the children's internal /r/ representations are too allowing.

Although the data in Shuster (1998) did not support her hypothesis, the study is an example of how corrected recordings of deviant speech can be used to gain insight to children's speech perception in relation to their speech production. However, in Shuster's experiment, the recordings were edited by hand, by LPC manipulation and subjective auditory analysis. Assumably, it would be even more valuable if the correction was generated automatically, as it would not only be time-saving, but it would also allow the speaker to receive immediate feedback to his/her speech production. Immediate and automatic modification of recorded deviant speech would also allow examination of the commonly held suggestion that when children with deviant speech monitor their own speech production, they perceive what they intended to produce, rather than the phonetic product of their intent (e.g. Locke & Kutz, 1975).

A method for automatic re-synthesis of recorded child speech has recently been developed at KTH (Strömbergsson, 2011). Here, a unit selection approach is used to replace an initial voiceless plosive in a recorded speech sample by another voiceless plosive, retrieved from a recording of another child. However, although previous results from an evaluation of this re-synthesis method with adult listeners are

promising, a naturalistic setting with children as users puts higher demands on the technique, and would be more revealing of its potential use.

Purpose

The purpose of the present study is to evaluate the modified re-synthesis technique, by letting 4-6 year-old children with typical speech judge phonological accuracy and speaker identity of their own modified speech. Later, the collected data will serve as control data in a planned study of how children with deviant speech perceive synthetically corrected versions of their own recorded deviant speech.

Method

At the time of writing, 9 children with normal hearing and typical speech and language development had participated in the study. The children were between 4 and 6 years old (ranging from 4;0 to 6;0, M = 61 months, SD = 9.0 months). All children were recruited through pre-schools in Stockholm, and all had at least one native Swedish parent.

All children participated in two 20-minute test sessions, in which they performed the following tasks:

1. External discrimination (ED)
2. Internal discrimination, modified speech (IDMod-direct)
3. Test of Phonological Working Memory (PWM; Radeborg et al, 2006)
4. Internal discrimination, modified speech – delayed (IDMod-delayed)
5. Test of Reception of Grammar (TROG; Bishop, 2003)
6. Assessment of phonological production (Fonemtestet; Hellquist, 1995)
7. Internal discrimination, modified speech – speaker identity (IDMod-identity)

Tasks 1-4 were performed in the first test session, and tasks 5-7 in the second session. Tasks 3, 5 and 7 are standard Swedish tests for assessment of speech and language abilities, that were used to ascertain that all participating children had age adequate speech and language. The experiment took place in a separate room with limited noise at the different preschools. The children were fitted with a headset and the experimenter with headphones to supervise the recordings and listening tasks.

In the ED task, the children were presented with 2 original and 10 synthetically modified recordings of other children. These stimuli were used in a previous study (Strömbergsson, 2011), where adult listeners were in high agreement of their phonological accuracy. (Per item kappa was 0.85 or more for all selected items.) The task for the children was to judge the accuracy of each stimulus, by pointing either to a picture illustrating the word on the screen when they heard the depicted word being said, or to a picture of a large X when they perceived the word as "wrong". An example screen shot is presented in Figure 1.

In the three IDMod tasks, a recording script of 8 words was used (see Appendix). In each IDMod task, the words in the script were repeated twice, resulting in 16 stimuli per task.

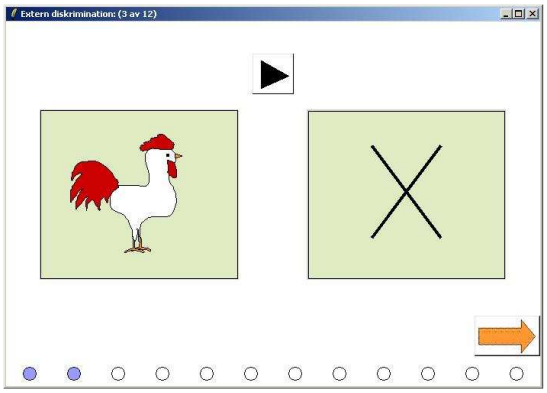


Figure 1. The task setup for judging phonological accuracy. The children pointed to the rooster if they perceived the audio stimulus as "tupp" (rooster), and to the X-symbol when the perceived the stimulus as "wrong" or "not rooster".

In the IDMod-direct task, the children were recorded when producing the words in the script. Immediately after having recorded a word, the recording was either a) re-synthesized with a different initial consonant (6 stimuli), b) re-synthesized with a phonologically equivalent initial consonant produced by another child (6 stimuli), or c) left unchanged (4 stimuli). (The number of stimuli beginning with /t/ and stimuli beginning with /k/ was equal in all three conditions.) The children were told that The Parrot would imitate what they had just said, but that The Parrot would not always get it right. The task for the children was to, for each word in the script, judge if The Parrot could say it "right" or "wrong", in the same way as illustrated in Figure 1.

In the IDMod-delayed task, the children were again presented with the 16 stimuli that were recorded/generated in the IDMod-direct task. Again, their task was to judge if The Parrot's production was "right" or "wrong".

The IDMod-identity task was designed to test whether the children could detect the synthetic modification of their recordings. In the same way as in the IDMod-direct task, the children were first recorded when producing a word, which then immediately was either a) re-synthesized with a different initial consonant, b) re-synthesized with a phonologically equivalent initial consonant produced by another child, or c) left unchanged. This time, the children were told that The Imitation Monkey ("Härmapan") would imitate what they had just said, but that he would only do it occasionally. The task for the children was to point to a picture of a child if they believed that the recording sounded exactly as they had produced it, or to a picture of The Imitation Monkey if they perceived the stimulus as being different from how they produced it.

For the re-synthesis, a unit selection approach was used, with a speech database of 1444 child recordings. Details are described in Strömbergsson (2011).

Results

The children's performance on judging the phonological accuracy of original and re-synthesized recordings of other children, the ED task, was high; the mean score was 94% correct (range 83% to 100% correct). Of the 108 responses, only 6 were misjudged; all of these were responses to re-synthesized stimuli. Agreement among the children, as measured by Fleiss' kappa, was 0.79.

Overall, the children's performance on judging phonological accuracy of original and re-synthesized recordings of their own speech was also high; the mean score on the IDMod-direct task was 95% correct (range 88% to 100% correct), and on the IDMod-delayed task 93% correct (range 75% to 100% correct).

Cohen's kappa was computed to examine the test-retest reliability in the IDMod-direct and IDMod-delayed tasks, i.e. the extent to which the children gave the same response to the same items in the two different test conditions (direct vs. delayed). The resulting kappa of .83 indicates high agreement.

To explore whether the children performed differently on modified (impaired or un-impaired) stimuli than on original stimuli, a Pearson's χ^2-test was conducted. However, no such dependence was found, neither in the direct condition (p = .68, Fisher's exact test, FET), nor in the delayed condition (p = .19, FET). (Since the number of misjudged original stimuli was smaller than 5 in both test conditions, Fisher's exact test was used.)

To explore the association between phonological accuracy of the stimuli and the children's response accuracy, another Pearson's χ^2-test was conducted. This would reveal if the children performed differently on incorrect stimuli (i.e. modified impaired stimuli) than on correct stimuli (i.e. original recordings and modified un-impaired stimuli). However, this analysis showed no such dependence in the direct listening condition (p = .35, FET), whereas in the delayed listening condition, the dependence was significant (p = .02, FET). Of the 11 incorrect responses (of 144 responses in total), 8 were responses to impared stimuli.

The children's ability to detect whether a stimulus had been modified or not was targeted in the IDMod-identity task. However, at the time of writing, only 4 children had completed this task, which was considered insufficient for statistical analysis. An inspection of the present data, however, as displayed in Table 1, reveals that original recordings were always perceived correctly as un-modified, and that impaired stimuli were almost always detected as modified (21 of 24). Modified un-impaired stimuli, on the other hand, were almost always incorrectly percieved as un-modified (21 of 24).

Table 1. The children's accuracy in detecting stimulus modification.

	Accuracy		
Stimulus type	Wrong	Correct	Total
Original	0	16	16
Mod: correct	21	3	24
Mod: impaired	3	21	24
Total	24	40	64

Discussion

The study described here is still ongoing, and the results presented are preliminary. However, as a first child evaluation of the re-synthesis technique, the results are promising, as they indicate that the children accept the intended phonological form of the re-synthesized stimuli,

seemingly without detecting when re-synthesis has taken place.

The children were able to judge phonological accuracy in original and re-synthesized speech with around 95% accuracy. As all participating children follow typical speech and language development, one might have expected 100% accuracy. However, there are some possible explanations to the children's sub-optimal performance. One is, of course, that the re-synthesis process might be disturbing. But although this might be a possible explanation when the children judge the phonological accuracy in recordings of other children (the ED task), it does not seem to be the case when they judge their own original and modified recordings (the IDMod-direct and IDMod-delayed tasks). Here, the children's performance is not dependent on whether the stimuli are original or modified recordings. Instead, there is a tendency in the children to misjudge impaired stimuli as correct. As a picture is shown when the children hear the stimulus, the children might be biased towards hearing what they expect to hear, i.e. the word that is depicted on the screen. Hopefully, more data will allow firmer conclusions.

The children's agreement on the external discrimation task is lower than the agreement among the adult listeners in Strömbergsson (2011). This might reflect a real difference between children and adults, but the difference might also level out as the number of participants increase.

Although only few children had completed the modification detection task, the present data provides more support to the suggestion that the re-synthesis of recordings is too subtle for the children to perceive when the phonological identity of the initial consonant is unchanged. When the phonological identity *is* changed, however, the children almost always perceive the change. This is indeed promising for the intended use of the re-synthesis technique with children with deviant speech, as the idea is to generate convincing corrections of the child's deviant speech; convincing in the sense that the speaker identity and naturalness of the utterance should not be affected, when the phonological form is altered.

This ongoing experiment will continue with the inclusion of more children with typical speech and language development, and will also be extended to include children with deviant speech.

Acknowledgements

This work was funded partly by The Swedish Graduate School of Language Technology, and partly by the Promobilia foundation.

References

Aungst L F, Frick J V (1964). Auditory Discrimination Ability and Consistency of Articulation of /r/. *J Speech Hear Dis*, 29: 76-85.

Bishop D V M (2003). 'Test for Reception of Grammar Version 2, TROG-2. The Psychological Corporation, London, UK.

Hellquist B (1995). Fonemtest. Kortversionen. [Phoneme Test. The short version]. Pedagogisk Design, Malmö.

Hoffman P R, Norris J (2005). Intervention: Manipulating Complex Input to Promote Self-Organization of a Neuro-Network. In: A G Kamhi & K E Pollock, eds, *Phonological Disorders in Children*. Paul H. Brookes Publishing Co, 139-156.

Kuczaj S A (1983). '"I mell a kunk!" - evidence that children have more complex representations of word pronunciations which they simplify. *J Psycholinguistic Res*, 12: 69-73.

Locke J L (1980). The Inference of Speech Perception in the Phonologically Disordered Child. Part II: Some Clinically Novel Procedures, Their Use, Some Findings. *J Speech Hear Disord*, 45: 445-468.

Locke J L, Kutz, K J (1975). Memory for Speech and Speech for Memory. *J Speech Hear Res*, 18: 176-191.

Radeborg K, Barthelom E, Sjöberg M, Sahlén, B (2006). A Swedish non-word repetition test for preschool children. *Scandinavian Journal of Psychology*, 47: 187-192.

Shuster L I (1998). The Perception of Correctly and Incorrectly Produced /r/. *J Speech Lang Hear Res*, 41: 941-950.

Strömbergsson S (2011). Segmental re-synthesis of child speech using unit selection. In: Proceedings of ICPhS 2011, Hong Kong.

Appendix

Orthography	Transcription	In English
k	/'ko:/	*(the letter k)*
kaka	/'kɑ:ka/	*cake*
kulle	/'kʉlə/	*hill*
kung	/'kʊŋ/	*king*
tak	/'tɑ:k/	*roof*
tumme	/'tʉmə/	*thumb*
tupp	/'tʉp/	*rooster*
tåg	/'to:g/	*train*

Cortical N400-Potentials Generated by Adults in Response to Semantic Incongruities

Eeva Klintfors, Ellen Marklund, Petter Kallioinen, Francisco Lacerda
Department of Linguistics, Section for Phonetics, Stockholm University

Abstract

Eight adult participants were investigated in a pre-experiment for the future assessment of semantic N400 effects in children. The materials were words presented in semantically incongruent vs. congruent picture contexts. For example, the word duck was played while a picture of a tree was shown in the incongruent test condition vs. the word duck was played while a picture of a duck was shown in the congruent test condition. A larger N400 effect was expected in response to the incongruent audio-visual pairings. The results showed in time extended peak-to-peak differences between congruent and incongruent audio-visual pairings at the centroparietal, parietal and parieto-occipital recording sites. This study was performed to validate the current materials to be used to answer questions on appearance of the N400 component in children.

1 Introduction

This pre-experiment on cortical responses to semantic incongruities in adults was performed within the project called "Early Development of Hemispheric Specialization for Speech Processing"[1]. The main aim of the project is to study establishment of language-related specialization in the brain and its specific links to different phases of language development early in life. Two EEG-studies (electroencephalography) are planned to be performed within the projocet: (1) Swedish materials, materials from a foreing language and "rotated speech" will be used to study lateralization of speech and its prosodic components in a developmental perspective in young children, and (2) Semantically incongruent audio-visual (AV) materials will be used to study the development of cortical activites that are likely to mediate semantic processing, as revealed by the so called N400 incongruity effect.

It is our goal that the basic research data collected from typically developing children will in future be applicable in establishing potential early signs of deviant language development in children with cognitive disorders, such as autism spectrum discorders (ASD).

Thus, neurophysiological experiments with adults were performed as a first step to reach the project goals. The rationale for investigating N400-potentials in response to semantic incongruitites across the AV domains in adults is to verify use of ERPs in response to the current materials. Also these data function as a baseline against which results from typically, as well as atypically developing children may be tested.

The adults in the current study were tested on auditory (A) and visual (V) materials in two test conditions: one using congruent, and the other using incongruent A-V pairings of objects.

2 Background

Emergence of linguistically induced semantic representations is known to appear during the second year of life (Bates et al., 1994). The child naming objects in its environment is naturally the most obvious demonstration of an indivual having conceptual compentence. However, the mechanisms involved in integration of meaning at a higher level are not well understood nor easily observed in children. These processes are presumably best investigated with help of imaging studies to complement results from behavioural studies.

Event-related potentials (ERPs) are a series of positive and negative voltage deflections in the ongoing EEG that are time-locked to sensory, motor, or cognitive events. The ERP component called N400 (N indicating that the peak is negative at around 400ms after the stimulus) is presumably the best known evoked response potential that is sensitive to variations in the semantic contents of words and sentences

[1] Financed by the Swedish Research Council.

(Bentin & McCarthy, 1994; Hagoort & Brown, 2002; Kutas, 1997). The N400 effect functions in the way that presentation of A-V pairings induce it in general, but the effect is strongest in response to incongruent materials (i.e. when the word presented is semantically incongruent with the picture).

What is known about the N400 effect in adults and children? First, semantically anomalous, relative to well-formed sentences produce a larger N400 effect both in children and adults (N=130, 5 to 26 years) (Holcomb, Coffey, & Neville, 1992). Yet N400 activation in younger children (5 to 14 years) is widely distributed, including the midline frontal och frontalcentral recording sites, relative to activation in adults and older children where it is limited to posterior areas only. Second, the N400 amplitude may be significantly higher for children (5 to 13 years) relative to adults. In addition, significant effects for delayed peak latency and extended duration of the effect in children are found (Atchley et al., 2006; Friederici & Hahne, 2004; Juottonen, Revonsuo, & Lang, 1996). In young infants, semantic processing mechanisms indexed by N400 are suggested to be present at 14 and 19 months of age (Friedrich & Friederici, 2004; Friedrich & Friederici, 2005).

2.1 The aim of the current study

The present study investigates adult N400-response using a child-adapted experiment design, creating a baseline for future studies. A picture-word paradigm with coloured pictures of familiar objects and slowly spoken words, either naming the object in the picture (A-V congruent condition) or naming another object (A-V incongruent condition) is used. The latter condition is expected to elicit the most significant N400-responses (see Figure 1). The distribution of N400 is expected to be maximal at posterior midline scalp locations, in specific the centroparietal (CPz), parietal (Pz), and parieto-occipital (POz) sites.

3 Method

The participants were 8 native speakers of Swedish (3 male and 5 female). Mean age was 32 years, ranging from 21 to 55 years. Seven of the participants were right-handed and all of them reported normal hearing. One of the participants was familiar with the materials and the research question of the study. The stimuli consisted of cartoon-like pictures of objects, and recordings of the labels (words) of the objects (figure 1).

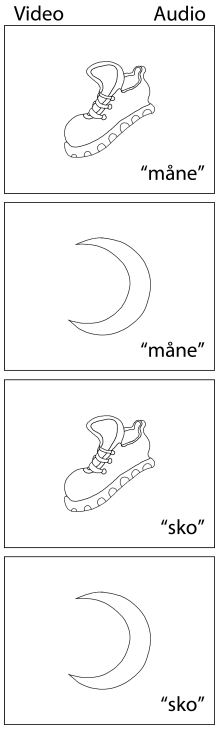


Figure 1. Examples of experiment trials. The word was uttered 1000 ms after the picture. The two middle trials belong to the congruent condition (audio and video match) and the top and bottom trials belong to the incongruent condition (audio and video do not match). Trials were presented in random order.

In earch trial, the picture was first presented, and after 1000 ms, the label was presented. In the congruent condition the picture of the object was paired with the correct label. In the incongruent condition, each picture of an object was paired with an incorrect label, controlled so that the correct label for the image and the incorrect label used differed in their first phonemes, with regards to articulation-type and voicing. A total of 48 object labels were used in the study, each presented once in the congruent condition and once in the incongruent condition, resulting in a total of 98 trials. The trials were presented in random order. The words were chosen from the Swedish Early Communicative Developmental Inventory (SECDI; Eriksson & Berglund, 1996) for compatibility with future infant studies, in which age-appropriate words will be chosen. The words were read by a female native speaker of Swedish and recorded in an anechoic chamber. The materials were digitally saved to a hard drive, segmented and manually matched for

loudness. Single word duration ranged from 526 to 1178 ms.

A net of electrodes was placed on the participant's head, and stimuli were presented using a computer screen and loudspeakers. Participants were asked to listen attentively to the stimuli, but were given no explicit task. The experiment session lasted approximately 14 min.

3.1 Data preparation

Net Station tools cleaned the data with a band pass filter from 0.3–30 Hz noise, segmented the responses into 96 segments of 1100 ms each (100 baseline before stimuli onset, 1000 ms after stimuli onset), and removed unusable channels and segments (*e.g.* eye blinks) before collating segment averages across stimuli conditions, and referencing the EEG-voltage measurements to a baseline prior to stimuli onset.

4 Results

Two participants (N=2) were excluded from the analysis based on the characteristic brain wave response, known as P300, often elicited by an infrequent, task-relevant stimulus. The P300 effect is present for participants number 3 and 6 illustrated by the upgoing thin line corresponding to semantically incongruent trials towards the end of the wave in Figure 2.

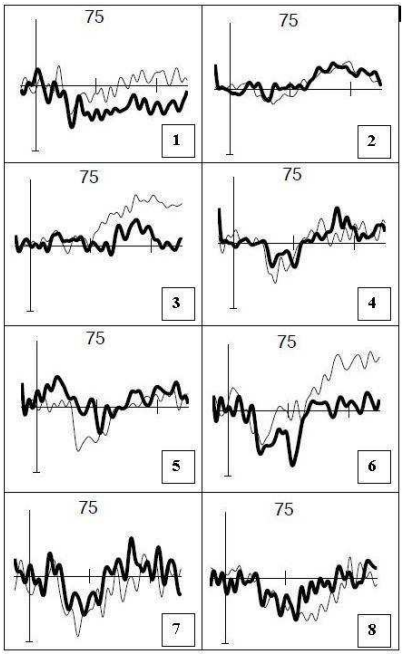


Figure 2. The P300 effect, considered to reflect response to non-standard items intermixed with high-probability items, is illustrated by the up-going thin line for participant number 3 and 6.

Figure 3 displays average ERPs elicited by congruent (thick lines) *vs.* incongruent (thin lines) trials for the participants (N=6) included in the analyses. The time period begins at 0 ms corresponding to the stimulus onset and ends 1000 ms post-stimulus. Waveforms for five electrode sites are depicted (Fz=11, FCz=6, CPz=55, Pz=62 and POz=72) to highlight the scalp distributions for the N400 waveforms. The incoruity effect, illustrated by the diverging wave-forms, was most prominent at the cenroparietal (CPz=55), parietal (Pz=62), and parieto-occipital (POz=72) sites. The incongruity effect was present approximately within 300 to 800 ms at the CPz (55), and within 200 to 800 ms at the Pz (62) and POz (72).

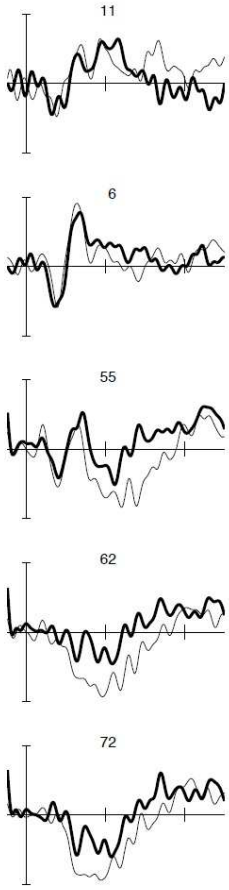


Figure 3. Averaged ERP wave-forms generated by the congruent (thick lines) vs. incongruent (thin lines). The short vertical marks along the time line correspond to 400 ms and 800 ms post stimulus. The N400 effect is most prominent (illustrated by the diverging wave-forms) within 300 to 800 ms at the CPz=55, and within 200 to 800 ms at the Pz=62 and POz=72 eclectrodes.

5 Discussion

As shown in previous studies, adult participants (N=6 of total N=8) in the current study elicited a N400 effect in response to semantic incongruities at the midline posterior scalp locations.

Two participants were excluded from the analysis based on a prominent P300 effect that was suspected to cancel out their N400 incongruity effect. We used this rejection criterion because P300 is thought to reflect processes involved in stimulus categorization. For example, P300 is often elicited in response to low-probability target items that are inter-mixed with high-probability non-target (or "standard") items. In the current study the 48 pictures of objects shown once in the congruent condition and once in the incongruent condition presented in random order. Despite of the fact that the congruent vs. incongruent trials were presentend in random order, we suspect that subsequent presentations of "standard" congruent stimuli might for the participants in question have caused a P300 effect in response to seemingly novel low-probability incongruent trials. In the current experiment participants were asked to listen attentively to the stimuli, but they were not given any explicit task. In future studies P300 effect is presumably best eliminated by giving participants a pseudo task to perform (such as to press a button in response to every fifth trial).

The current study is a pre-experiment on eight adult participants to study N400 effect in children. Therefore it is premature to speculate on implications of the results given the obvious need for replication and extension of the current research. Also, additional analyses of cortical responses *within* the incongruent condition are to be performed on animate *vs.* inanimate objects. However, if these results prove robust then they could be very informative and function as a baseline for N400 studies in children with typical language development *vs.* in children with ASD.

Typically developing infants are expected to have a higer N400 component amplitude, more delayed component latency, and more widely distributed scalp distribution relative to adults. The onset of speech and other linguistic milestones are typically delayed in children with ASD. Autism is of particular interest in this project because it offers the opportunity of testing a theoretical perspective proposing that correlated sensory information is the very key to the development of linguistic referential function. The activation pattern found in typically developing children is not expected to show as clearly (if at all) in children with ASD.

References

Atchley RA, Mabel RL, Stacy BK., Kwasny KM, Sereno JA, Jongman A (2006). A comparison of semantic and syntactic event related potentials generated by children and adults. *Brain and Language*, 99: 236-246.

Bates E, Marchman V, Thal D, Fenson L, Dale P, Reznik S, Hartung J (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, 21: 85-123.

Bentin S, McCarthy G (1994). The effects of immediate stimulus repetition on reaction times and event-related potentials in tasks of different complexity. *Journal of Experimental Psychology: Learning Memory, and Cognition*, 20: 130-149.

Eriksson M, Berglund E (1996). Swedish Early Communicative Development Inventory - words and gestures. *First Language*, 19: 55-90.

Friedrich M, Friederici AD (2004). N400-like Semantic Incongruity Effect in 19-Month-Olds: Processing Known Words in Picture Contexts. *Journal of Cognitive Neuroscience*, 16: 1465-1477.

Friedrich M, Friederici AD (2005). Lexical priming and semantic integration reflected in the event-related potential of 14-month-olds. *Neuroreport*, 25: 653-656.

Friederici AD, Hahne A (2004). Development patterns of brain activity reflecting semantic and syntactic processes. In J Weissenborn & B Houle, eds, *Approaches to bootstrapping: Phonological, lexical, syntactic, and neurophysiological aspects of early language acquisition*. Amsterdam/Phila-delphia: John Benjamin, 231-246.

Hagoort P, Brown CM (2002). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia*, 38: 1518-1530.

Holcomb PJ, Coffey SA, Neville HJ (1992). Visual and auditory sentence processing: A developmental analysis using event-related brain potentials. *Developmental Neuropsychology*, 8: 203-241.

Juottonen K, Revonsuo A, Lang H (1996). Dissimilar age influences on two ERP waveforms (LPC and N400) reflecting semantics context effect. *Cognitive Brain Research*, 4: 99-107.

Kutas M (1997). Views on how the electrical activity that the brain generates reflects the functions of different language structures. *Psychophysiology*, 34: 383-398.