ISSN 0280-526X

Lund University
Centre for Languages and Literature

General Linguistics
Phonetics

Working Papers

54. 2010

Proceedings from *Fonetik 2010*
Lund, June 2–4, 2010

Lund University
Centre for Languages and Literature

General Linguistics
Phonetics

Working Papers

54. 2010

Proceedings from *Fonetik 2010*
Lund, June 2–4, 2010

Edited by Susanne Schötz and Gilbert Ambrazaitis

Working Papers
Department of Linguistics and Phonetics
Centre for Languages and Literature
Lund University
Box 201
S-221 00 LUND
Sweden
Fax +46 46 222 32 11
http://www.sol.lu.se/

This issue was edited by Susanne Schötz and Gilbert Ambrazaitis

ISSN 0280-526X

Printed in Sweden, Mediatryck, Lund, 2010

Preface

This volume of the Working Papers in General Linguistics and Phonetics contains the Proceedings of FONETIK 2010, the 23rd Swedish Phonetics Conference.

FONETIK 2010, held at the Centre for Languages and Literature, Lund University, on June 2-4, is one in the series of annual conferences for phoneticians and speech scientists in Sweden which regularly also attract participants from Denmark, Finland and Norway and sometimes from other countries as well. There are 23 contributions represented in this volume. A large variety of topics are covered in the papers, and we believe that the volume gives a representative overview of current phonetics research in Sweden.

We would like to thank all contributors to the Proceedings. We would also like to acknowledge the valuable support from The Swedish Phonetics Foundation (Fonetikstiftelsen) and from the Centre for Languages and Literature.

Lund, May 2010

The Organizing Committee

Gilbert Ambrazaitis, Gösta Bruce, Johan Frid, Susanne Schötz, My Segerup, and Mechtild Tronnier

Previous Swedish Phonetics Conferences (from 1986)

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

Contents

Expression of emotions in spoken Swedish – a corpus study **1**
Åsa Abelin

Research focus: Interactional aspects of spoken face-to-face communication **7**
Jonas Beskow, Jens Edlund, Joakim Gustafson, Mattias Heldner, Anna Hjalmarsson & David House

Very short utterances in conversation **11**
Jens Edlund, Mattias Heldner, Samer Al Moubayed, Agustín Gravano & Julia Hirschberg

An acoustic analysis of purring in the cheetah *(Acinonyx jubatus)* and in the domestic cat *(Felis catus)* **17**
Robert Eklund, Gustav Peters & Elizabeth D. Duthie

Effects of open and directed prompts on filled pauses and utterance production **23**
Robert Eklund & Mats Wirén

Dynamic vocal tract length normalization in speech recognition **29**
Daniel Elenius & Mats Blomberg

Vowel dependence for electroglottography and audio spectral tilt **35**
Laura Enflo

Automatic estimation of pitch range through distribution fitting **41**
Johan Frid & Gilbert Ambrazaitis

Quantity in Central Standard Swedish and Fenno-Swedish **47**
Pétur Helgason, Catherine Ringen & Kari Suomi

Fo curves – smooth, seamless yet pulsed? **51**
Björn Lindblom

Coarticulation as incomplete interpolation **57**
Björn Lindblom, Diana Krull & Harvey M. Sussman

Voice similarity – a comparison between judgements by human listeners and automatic voice comparison **63**
Jonas Lindh & Anders Eriksson

Cross-modal magnitude matching as a precursor of multi-modal speech perception **69**
Ellen Marklund, Anna Ericsson & Francisco Lacerda

Phonetic and phonological aspects in relation to theory of mind and possible consequences for speech training in autism 75
Pia Nordgren

The effect of speaking rate on the perception of vowel-initial and vowel-final [h] 79
Jörgen L. Pind

Synthesising intonational varieties of Swedish 85
Susanne Schötz, Jonas Beskow, Gösta Bruce, Joakim Gustafson, Björn Granström & My Segerup

English *of* in L1 and L2 speakers' read and spontaneous speech 91
Helena Spilková and Wim A. van Dommelen

How widespread is preaspiration in Italy? A preliminary acoustic phonetic overview 97
Mary Stevens

Durations of phonologically long segments in native and foreign accented Swedish 103
Bosse Thorén

Bertil Malmberg, first chair of phonetics in Sweden – peer evaluations 107
Paul Touati

Earwitnesses: The effect of type of voice lineup in identification accuracy and the realism in confidence judgments 113
Elisabeth Zetterholm, Farhan Sarwar & Carl Martin Allwood

Post surgery effects on VOT for Parkinson Disease STN/DBS patients 119
Joel Åkesson, Jonas Lindh & Lena Hartelius

A preliminary study of audiovisual integration of roundedness in front vowels: limitation due to discrepancy in jaw depression 125
Niklas Öhrström

Expression of emotions in spoken Swedish – a corpus study

Åsa Abelin
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden

Abstract

How are emotions expressed in spoken interaction? Prosodic emotional expressions and emotional words were searched for in the Gothenburg Spoken Language Corpus. Analyses were also made with respect to different activities, e.g. business negotiation, lecture and discussion. Some emotions are expressed with prosody, e.g. surprise, while other emotions are more often expressed with adjectives, e.g. anger and fear. Different emotional expressions are displayed in different social activities. Analyses were also made of prosodic emotional expressions in relation to carrier phrase, showing that the linguistic meaning of some words cannot be determined without the prosody. Preliminary acoustic analyses were made.

Introduction

Communication of emotions is vital in spoken interaction; it is difficult to imagine two, or more, persons interacting face-to-face, or by telephone, without them expressing different emotions almost constantly. The prosodic expression of emotions influences the way we interpret each other’s words, in fact it can be claimed that lexical and emotional meaning cannot always be separated, other than analytically: What is the meaning of the Swedish word “nähä”? That cannot be determined without hearing it spoken since the meaning depends on the prosody; if there is an intonation rise on a final stressed syllable (with a long vowel) it means “is that true”, with interest or surprise, but if there is stress on the first syllable (with accent II) and falling intonation on both syllables it will mean “no I won’t do it” expressing displeasure. Saying that “nähä” is polysemous is to take a written language perspective on word meanings. Some words simply cannot be defined without hearing them spoken.

This paper explores the possibilities of studying emotional expression in the Gothenburg Spoken Language Corpus (GSLC), see Allwood et al (2002). It will mainly treat the use of emotion adjectives and listeners’ interpretations of emotional prosody.

The research questions are:

Which emotions do we usually express in communication with each other, with prosody and with words?

How does use of emotional expressions differ in different social activities?

What is the interaction between lexicon and prosody?

The question of the nature of human emotions has been discussed to a great extent (see e.g. Cornelius, 2000, Coddy & Cornelius, 2003, Scherer, 2003). Speakers use words, prosody, body language or even syntax to express emotions; in this article prosody and words will be studied. Joy, anger, surprise, fear, sadness and disgust have often been described as basic emotions or the big six (cf Ekman, 1992), implying that these emotions are in some way more fundamental than others.

We can express emotions with adjectives: I am happy, angry, sad, afraid, surprised, tired and so on. But often we express the emotions with prosody.

Method

The methods used were the following: Emotional expressions were searched for in Gothenburg Spoken Language Corpus GSLC (www.ling.gu.se/~leifg/tal) with a web based tool, producing e. g. frequencies and concordances. Apart from the transcriptions the corpus contains a rough coding of the interlocutors’ different emotional states, as especially noted by transcribers. (Coding emotions really well is of course a much more complex process, see e. g. Devillers et al, 2005, 2006.) The emotion codings are marked as ”mood:” in an information line placed under the transcription. Searches were made with the corpus browser for e.g. ”mood:” or ”mood: surprised”. This was done in order to study emotions expressed with prosody. In order to study emotions expressed with adjectives, words for basic emotions were searched in the transcription line, e.g. ”förvånad” (surprised). Analyses were made with respect to different social activities, e.g. ”discussion” or ”travel agency negotiation” (these are 3 out of 25 social activities in the corpus). Analyses of the relation between words and interpreted mood were made. Finally, preliminary acoustic analyses of some emotional expressions were also made. Since the corpus is not very large, especially when broken down into different activities, the study should be seen as exploratory.

Results

Frequencies of different interpreted emotions

Which emotions are found, in the interpreted mood line, expressed prosodically? In comparison to the number of words in the material, there are rather few instances of different emotions noted by the transcribers. Nevertheless it is interesting to study the ones noted, to see how they are distributed over different emotions, different activities and how they combine with different expressions. There are 566 ”mood”, of which 235 express the emotions in Table 1. The remaining emotions are distributed over singular occurrences.

Table 1. Frequencies for the emotions marked in the “mood” line. Basic emotions in italics.

Emotion	Frequency
surprised	80
hesitating	56
irritated	35
happy	28
uncertain	18
sad	7
angry	5
amused	6
disgust	0
fear	0

The basic emotions, the big six, are marked in italics in Table 1. There is evidence for four of the six basic emotions, *surprised*, *happy*, *angry* and *sad*. There is also strong evidence for other emotions, *hesitating* and *irritated*. In fact only the basic emotion *surprised* is more common than these two.

Frequencies of emotional adjectives

Looking at the actual transcription we can see which of the basic emotions are expressed with emotion adjectives in the conversations. The reason to choose the basic emotion adjectives is to have a starting point, because emotion words in the transcriptions must be searched by word, not as a group.

Table 2. Emotion adjectives in the transcriptions.

Emotion	Frequency
rädd (afraid)	193
glad (happy)	48
arg (angry)	36
förvånad (surprised)	21
avskyr (detests)	3
ledsen (sad)	2

Comparison of emotions expressed with prosody or with adjectives

Now we can compare the frequencies of mood-coding (prosody) with the frequencies of the adjectives. Anger and fear are not frequent in the mood-coding, i.e. usually not expressed prosodically. But anger and fear are expressed or talked about with words (usally in the past or future tense). Surprise is expressed much more frequently with prosody than with words. Joy is expressed either with words or prosody (or both).

Frequencies of emotional prosody in different social activities

The next question is how common prosodically expressed emotions – ”moods”, and emotional adjectives are in different activities. Table 3 shows the frequencies for interpreted emotions ”moods” in different social activities.

Table 3. Frequencies for prosodically expressed emotions in different social activities.

Travel agency	Lecture	Discussion
happy 13	feigning anger 1	surprised 22
irritated 6	feigning humility 1	sarcastic 12
surprised 6	ironic 1	hesitating 8
uncertain 2	jokingly solemn 1	upset 5
cheerful 2		happy 3
doesn't like 1		accusing 3
upset 1		emphatic 3
amused 1		coldly 2
embarrassed 1		irritated 5
concerned 1		

Table 3 can be summarized in the following way: travel agency dialogues display more *happiness*, lectures show no real emotions, and discussions contain more *surprise*. An explanation for the differences between the acitvities could be that travel agency interaction is interaction between salesmen and customers which try to show positive emotions and friendliness towards each other in order to sell/buy, the lectures convey no emotions since lectures have an informative function and is mainly one-way communication.

Mood in relation to carrier phrase

The occurrence of different emotional prosody in relation to carrier words was also analyzed. The question is whether emotions are expressed simultaneously in different modalities, e.g. with the words *I am angry*, simultaneously expressed with angry prosody, or whether prosody expresses emotions independently of the lexical content of the words. The findings, in Tables 4 and 5, suggest that prosody can combine with lexical information or be independent of it. Further studies will treat the strength of the prosodic expression in relation to lexical content. (cf. Dubost & Su, 1999, on Mandarin Chinese).

Surprised

The material is quite small if divided into both emotions and social activities, so we will look at all three activities, first at the emotion surprised.

Table 4. Mood: surprised in relation to carrier phrase, in all activities. Utterances occurring more than once are listed.

Utterance	Frequency
jaså (is that so?)	10
jaha (oh well)	6
aha (I see)	2
va (what)	2
nähä (is that true)	2
nä (is that true)	2
with question word order	17

The most common expressions marked as mood: surprise in all activities is *jaså*, followed by *jaha*. The meaning dimension ”unexpected new information” existing in surprise, exists as well in jaha (oh well), oj oj oj (oh dear). It could exist in hallå (hallo) depending on the prosody, again an example of a word which cannot really be semantically defined without its emotional prosody. The meaning dimension of ”unexpected new information” in surprise is also found in *jaså* (is that so?), *va* (what), *aha* (I see). It can also be found in nä (no) depending on the prosody. The meanings of *nähä* and *nä* depend on the prosody.

Happy

24 different utterances were marked once as mood: happy. The meaning dimension ”positive” exists in many of the words, e.g. in *tack* (thanks) (2) and *bra* (good) (2), älskling (darling) and is not contradictory to the meanings of *hej* (hi) (3), *joo* (yes), *ja* (yes). In some of the words expressed with happiness there is no general positive meaning dimension, like *fishing hat* or *week.*

Irritated and uncertain

The emotions irritated and uncertain were also studied in all activities. 35 utterances were marked as ”mood: irritated” The only one occurring more than once is: ah (ah!). 10 utterances were marked as ”mood: uncertain” The utterances occurring are: mhm (feedback), så där (so so), jaa (yes), utveckling (development), har jag sagt fel nu (did I say something wrong now), hm (”thinking”).

Generally, there is a relation between coded, perceived, emotion and emotional dimension in the lexical items in many cases, but in many other cases not. Examples of a relation are *jaså (is that so?)* with surprise or *bra* (good) with happiness. Examples of no relation are *fiskemössa* (fishing hat) with happiness or *radion* (the radio) with surprise.

Utterances ocurring with different moods

Turning the perspective around, taking the point of departure in which utterances occur for several of these four ”moods”, we find the following:

Table 5. Utterances occurring for several of the four “moods”.

Utterance	Mood
ja (yes)	uncertain, happy
jaa (yes)	uncertain, happy
mhm (yes, feedback)	uncertain, happy, surprised
nä/nej	irritated, surprised

So, some words (feed back words), are to a large extent compatible with different emotional prosodies.

Emotional adjectives ocurring in different social activities

The adjectives listed in Table 2 can also be studied in relation to activities. In the three activities travel agency, lecture and discussion, the adjectives studied above: glad (happy), förvånad (surprised), arg (angry) and rädd (afraid) are all common in discussions but almost absent in Travel agency and Lecture. The most common one was *arg* (angry). See Table 6.

Table 6. Adjectives used in three different social activities.

	Travel agency	Lecture	Discussion
glad (happy)	0	0	11
förvånad (surprised)	1	0	9
arg (angry)	0	1	21
rädd (afraid)	0	0	17

There were thus frequency differences of emotion adjectives for different activities: discussions contain many emotion adjectives while travel agency and lecture contain almost none. Compared with Table 3 for emotional prosody in different activities, the biggest difference is in the activity Travel agency: here emotional expression is conveyed by other means than by adjectives. In lectures there are hardly any emotional expressions, neither with adjectives or prosody and in discussions emotions are expressed with prosody as well as with emotional adjectives.

Preliminary acoustic analyses

Preliminary acoustic analyses were made in order to check the quality of the recordings, which were all made in naturalistic settings. The utterances to be analyzed e.g. hej, ja, mhm, nej (cf. Table 5) were found in the GSLC with the help of the corpus browser (searching for word and interpreted mood). Then the transcription and recording numbers were identified and the utterances to be analyzed were copied from the recordings. The recordings were noisy but there were no problem extracting the F0 curves. The F0 of hej (hi), interpreted as happy, uttered in one travel agency dialogue, is shown in Fig. 1.

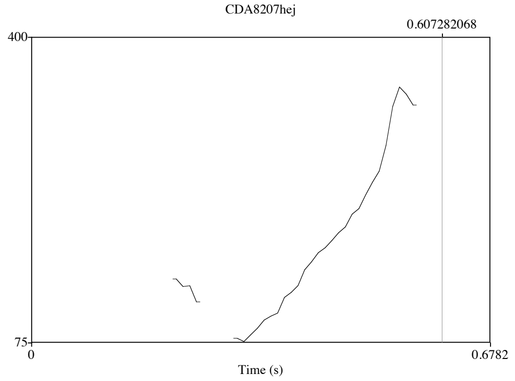


Figure 1: *F0 curve of happy hej (hi).*

The F0 curve of ja (yes) interpreted as happy, uttered in one travel agency dialogue, is shown below in Fig. 2.

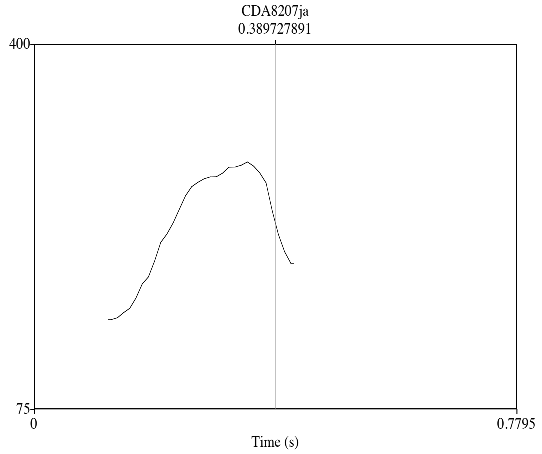


Figure 2: *F0 curve of happy ja (yes).*

The diagrams show the (expected) rising F0 curves with good quality. The program used was Praat 5.0.22.

Summary and discussion

The most common emotion expressed by prosody is surprise. Thereafter comes hesitation and irritation. In fourth place we find happiness. Almost no talker expressed fear or real anger, as experience by the interpreters. There were also frequency differences of emotions for different activities: discussions contain more surprise, travel agency dialogues display more happiness and lectures showed no emotions. This could be due to communicative needs and display rules.

The occurrences of adjectives for the basic emotions, the big six, were as follows: rädd (afraid) was most commonly used. Thereafter comes glad (happy), but almost no one uses the word ledsen (sad). There were also frequency differences of emotion adjectives for different activities: discussions contain many emotion adjectives while travel agency and lecture contain almost none. Sometimes lexical and prosodic content seem to interact, sometimes not. Some examples are the following, where the same semantic dimensions can be found in the meaning of the word and the emotion interpreted:

jaså (is that so) – surprised
bra (good) – happy
har jag sagt fel nu (did I say something wrong) – uncertain

But, on the other hand, in the following examples we see no common semantic dimension between the word and the interpreted emotion: fiskemössan (fishing hat) – happy, radion (the radio) – surprised.

Certain utterances are used for many emotions:

utterance	”mood”
ja	uncertain, happy
jaa	uncertain, happy
m(hm)	uncertain, happy, surprised
nä/nej	irritated, surprised

So, some words (feed back words), are to a large extent used with different emotional prosody. Acoustic analyses of the prosody in the recordings is being initiated.

Corpus studies can give us knowledge about:

- Which emotional expressions are used in which situations
- How often they are used
- How we express the emotions – with prosody, words etc.
- How prosody and words interrelate.
- The actual manifestation of emotional prosody in natural spoken interaction

Acknowledgements

The program used for the acoustic analysis was Praat 5.0.22.

References

Allwood J, Grönkvist L, Ahlsén, E, Gunnarsson M (2002). Göteborgskorpusen för talspråk (The Gothenburg Spoken Language Corpus, GSLC). *Nydanske studier*, Copenhagen: Akademisk Forlag, 39-58.

Boersma P (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10: 341-345.

Cornelius R R (2000). Theoretical approaches to emotion. *Proceedings of the ISCA Workshop on Speech and emotion*. Ireland, 3-10.

Cowie R, Cornelius R R (2003). Describing the emotional states that are expressed in speech. *Speech Communication* 40: 5-32.

Devillers L, Cowie R, Martin J-C, Douglas-Cowie E, Abrilian S, McRorie M (2006). Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. *5th international conference on Language Resources and* Evaluation (LREC 2006), Genoa, Italy, 1105-1110.

Devillers L, Vidrascu L (2006). Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs. *Proceedings of Ninth international conference on spoken language processing (ICSLP)*, Pittsburgh, PA, USA, 801-804.

Dubost J-M, Su T (1999). Prosodic differences and similarities between Mandarin and French in declarative, interrogative, surprise and doubt expressions. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 1561-1564.

Ekman P (1992). An argument for basic emotions. *Cognition and emotion* 6: 169-200.

Scherer K (2003). Vocal communication of emotion: a review of research paradigms. *Speech communication* 40 (1): 227-256.

www.ling.gu.se/~leifg/tal

Research focus: Interactional aspects of spoken face-to-face communication

Jonas Beskow, Jens Edlund, Joakim Gustafson, Mattias Heldner, Anna Hjalmarsson, David House
KTH Speech, Music and Hearing

Abstract

We have a visionary goal: to learn enough about human face-to-face interaction that we are able to create an artificial conversational partner that is human-like. We take the opportunity here to present four new projects inaugurated in 2010, each adding pieces of the puzzle through a shared research focus: interactional aspects of spoken face-to-face communication.

Introduction

We have formulated a visionary goal: *to learn enough about human face-to-face interaction that we are able to create an artificial conversational partner that is human-like,* or as Cassell (2007) puts it, ”acts human enough that we respond to it as we respond to another human”.

Note that this is not a case of implementing science fiction – it is about testing scientific hypotheses. Paraphrasing the methods used in cognitive science, we implement models of human behaviour and put these to the test in interactions with humans. This is how we validate our understandings.

While our dream of human-likeness may eventually turn out to be overly ambitious, it has already been useful to us by guiding our research efforts towards black holes in our knowledge about human conversation. It is for example evident that state-of-the-art speech technology neither sounds like a conversational partner, nor understands fundamental aspects of human conversational behaviour.

A massive amount of cross-disciplinary research is needed to realize the visionary goal. Although we have a long way to go, we learn useful lessons from every step. We are currently involved in a half-dozen national and international research projects orchestrated towards our goal.

In this paper, we take the opportunity to give a brief overview of four new projects with national funding sharing a common research focus: *interactional aspects of spoken face-to-face communication*.

Background

Speech in conversation differs from speech in situations where there is no conversational partner present, for example read-aloud speech. This becomes apparent when listening to people acting or reading dialogues aloud. We can often tell that a conversation is acted, even if the script is meticulously written with all imaginable detail. Something special happens when there is someone to interact with, and what is being said is planned there and then.

Ironically, we know less about the primary use of speech – in face-to-face communication – than about many other kinds of speech (see e.g. Heldner & Edlund, 2007). This is, among other things, reflected in how the talking computers we encounter today behave. For one thing, they simply do not sound like they are having a conversation even if they say the same thing a human in the same situation would. In addition, face-to-face conversation involves other sources of information, perhaps most notably a visual channel with gaze, nods, other gestures, posture, proxemics etc. that forms an intrinsic part of the communication. Thus, further investigations about how humans converse are motivated from basic and applied research perspectives alike.

An initial requirement for making substantial progress is access to spontaneous conversations. We have recently collected about 60 hours of audio, video and motion capture data in conversations within the project *Spontal: Multimodal database of spontaneous speech in dialog* (Edlund, et al., 2010). The projects presented here all explore this dataset.

Current projects

The following is a brief overview of four new projects within the research theme interactional aspects of spoken face-to-face communication.

Prosody in conversation

Riksbankens Jubileumsfond (RJ) has granted the project *Prosody in conversation (Samtalets prosodi)* 5.2 MSEK for the years 2010-2012 (contract P09-0064:1-E). Applicant: Mattias Heldner.

The project investigates how people talking to each other jointly decide *who should speak when*, and the role of prosody in making these joint decisions. While prosody is by no means the only relevant information for this joint interaction control, we believe that it plays a central role (see e.g. Edlund & Heldner, 2005, and references mentioned therein). A detailed model of the prosody involved in interaction control is crucial both for *producing* appropriate conversational behaviour and for *understanding* human conversational behaviour. Both are required in order to reach our visionary goal, and represent the artificial conversational partner in the roles of speaker and listener in a conversation, respectively.

One line of inquiry within the project is the quantitative acoustic analysis of prosodic features in genuine spoken face-to-face conversations. The project focuses on local intonation patterns in the immediate vicinity of interactional events, such as transitions from (i) speech to pauses (within-speaker silences); (ii) speech to gaps (between-speaker silences, i.e. at speaker changes); and (iii) speech by one speaker to speech by another speaker (i.e. overlapping speech in speaker changes). In addition, we analyze selected interactional phenomena occurring on a longer time scale, such as pitch similarity across these interactional events and the overall tendency of interlocutors to grow increasingly similar as the conversation proceeds. This increasing interlocutor similarity reported in the literature under many names (e.g. entrainment, alignment, accommodation; see e.g. Edlund, Heldner, & Hirschberg, 2009 for an overview) has been reported for a great number of linguistic features, but we limit ourselves to prosody in this project.

In addition, the results of the acoustic analyses are fed into a second line of inquiry: studies of the effects of using or introducing such prosodic features in a conversation. These studies will include listening experiments where manipulations of genuine conversations by means of re-synthesis are used as stimuli. Furthermore, there will be pragmatic experiments where the conversational behaviour in response to the use of such prosodic features in artificial speech is analyzed. Finally, there will be analyses of conversational behaviour in response to real-time (or minimum delay) manipulations of genuine conversations, such as deletions, insertions or manipulations of features.

The rhythm of conversation

The Swedish Research Council (VR) HS, has granted the project *Rhythm of conversation (Samtalets rytm)* 2.9 MSEK for the years 2010-2012 (contract 2009-1766). Applicant: Mattias Heldner.

The project *Rhythm of conversation* investigates how a set of rhythmic prosodic features contributes to the joint interaction control in conversations. Of particular interest is acoustic descriptions of features related to variations in speech rate (i.e. accelerations and decelerations in speech rate) and loudness (i.e. increases and decreases in loudness), and how these are used for interactional purposes.

Loudness is generally perceived as an important component in the signalling of prosodic functions such as prominence and boundaries (cf. Lehiste & Peterson, 1959). Attempts to capture this impression in acoustic analyses, however, regularly show only weak correlations with these functions (e.g. Fry, 1955; Lieberman, 1960). This has lead much prosodic research to concentrate on melodic prosodic aspects – sometimes to the extent that prosody is equated with pitch. Recent work indicates, however, that loudness may be a strong correlate of such functions, when measured as subjective loudness rather than as physical intensity (Kochanski, Grabe, Coleman, & Rosner, 2005). This is highly unexplored and something we pursue in connection with rhythm as an interactional phenomenon.

We want to find out, for example, whether the speech rate and loudness variations (prosodic features that are complementary to those studied in *Prosody in conversation*) before pauses (i.e. within-speaker silences) are different from those before gaps (between-speaker silences), or whether they display differences before backchannel-like utterances compared to other utterances.

Introducing interactional phenomena in speech synthesis

The Swedish Research Council (VR) NT, has granted the project *Introducing interactional phenomena in speech synthesis (Talsyntes för samtal)* 2.1 MSEK for the years 2010-2012 (contract 2009-4291). Applicant: Joakim Gustafson.

The project recreates human interactional vocal behaviour in speech synthesis in three phases. The first deals with what Allwood (1995) calls feedback morphemes and Ward (2000) conversational grunts (e.g. mm, eh). We also include audible breathing, following Local & Kelly (1986) who hold breath as a strong interactional cue. These tokens are traditionally missing in speech synthesis. We remedy this by (1) annotating instances of them in the Spontal corpus (and possibly other corpora), (2) synthesizing the missing tokens using several methods, and (3) evaluating the results in a series of experiments comparing synthesized versions with the originals as well as evaluating their perceived meaning and function.

The second phase is similarly structured, but targets events that occur in the transitions between speech and silence and back – transitions that vary depending on the situation. We focus on three transition types: *normal* (the target of current syntheses), *hesitant* and *abrupt*. Pauses and retardations are strong cues for hesitation, and glottal stops or unreleased plosives are frequently a result of an interruption (Local & Kelly, 1986). Speech that has been halted on a glottal stop or an unreleased plosive can be restarted by releasing the stop. This gives rise to different acoustic effects which we recreate and evaluate.

In the third phase, we evaluate reactions to a dialogue system making use of the synthesized cues developed in the first two phases. In semi-automatic dialogue systems modelling speaking and listening as parallel and mutually aware processes, we use two scenarios to verify and validate our results: the attentive speaker – an interruptible virtual narrator making use of synthesized cues for hesitation and end-of-contribution; and the active listener – an information gathering system, aiming to encourage the user to continue speaking (cf. Gustafson, Heldner, & Edlund, 2008).

Intonational variation in questions in Swedish

The Swedish Research Council (VR) HS, has granted the project *Intonational variation in questions in Swedish (Variation i frågeintonation i svenska)* 2.6 MSEK for the years 2010-2012 (contract 2009-1764). Applicant: David House.

The project investigates and describes phonetic variation of intonation in questions in spontaneous Swedish conversation, with an initial premise that there does not exist a one-to-one relationship between intonation and sentence type (Bolinger, 1989). The Spontal database is used to find a general understanding of the role of questions in dialogue and an explanation of why descriptions of question intonation has proven so difficult. We expect to find certain patterns of intonation that correlate with for example dialogue and social function.

We will test several hypotheses from the literature. One example is the hypothesis that there is a larger proportion of final rises and high pitch in questions which are social in nature than in those which are information oriented. Another example concerns the type of visual gestures that accompany questions (McNeill, 1992): we will test the hypothesis that iconic and emblematic gesture types co-occur more often with information-oriented questions while dialogue gestures such as nods and facial gestures co-occur more frequently with social-oriented questions.

Finally, our results will be analyzed within the framework of biological codes for universal meanings of intonation proposed by Gussenhoven (2002). Gussenhoven describes three codes, or biological metaphors: a frequency code, originally proposed by Ohala (1983), implying that a raised F0 is a marker of submissiveness or non-assertiveness and hence question intonation; an effort code, in which higher F0 requires increased articulation effort which highlight important focal information; and a production code associating high pitch with phrase beginnings (new topics) and low pitch with phrase endings. A biological explanation for the pragmatic functions of intonation is of important theoretical interest for the project, and leads further into investigations of the relationships between intonation and visual gestures in a framework of biological codes.

Summary

We have proposed an ambitious and visionary goal for our research: to learn enough about human face-to-face interaction that we are able to create an artificial conversational partner that is human-like in the sense that people interacting with it respond to it as they do to other humans. This visionary goal has been instrumental in the prioritization and formulation of a current research focus for our group: investigations of interactional aspects of spoken face-to-face communication. We have described four new externally funded projects that are representative of and will advance the research frontier within this common research focus.

While these projects do not in themselves have either the resources or the scope to reach our visionary goal, they each add a piece of the puzzle, and we are confident that they will help identify future areas for research contributing towards the long-term goal. The visionary goal requires a wider scoped platform for future grant applications. The joint effort of these projects forms a strong point of departure by providing critical mass of research expertise in the area.

Acknowledgements

This research is carried out at KTH Speech, Music and Hearing. Funding was provided by Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation;* the Swedish Research Council (VR) projects 2009-1766 *The rhythm of conversation*, 2009-4291 *Introducing interactional phenomena in speech synthesis*; 2009-1764 *Intonational variation in questions in Swedish*; and 2006-7482 *Spontal: Multimodal database of spontaneous speech in dialog*.

References

Allwood, J (1995). An activity based approach to pragmatics. In: H Bunt & B Black, eds, *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam, The Netherlands: John Benjamins.

Bolinger, D (1989). *Intonation and its uses: Melody in grammar and discourse*. London, UK: Edward Arnold.

Cassell, J (2007). Body language: Lessons from the near-human. In: J Riskin, ed, *Genesis Redux: Essays in the History and Philosophy of Artificial Life*. Chicago, IL, USA: The University of Chicago Press, 346-374.

Edlund, J, Beskow, J, Elenius, K, Hellmer, K, Strömbergsson, S, & House, D (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In *Proceedings of LREC 2010*. Valetta, Malta.

Edlund, J, & Heldner, M (2005). Exploring prosody in interaction control. *Phonetica,* 62: 215-226.

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

Fry, D B (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America,* 27: 765-768.

Gussenhoven, C (2002). Intonation and interpretation: phonetics and phonology. In: B Bel & I Marlien, eds, *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence, France, 47-57.

Gustafson, J, Heldner, M, & Edlund, J (2008). Potential benefits of human-like dialogue behaviour in the call routing domain. In *Perception in Multimodal Dialogue Systems*. Berlin, Germany: Springer, 240-251.

Heldner, M, & Edlund, J (2007). What turns speech into conversation? A project description. In *TMH-QPSR 50: Fonetik 2007*. Stockholm, Sweden, 45-48.

Kochanski, G, Grabe, E, Coleman, J, & Rosner, B (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America,* 118: 1038-1054.

Lehiste, I, & Peterson, G E (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America,* 31: 428-435.

Lieberman, P (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America,* 32: 451-454.

Local, J K, & Kelly, J (1986). Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies,* 9: 185-204.

McNeill, D (1992). *Hand and mind – What gestures reveal about thought*. Chicago, IL, USA: University of Chicago Press.

Ohala, J J (1983). Cross-language use of pitch: an ethological view. *Phonetica,* 40: 1-18.

Ward, N (2000). The challenge of non-lexical speech sounds. In *Proceedings of ICSLP 2000*. Beijing, China, 571-574.

Very short utterances in conversation

Jens Edlund[1], Mattias Heldner[1], Samer Al Moubayed[1], Agustín Gravano[2] and Julia Hirschberg[3]
[1] KTH Speech, Music and Hearing
[2] Computer Science Department, University of Buenos Aires
[3] Department of Computer Science, Columbia University

Abstract

Faced with the difficulties of finding an operationalized definition of backchannels*, we have previously proposed an intermediate, auxiliary unit – the* very short utterance *(VSU) – which is defined operationally and is automatically extractable from recorded or ongoing dialogues. Here, we extend that work in the following ways: (1) we test the extent to which the VSU/NONVSU distinction corresponds to backchannels/non-backchannels in a different data set that is manually annotated for backchannels – the Columbia Games Corpus; (2) we examine to the extent to which VSUs capture other short utterances with a vocabulary similar to backchannels; (3) we propose a VSU method for better managing turn-taking and barge-ins in spoken dialogue systems based on detection of backchannels; and (4) we attempt to detect backchannels with better precision by training a backchannel classifier using durations and inter-speaker relative loudness differences as features. The results show that VSUs indeed capture a large proportion of backchannels – large enough that VSUs can be used to improve spoken dialogue system turntaking; and that building a reliable backchannel classifier working in real time is feasible.*

Introduction

A large number of vocalizations in everyday conversation are traditionally not regarded as part of the information exchange. Examples include confirmations such as *yeah* and *ok* as well as traditionally non-lexical items, such as *uh-huh*, *um*, and *hmm*. Vocalizations like these have been grouped into different categories and given different names: for example *backchannels* (i.e. back-channel activity, Yngve, 1970), *continuers* (Schegloff, 1982), *feedback* and *grunts*, and attempts at formalizing their function and meaning have been made (e.g. Ward, 2004). The working definitions of these overlapping concepts, however, are imprecise, and different labeling schemes treat them quite differently. The schemes are also often complex. Faced with these difficulties and inspired by others, for example Shriberg et al. (1998), we previously proposed an intermediate, auxiliary unit – the *very short utterance* (VSU) – which is defined operationally and is automatically extractable from recorded or ongoing dialogues (Edlund, Heldner, & Pelcé, 2009). VSUs are intended to capture a large proportion of the interactional dialogue phenomena commonly referred to as backchannels, feedback, continuers*, inter alia*, at zero manual effort.

VSUs and backchannels

The data we used for our first examination of VSUs, however, was not annotated for backchannels, and automatically identified VSUs were instead compared to annotation for degree of informational content of the same utterances, under the assumption that utterances with low informational content would be representative for backchannels. The first contribution of this paper is to report the extent to which the distinction of VSU/NONVSU, as defined in (Edlund, et al., 2009) captures the distinction between backchannels and non-backchannels as annotated in the Columbia Games Corpus. We also include a more fine-grained analysis of the non-backchannels captured by the VSUs.

Although VSUs may be a useful compromise when no manual annotation is available, we would ideally like to be able to do without them and detect backchannels and other feedback directly. As a first step towards this, we also train a classifier of backchannels on duration and inter-speaker relative loudness and report the preliminary results.

VSUs and spoken dialogue systems

In spite of the difficulties involved in defining backchannels, there is little controversy surrounding the utility of modeling them. They behave differently from other utterances, and so are interesting both for models of human conversation and for spoken dialogue systems aiming at more human-like behavior. Commonly reported characteristics include the fact that they can be spoken in overlap without disrupting the original speaker (hence the term ‘backchannel’). For a spoken dialogue system designer to build systems that encourage users to talk to a system as they would to another human, this phenomenon needs to be managed so that such a system can receive continuous feedback from its users, often in the form of backchannels. As most systems today deal with any user vocalization occurring during system speech as if it were a barge-in, causing the system to stop speaking immediately, the effects of such feedback to current systems would be peculiar and unwanted. The second contribution of this paper is to propose a method to better deal with turntaking in spoken dialogue systems by continuously detecting VSUs and to quantify the potential gain of using such a method.

Method

Columbia Games Corpus

The data used in this work is drawn from the Columbia Games Corpus (CGC), a collection of spontaneous task-oriented dialogues by native speakers of Standard American English, and its associated annotations. This corpus contains recordings made using close-talking microphones, with speakers recorded on separate channels, 16 bit/48 kHz, in a sound-proof booth. Speakers were asked to play two types of collaborative computer games that required verbal communication. The speakers did not have eye contact. There were 13 subjects (7 males and 6 females) and they formed 12 different speaker pairs. Eleven of the subjects spoke with two different partners in two separate sessions. The recording sessions lasted on average 45 minutes, and the total duration of the corpus is 9 hours 8 minutes.

The corpus has been orthographically transcribed and manually annotated for a number of phenomena. For the present study, we have used the labeling of single *affirmative cue words* (i.e. lexical items potentially indicating agreement such as *alright, gotcha, huh, mm-hm, okay, right, uh-huh, yeah, yep, yes, yup*) with their communicative function, by three trained annotators, and the labeling of turn-exchanges, by two trained annotators. The function labels for affirmative cue words are *backchannel*, *affirmation/agreement, cue phrase beginning discourse segment, cue phrase ending discourse segment, pivot beginning* and *pivot ending*. Turn exchanges were labeled by first identifying *Interpausal Units* (IPUS), maximal sequences of words surrounded by silence longer than 50 ms (cf. talkspurts in Brady, 1968). A turn was defined as a maximal sequence of IPUS from a single speaker, so that between any two adjacent IPUS there is no speech from the interlocutor (cf. talkspurts in Norwine & Murphy, 1938).

All turn transitions in the corpus were classified using a labeling scheme adapted from (Beattie, 1982) that identifies, inter alia, *smooth switches* (S) — transitions from speaker A to speaker B such that (i) A manages to complete her utterance, and (ii) no overlapping speech occurs between the two conversational turns; *pause interruptions* (PI), defined as cases similar to smooth switches except that A does ***not*** complete her utterance; and *backchannels* (BC), defined as an utterance produced a “response to another speaker’s utterance that indicates only *I’m still here / I hear you and please continue*”, with no attempt to take the turn. Speech from A following backchannels from B was labeled separately as X2[1].

Data

For the present study, we used the annotations of turn transitions in silences in the Columbia Games Corpus. We contrasted backchannels with a collapsed non-backchannel category including smooth switches, pause interruptions and utterances following backchannels (S+PI+X2). In addition, we contrasted backchannels with a collapsed category including all other single affirmative cue words (AFFCUE). The backchannel category (BC) in both comparisons was identical, while the other discourse functions of affirmative cue words comprised a subset of the smooth switches plus pause interruptions category.

[1] http://www.cs.columbia.edu/speech/games-corpus/ has further details and annotation manuals.

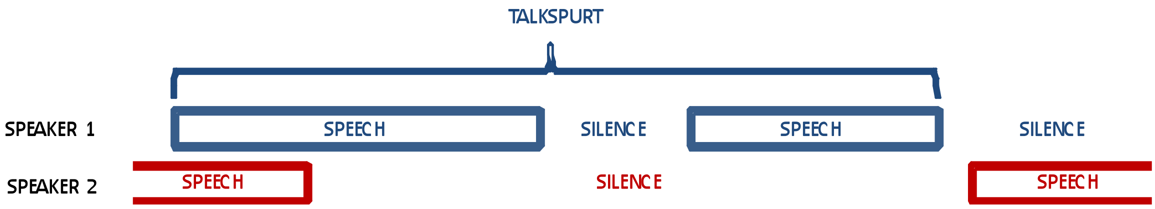


Figure 1. Schematic illustration of a talkspurt as used in the current work.

Talkspurt durations

We are interested in how long a speaker of a backchannel or a non-backchannel goes on speaking, on average, until the other speaker takes the turn. For this, we need durations of the talkspurt defined by Norwine & Murphy: "A *talkspurt* is speech by one party, including her pauses, which is preceded and followed, with or without intervening pauses, by speech of the other party perceptible to the one producing the talkspurt" (Norwine & Murphy, 1938), or a *turn* in CGC. We note that this definition differs from that used by Brady (1968), in which a talkspurt is a sequence of speech activity flanked by silences in one speaker's channel. Brady's definition has been used by ourselves in previous work (e.g. Edlund & Heldner, 2005; Laskowski, Edlund, & Heldner, 2008), but Norwine & Murphy's concept is better suited for our current purposes, and we adopt their definition in what follows.

Identifying VSUs

The objective of automatically defining VSUs draws on the observation that backchannels are limited in duration and quiet. Thus, we extracted DURATION from start to finish (see Figure 1) for each talkspurt. A talkspurt was classified as a VSU if the talkspurt's DURATION was shorter than a given threshold. In addition, we extracted loudness differences across turn exchanges using the method and frequency weighting proposed by the ITU (International Telecommunication Union, 2006). In (Edlund, et al., 2009) we also used voicing ratio which helped filter out mistakes made by the automatic speech detector. As the CGC data is manually annotated, we left the voicing ratio parameter out for simplicity.

Backchannel classifier training

For training, the K-Nearest Neighbors (K-NN) method was used. K-NN is a non-parametric non-linear classifier which does not build a model for training data, but builds a local model for each test sample using that sample's neighborhood. The study in (Atkeson, Moore, & Schaal, 1997) gives a good overview of the method. Initially, the K-nearest neighbors to the test sample are collected using Euclidian distance on the features. Then, a weighted average voting of these neighbors decides which class the test sample belongs to. In our case we use a binary classification of BC/NONBC, and the method would give a number between 0 and 1, which can be taken as the probability of the class 0 or 1. Classifiers were trained using duration only (DUR) as well as duration and inter-speaker loudness difference (DUR+LOUDDIFF).

Results

Table 1 shows the confusion matrix for VSU/NONVSU versus the manual annotation of BC/NONBC using the same threshold for VSU as in our previous study: 1 s. We note in particular that all BCs in the material are also VSUs with this threshold. Figure 2 shows the underlying data – the histograms over the durations of BC and NONBC. Table 2 shows a cross tabulation of VSU/NONVSU (again using a 1 s threshold) versus the manual annotation of BC/AFFCUE/OTHER, and the underlying data for this three-way split – the histograms over the durations of BC, AFFCUE and NONBC appear in Figure 3.

Table 1. Confusion matrix for VSU/NONVSU versus the manual annotation of BC/NONBC.

	VSU	NONVSU	TOTAL
BC	553	0	553
NONBC	1208	2600	3808
TOTAL	1761	2600	4361

Table 2. Cross tabulation of VSU/NONVSU versus a manual annotation of BCs versus AFFCUE and OTHER.

	VSU	NONVSU	TOTAL
BC	553	0	553
AFFCUE	699	768	1467
OTHER	509	1832	2341
TOTAL	1761	2600	4361

We note that 31% of the VSUs are backchannels and 40% are AFFCUES. Inspection of the remaining 29% of VSUs labeled as OTHER revealed that a large proportion of them also have feedback functions. The 25 most frequent tokens are *mm*, *no*, *oh*, *got it*, *um*, *oh okay*, *hm*, *I'm gonna pass*, *uu*, *I have to pass*, *cool*, *great*, *nope*, *and*, *sure*, *that's it*, *and then*, *don't have it*, *exactly*, *I don't have that*, *oh right*, *oh yeah*, *so*, and *sorry*. These comprise about one third of all OTHER VSU tokens.

Moving on to the BC/NONBC classifiers, we observe that the classifiers trained on duration only (DUR) and duration plus relative loudness (DUR+LOUDDIFF) showed similar performance, with a slight advantage for the combination of the two. Figure 4 shows the ROC curves for the DUR classifier and the DUR+LOUDDIFF classifier.

The DUR+LOUDDIFF classifier was applied using a 77% training set, and 23% test set size, resulting in 1000 test samples (890 NONBC, and 110 BC). Using an optimized threshold of 0.0995, the overall accuracy of the system resulted in 73% correct classification. Table 3 presents the confusion matrix between the two classes on the test set.

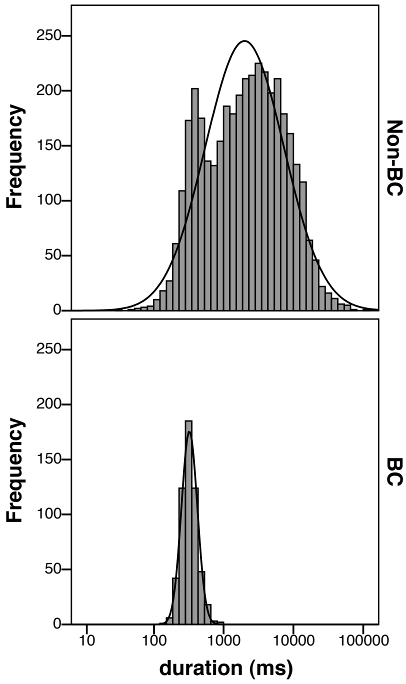


Figure 2. Histograms over durations in ms of manually annotated NONBCs (top) and BCs (bottom) in CGC.

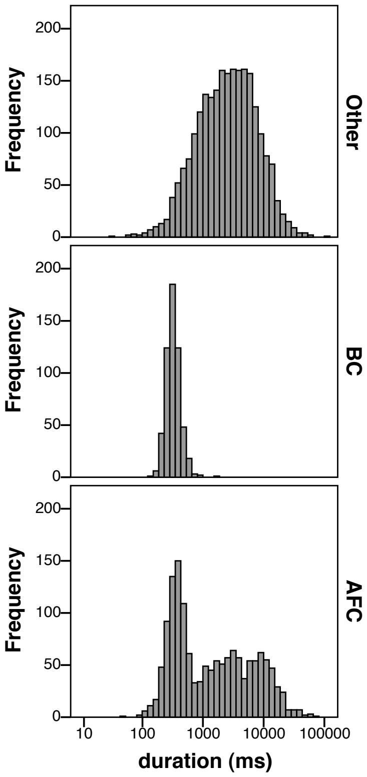


Figure 3. Histograms over durations in ms of manually annotated AFFCUE (bottom), BC (middle) and OTHER (top) in CGC.

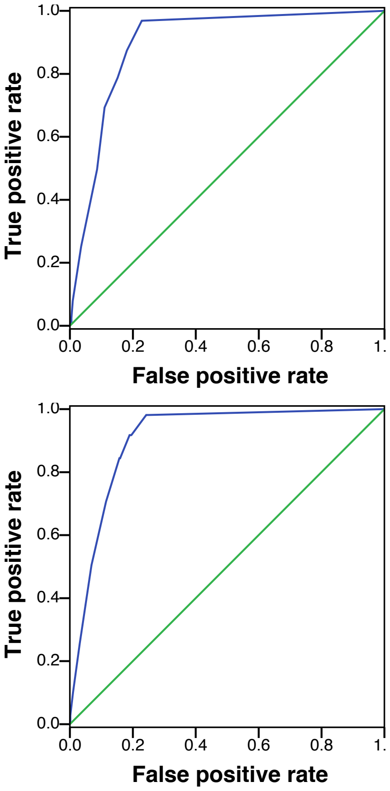


Figure 4. ROC curves for the DUR classifier (top) and the DUR+LOUDDIFF (bottom). The areas under the curves are 0.896 and 0.908, respectively.

Table 3. Confusion matrix between the two classes on the manually annotated test set.

		PREDICTED	
		NONBC	BC
TRUE	NONBC	70.7865	1.8182
	BC	29.2135	98.1818

Discussion

The results in Tables 1 and 2 show that the 1 s duration threshold we previously used to automatically identify VSUs indeed captures a large portion – all, in fact – of the manually annotated backchannels in the Columbia Games Corpus. The duration distributions of BCs and NONBCs, respectively, in Figure 2 suggests that more precise discrimination of backchannels is possible in this material using duration alone, as the majority of backchannels are considerably shorter than 1 s. A shorter threshold would eliminate much of the NONBCs identified as VSUs.

If we keep the 1 s threshold, we note that amongst the talkspurts that are VSUs, yet are not backchannels, many belong to the group of other affirmative cue words, so that the majority (71%) of talkspurts identified as VSUs are either BCs or AFFCUE. Of the remaining 29%, the 25 most frequent words can all be ascribed feedback functions. The 1 s VSUs, in other words, capture almost exclusively short feedback utterances, and using a lower threshold will increase the proportion of manually annotated backchannels.

Although VSUs appear to be a good approximation of backchannels (and other feedback, depending on the duration threshold), we would prefer a classifier that could, in real-time and immediately at the beginning of a talkspurt, identify backchannels. The BC/NONBC classifier is a first step towards this, and its performance suggests that using duration and inter-speaker relative loudness only to train a classifier seems viable. The ROC curves in Figure 4 are promising, and the result of running the combined classifier on unseen test data suggest that BCs can be detected reliably using these features.

Finally, from a speech technology perspective, the histogram in Figure 3 is encouraging. We see that the vast majority of backchannels are shorter than 500 ms, which makes the following strategy for barge-ins tractable:

When user speech is detected during a system utterance, do the following:

- Go on speaking for 300-500 ms.
- If the user has stopped speaking after 300-500 ms has passed, the vocalization was likely a backchannel, so just go on.
- If, on the other hand, the user is still speaking after 300-500 ms, the vocalization is highly unlikely to be a backchannel, so consider stopping (for a polite system) or raising the system’s voice (for urgent messages or for an impolite system).

If we allow for 200 ms to detect silence, approximating the detection thresholds for humans (cf. Izdebski & Shipp, 1978; Shipp, Izdebski, & Morrissey, 1984) as well as for many voice activity detectors (e.g. VADER[2]),

[2] See the CMU Sphinx Speech Recognition Toolkit: http://cmusphinx.sourceforge.net/

the system should be able to make an informed decision at the expense of occasional latencies. On the the Columbia Games Corpus data, such a system would never mistake a backchannel for a barge-in, at the expense of 500-700 ms response delays *occurring only when the user barges in*. This delay corresponds roughly to two or three syllables of speech.

Conclusions

We have shown that the VSU – our previously proposed automatically extractable auxiliary unit – does indeed capture, with zero manual labor, a large proportion of talkspurts annotated as backchannels in the Columbia Games Corpus. Furthermore, a large proportion of those VSUs that are not backchannels are instead different forms of affirmative cue words and other types of feedback which from a dialogue system point of view may be treated in a similar manner. We have trained a BC/NONBC classifier on duration and inter-speaker relative loudness, and found that it finds backchannels with high accuracy and that adding the relative loudness may yield a performance improvement, which is consistent with the claim that backchannels are quiet. This is a first step towards eliminating, at least in part, the intermediate VSU classification.

We have also suggested a method which utilizes the shortness of backchannels to avoid having a barge-in sensitive spoken dialogue system halt abruptly at each backchannel. The cost of this method is acceptable at a latency of some 500-700 ms, applied only where the user speaks at the same time as the system.

We conclude by noting that these findings all point to backchannels being unobtrusive and acoustically not very prominent, and that they are all consistent with descriptions of them as being relatively brief, soft and quiet.

Acknowledgements

This research was carried out at the Department of Computer Science, Columbia University, New York. Funding was provided by the Swedish Research Council (VR) project 2009-1766 *The Rhythm of Conversation*, and by NSF IIS-0307905.

References

Atkeson, C, Moore, A, & Schaal, S (1997). Locally weighted learning. *Artificial intelligence review,* 11: 11–73.

Beattie, G W (1982). Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica,* 39: 93-114.

Brady, P T (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal,* 47: 73-91.

Edlund, J, & Heldner, M (2005). Exploring prosody in interaction control. *Phonetica,* 62: 215-226.

Edlund, J, Heldner, M, & Pelcé, A (2009). Prosodic features of very short utterances in dialogue. In: M Vainio, R Aulanko & O Aaltonen, eds, *Nordic Prosody: Proceedings of the Xth Conference, Helsinki 2008*. Frankfurt am Main: Peter Lang, 57-68.

International Telecommunication Union. (2006). Recommendation ITU-R BS.1770-1: Algorithms to measure audio programme loudness and true-peak audio level.

Izdebski, K, & Shipp, T (1978). Minimal reaction times for phonatory initiation. *Journal of Speech and Hearing Research,* 21: 638-651.

Laskowski, K, Edlund, J, & Heldner, M (2008). An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems. In *Proceedings ICASSP 2008*. Las Vegas, NV, USA, 5041-5044.

Norwine, A C, & Murphy, O J (1938). Characteristic time intervals in telephonic conversation. *The Bell System Technical Journal,* 17: 281-291.

Schegloff, E (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In: D Tannen, ed, *Analyzing Discourse: Text and Talk*. Washington, D.C., USA: Georgetown University Press, 71-93.

Shipp, T, Izdebski, K, & Morrissey, P (1984). Physiologic stages of vocal reaction time. *Journal of Speech and Hearing Research,* 27: 173-178.

Shriberg, E, Bates, R, Stolcke, A, Taylor, P, Jurafsky, D, Ries, K, et al. (1998). Can prosody aid in the automatic classification of dialog acts in conversational speech. *Language and Speech,* 41: 439-487.

Ward, N (2004). Pragmatic functions of prosodic features in non-lexical utterances. In *Proceedings of Speech Prosody 2004*. Nara, Japan, 325-328.

Yngve, V H (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting Chicago Linguistic Society*. Chicago, IL, USA: Chicago Linguistic Society, 567-578.

An acoustic analysis of purring in the cheetah (*Acinonyx jubatus*) and in the domestic cat (*Felis catus*)

Robert Eklund,[1,2,3] Gustav Peters[4] & Elizabeth D. Duthie[5]
[1] *Karolinska Institute / Stockholm Brain Institute, Stockholm, Sweden*
[2] *Voice Provider Sweden, Stockholm, Sweden*
[3] *Linköping University, Linköping, Sweden*
[4] *Forschungsmuseum Alexander Koenig, Bonn, Germany*
[5] *Dell Cheetah Centre, Parys, South Africa*

Abstract

This paper analyses purring in a cheetah and a domestic cat from an acoustic point of view. The results are discussed in the light of previously published studies.

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. Less known is that most other felids (cats) also purr, some relatively loudly. This paper will describe purring in the domestic cat, and compare that with purring in one of the biggest purrers, the cheetah.

Felids: an overview

Felids are among the most successful carnivores ever to develop within the mammal lineage. The number of felid species varies between the sources, and different attempts of classifications have been made based on morphology, behavior and genetics, but the number of species normally lies in the range of 35 to 40 species (Wilson & Reeder, 2005; Sunquist & Sunquist, 2002).

Almost all of the wild cats are considered endangered today (*IUCN Red List*).

The domestic cat

The domestic cat (*Felis catus*, Linneaus 1758) is by far the most well-known cat, and with an estimated number of 600 million individuals worldwide (Driscoll et al., 2009) it far outnumbers all its relatives. It was long believed that the cat was first domesticated in ancient Egypt around 3600 years ago, but recent evidence suggests that domestication took place much earlier, perhaps 10,000 years ago, in the Fertile Crescent. The closest relative of the domestic cat is considered to be the African (Middle Eastern) 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).

The cheetah

The cheetah (*Acinonyx jubatus*, Schreber 1776) is probably best known for being the fastest land animal in the world with an estimated top speed of around 105–110 km/h (Sharp, 1997; Sunquist & Sunquist, 2002, p. 23).

The cheetah is roughly the same size as a leopard (*Panthera pardus*) – with which it is most often confused – but is of a lighter and more slender build, has a smaller head, smaller teeth, and is a poor climber. The cheetah is also distinguished by dark tear-marks running down its eyes and has short fur. Cubs are also characterized by a pronounced mane at the nape and shoulders, which to some extent can persist into adulthood. The cheetah’s claws are only weakly retractile, a rare trait in felids, shared only with only three other species of cat, the fishing cat (*Prionailurus viverrinus*), the flat-headed cat (*P. planiceps*) and the Iriomote cat (*P. iriomotensis*) (Leyhausen & Pfleiderer, 1999).

Sexual dimorphism is not very pronounced in the cheetah. A male cheetah weighs 29–65 kg, a female 21–63 kg. A male is 172–224 cm nose-to-tail, a female 170–236 cm. A male has a shoulder height of 74–92 cm, and a female 67–84 cm (Hunter & Hamman, 2003).

A major characteristic of the cheetah is the lack of genetic variation (O'Brien et al., 1985), most likely due to a near-extinction event during the late Pleistocene (c. 10,000–12,000 years ago), when all but a handful of cheetahs went extinct – along with a large number of large mammals (Menotti-Raymond & O'Brien, 1993). Modern cheetahs show an extremely high frequency of spermatozoal abnormalities, and infant mortality is high (O'Brien et al., 1987).

Although the cheetah is a relatively large carnivore, there are no records of a wild cheetah ever killing a human being (Hunter & Hamman, 2003, p. 17).

Around 1900 the population of wild cheetahs was estimated to be around 100,000 (Marker-Kraus, 1997) while the estimated number of wild cheetahs today is around 4000 (Sunquist & Sunquist, 2002, p. 30).

For further information on the cheetah the reader is referred to Sunquist & Sunquist (2002, pp. 19–36) and Krausman & Morales (2005).

Previous research

This section summarizes previous research on egressive–ingressive phonation and purring.

Egressive–ingressive phonation

Although most vocalization in mammals and humans occurs on a pulmonic egressive airstream, pulmonic *ingressive* phonation is not uncommon, both in human speech and phonation and in animal phonation (Eklund, 2008), an example of which being felid purring.

Purring

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

The subdivision of the Felidae, the cat family, into "purring cats" on the one hand, and "roaring = non-purring cats" on the other, originally goes back to Owen (1834/1835) and was definitely introduced by Pocock (1916), based on a difference in hyoid anatomy. The "roaring cats" (lion, *Panthera leo*; tiger, *P. tigris*; jaguar, *P. onca*; leopard, *P. pardus*) have an incompletely ossified hyoid, which, according to this conception, enables them to roar but not to purr. On the other hand, the snow leopard (*Uncia uncia*, or *P. uncia*), as the fifth felid species with an incompletely ossified hyoid, purrs (Hemmer, 1972). All remaining species of the family Felidae ("purring cats") have a completely ossified hyoid which enables them to purr but not to roar. The two cat species studied here, the domestic cat and the cheetah, belong to the latter group.

However, there is no well-founded and unequivocal basis for a classification of the species in the family Felidae according to the absence/presence of purring and roaring, respectively, and differences in hyoid anatomy. Weissengruber et al. (2002) decidedly argued that the ability of a cat species to purr is not affected by the anatomy of its hyoid, i.e. whether it is fully ossified or has a ligamentous epihyoid, and that, based on a technical acoustic definition of roaring, the presence of this vocalization type depends on specific characteristics of the vocal folds and an elongated vocal tract, the latter rendered possible by an incompletely ossified hyoid.

The current classification of the Felidae is based on molecular characteristics (Johnson et al., 2006; O'Brien & Johnson, 2007) and groups the clouded leopards (*Neofelis nebulosa* and *N. diardi*) – with completely ossified hyoids – together with the five cat species in which it is incompletely ossified.

Data collection

Data were collected from two felids, one domestic cat, and one cheetah. The animals, the equipment and data post-processing are described in the following paragraphs.

Equipment

Both animals were recorded with the same equipment. A Canon HG-10 high-definition video camera was used with an external professional high-fidelity Audiotechnica AT813 cardoid-pattern, condenser mono microphone.

Moreover, a long extension cord was used so as to permit video capture from a distance and avoid the risk of the camera constituting a disturbing factor for the animals.

Plate 1. Cheetah data collection. Third author operated a high-definition camcorder while first author adjusted the microphone to different positions relative to the cheetah's muzzle. Egressive–ingressive phonation was synchronized for sound-only version by keeping the hand on the side of the animal's chest while saying "in" and "out" according to expanding (in-breath) or collapsing (out-breath) rib cage. Film available at http://purring.org

Cheetah data

The cheetah (Caine) was a male, 7 years old at the time of the recording (11 December 2009) and weighing 67.5 kilos (i.e. an exceptionally big cheetah). He was recorded in his enclosure at Dell Cheetah Centre, next to his shelter, in a setting very familiar to the animal. Caine is a constant purrer, and also a very loud purrer, and can easily be heard at a distance of more than 40 meters in an outdoor setting with background noises. Video duration was 1m55s. Film captures and a detailed description of the data collection are given above in *Plate 1*.

Domestic cat data

The domestic cat (Misha) was a female, 14 years old at the time of the recording (31 January 2010) and weighing 3.7 kilos. She was recorded in her home, with her owner holding and caressing her to elicit purring, in a setting Misha was used to. The recording procedures were similar to the ones described above, with the exception that the first author operated the video camera, and that good video footage was not obtained (nor aimed for). Except for a slight background noise from a radiator, the room was silent at the time of recording. Video duration was 1m47s.

Data post-processing

Audio tracks were excerpted from the films with TMPGEnc 4.0 Xpress. Working audio format was 44.1 kHz, 16 bit, mono.

Results

The results are presented in *Table 1*, and methodology, analysis parameters/phenomena and observations are described and discussed separately in the following paragraphs.

Analysis tools

Waveforms were created and analysed with Cool Edit, and both waveform and spectrogram analyses were carried out with WaveSurfer. In order to create discernible waveforms, the sound files were amplified with the *Amplify* function in Cool Edit.

In order to obtain number of respiratory cycles per phase and to calculate frequency, the number of respiratory cycles was counted manually from the waveform.

Statistics were calculated with SPSS 12.0.1.

Egressive–ingressive identification

The first analysis that was carried out was to ascertain that the egressive and ingressive phases were correctly identified in both animals. This was done by locating the parts of the recording sessions – in both the audio and video files – when the first author said the words "in" and "out" while holding his hand on the side of the chest of the animals, in synchrony with the breathing, as described in the *Plate 1* caption above. Identification proved completely unproblematic, and the rest of the files were labeled on the basis of sound and waveform characteristics.

Amplitude

It was reported by Frazer Sissom, Rice & Peters (1991) that purring is strongest right in front of muzzle, showing that the purring sound emanates from the mouth and nose. This was confirmed in both cases in that the strongest signal was obtained by holding the microphone right in front of the animals' muzzles.

Some previous sources report that ingressive phases are louder than egressive phases, e.g. Moelk (1944) and Peters (1981).

Table 1. Summary Table. For both the cheetah and the domestic cat results are given for durations, cycles per phase, and frequency. Results are presented independently for egressive and ingressive phases, and statistical tests are performed on differences between egressive and ingressive phonation.

	Cheetah		Domestic cat	
Phonation type	Egressive	Ingressive	Egressive	Ingressive
No. phases analysed	20	20	16	16
Mean duration (ms)	3049	2491	576	573
Mean duration egressive+ingressive	2770		575	
Standard deviation	270	294	90	72
Maximal duration	3450	2900	700	740
Minimal duration	2400	2000	360	450
Δ *t* test (paired-samples, two-tailed)	$p < 0.000$		$p = 0.926$	
Δ Wilcoxon (two related samples)	$p < 0.000$		$p = 0.615$	
Mean no. cycles/phase	63.45	45.35	12.43	13.19
Mean no. cycles/phase egressive+ingressive	54.4		12.8	
Standard deviation	4.83	3.50	1.21	1.64
Maximal no. phases/cycle	70	53	15	15
Minimal no. cycle/phase	53	40	10	10
Δ *t* test (paired-samples, two-tailed)	$p < 0.000$		$p = 0.118$	
Δ Wilcoxon (two related samples)	$p < 0.000$		$p = 0.071$	
Mean fundamental frequency (Hz)	20.87	18.32	21.98	23.24
Mean frequency egressive+ingressive (Hz)	19.6		22.6	
Standard deviation	1.34	1.37	3.36	3.58
Highest fundamental frequency	23.0	20.0	30.5	28.8
Lowest fundamental frequency	17.1	16.2	17.1	18.2
Δ *t* test (paired-samples, two-tailed)	$p < 0.000$		$p = 0.418$	
Δ Wilcoxon (two related samples)	$p = 0.001$		$p = 0.427$	

This was not confirmed in either of the two species of felid here. The egressive phases were clearly louder in the cheetah, while they were roughly equal in amplitude in the domestic cat, as is shown in *Figure 1a* and *Figure 1b*.

Cycles per phase

The number of cycles per egressive and per ingressive phase was much higher in the cheetah than in the domestic cat.

Phase durations

Phase durations were much longer in the cheetah than in the domestic cat. While the egressive and ingressive phases were roughly the same length in the domestic cat, egressive phases were significantly longer in the cheetah.

Phonation and frequency

Phonation in both animals was relatively regular, and not as noisy as described in some of the previous reports on purring, e.g. Frazer Sissom, Rice & Peters (1991, p. 76).

The transitions between egressive and ingressive phases were short in both animals, with durations in the range of 50–200 ms in the cheetah, and 30–50 ms in the domestic cat. Transitions from egressive to ingressive phonation for both the cheetah and the domestic cat are shown in *Figure 2*.

Turning to fundamental frequency, both animals purr well below the lowest note on the piano. Given a weight difference between the two animals with almost a factor twenty, the observation that the two animals are very close in fundamental frequency might perhaps seem somewhat surprising, but it is well established that fundamental frequency can be an unreliable predictor of body weight/size in mammals (Ey, Pfefferle & Fischer, 2007).

Discussion

Our results show that egressive phases in the cheetah are louder and longer (air is expelled slower than inhaled) than ingressive phases.

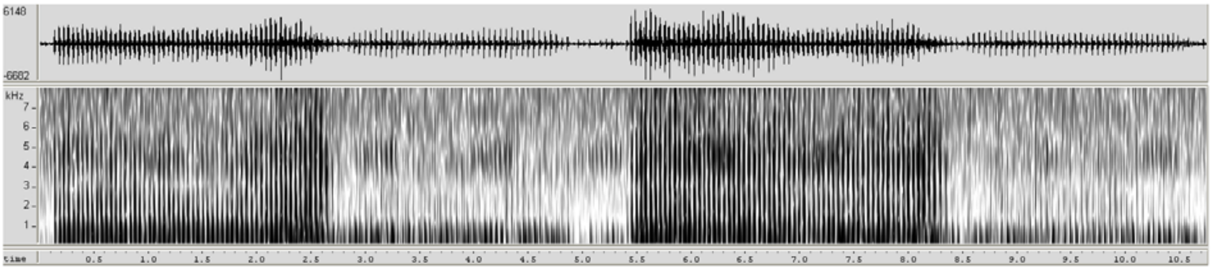


Figure 1a. Waveform and spectrogram of cheetah purring. Egressive–Ingressive–Egressive–Ingressive phases. Hamming window. Window duration 10.5 seconds.

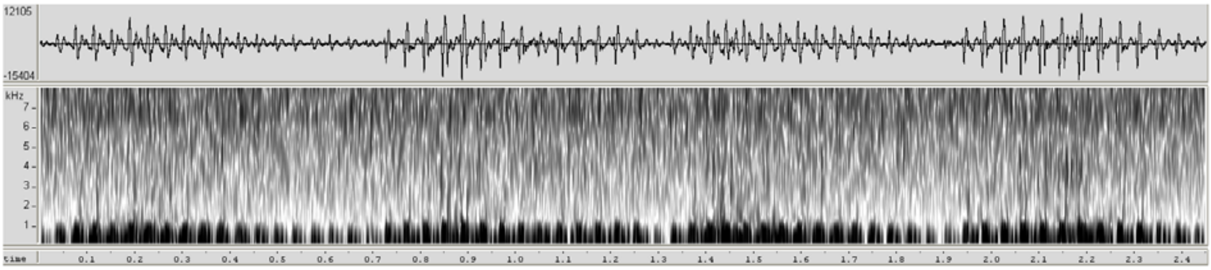


Figure 1b. Waveform and spectrogram of domestic cat purring. Egressive–Ingressive–Egressive–Ingressive phases. Hamming window. Window duration 2.4 seconds.

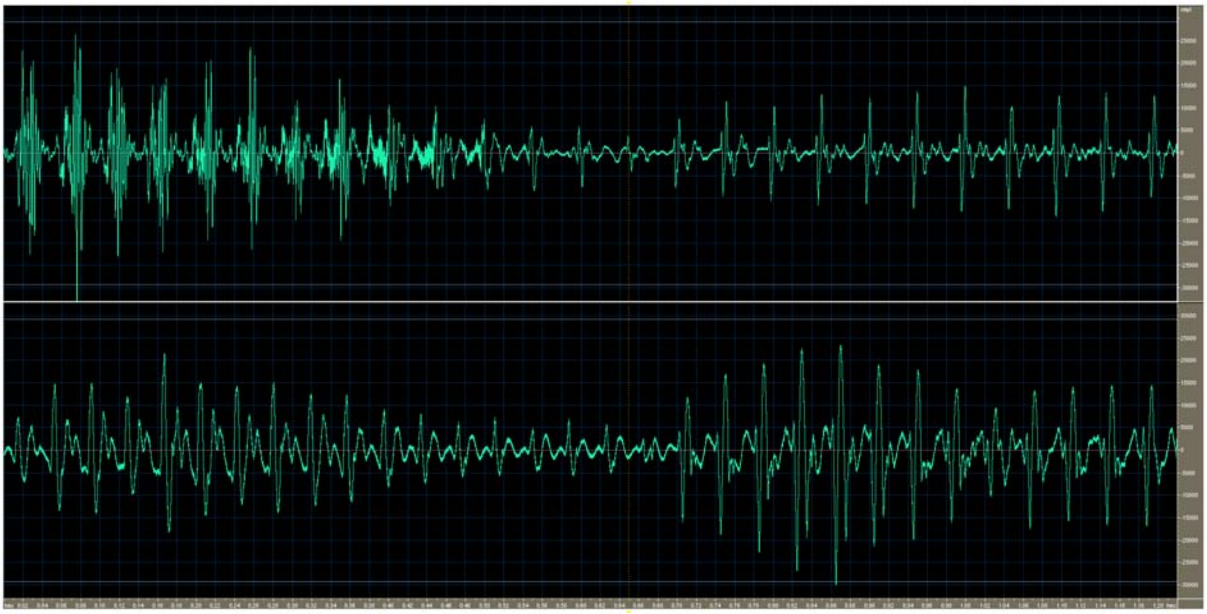


Figure 2. Waveforms showing transitions from egressive (left half) and ingressive (right half) phases of purring in both cheetah (upper waveform) and domestic cat (lower waveform). Window duration 1.2 seconds.

One possible explanation for this is that there might be some physiological asymmetry in the vocal organs involved, creating louder vibration in one direction than the other. This was shown to be case for human beings, by e.g. Catford (1988, p. 57) who pointed out that the asymmetrical shape of the glottis makes egressive phonation very different from ingressive phonation in human speech.

Our observation that egressive phases in the cheetah are louder then ingressive phases is contrary to the results presented in Peters (1981), and that there is no palpable difference in amplitude in the domestic cat at least does not confirm Moelk (1944) who also claimed that ingressive phases were louder. While the latter perhaps can be explained on a purely impressionistic basis in that ingressive phases indeed might sound louder in the domestic cat (due to their “harsher” sound quality), the observed differences between our results for the cheetah and some of the previous reports on felid purring (Frazer Sissom et al., 1991; Peters, 2002) in respect of relative amplitude of egressive and ingressive phases and their respective frequency components warrants further research.

Conclusions

The analyses in this paper both confirmed and differed from some of the previously published reports on purring in felids. Whether this hints at species or individual differences can only be answered by further analyses of felid purring.

Web resources

Data files are available at http://purring.org

Acknowledgements

The authors would like to thank Pieter, Estelle and Ilse Kemp at the Dell Cheetah Centre in Parys, South Africa (http://www.dccafrica.co.za). Also thanks to cat owner Miriam Oldenburg and her very cooperative cat Misha.

References

Catford, J. C. 1988. *A practical introduction to phonetics*. Oxford: Clarendon Press.

Driscoll, Carlos A., Juliet Clutton-Brock, Andrew C. Kitchen & Stephen J. O'Brien. 2009. The taming of the domestic cat. *Scientific American*, June 2009, pp. 68–75.

Driscoll, Carlos A., Marilyn Menotti-Raymond, Alfred L. Roca, Karsten Hupe, Warren E. Johnson, Eli Geffen, Eric H. Harley, Miguel Delibes, Dominique Pontier, Andrew C. Kitchener, Nobuyuki Yamaguchi, Stephen J. O'Brien & David W. Macdonald. 2007. The Near Eastern Origin of Cat Domestication. *Science*, vol. 317, pp. 519–523.

Eklund, Robert. 2008. Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *Journal of the International Phonetic Association*, vol. 38, no. 3, pp. 235–324.

Ey, E., D. Pfefferle & J. Fischer. 2007. Do age- and sex-related variations reliably reflect body size in non-human primate vocalizations? A review. *Primates*, vol. 48, no. 4, pp. 253–267.

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

Hemmer, Helmut. 1972. Uncia uncia. *Mammalian Species*, no. 20, pp. 1–5.

Hunter, Luke & Carl Hamman. 2002. *Cheetah*. Cape Town, South Africa: Struik Publishers.

Krausman, Paul R. & Susana M. Morales. 2005. Acinonyx jubatus. *Mammalian Species*, no. 771, pp. 1–6.

IUCN, International Union for Conservation of Nature and Natural Resources, Red List: http://www.iucnredlist.org/

Johnson, Warren E. Eduardo Eizirik, Jill Pecon-Slattery, William J. Murphy, Agostinho Antunes, Emma Teeling & Stephen J. O'Brien. 2006. The Late Miocene Radiation of Modern Felidae: A Genetic Assessment. *Science*, vol. 311, pp. 73–77.

Leyhausen, P. & M. Pfleiderer. 1999. The systematic status of the Iriomote cat (*Prionailurus iriomotensis* Imaizumi 1967) and the subspecies of the leopard cat (*Prionailurus bengalensis* Kerr 1792). *Journal of Zoological Systematics and Evolutionary Research*, vol. 37, pp. 121–131.

Marker-Kraus, Laurie. 1997. History of the Cheetah. *International Zoo Yearbook*, vol. 35, pp. 27–43.

Menotti-Raymond, Marilyn, Victor A. David, Solveig M. Pflueger, Kerstin Lindblad-Toh, Claire M. Wade, Stephen J. O'Brien & Warren E. Johnson. 2008. Patterns of molecular variation among cat breeds. *Genomics*, vol. 91, pp. 1–11.

Menotti-Raymond, Marilyn & Stephen J. O'Brien. 1993. Dating the genetic bottleneck of the African cheetah. *Proceedings of the National Academy of Sciences*, vol. 90, pp. 3172–3176.

Moelk, Mildred. 1944. Vocalizing In The House-Cat; A Phonetic And Functional Study. *The American Journal of Psychology*, vol. 57, no. 2, pp. 184–205.

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

O'Brien, Stephen J., David E. Wildt, Mitchell Bush, Timothy M. Caro, Clare FitzGibbon, Issa Aggundey & Richard E. Leakey. 1987. East African cheetahs: Evidence for two population bottlenecks? *Proceedings of the National Academy of Sciences*, vol. 84, pp. 508–511.

O'Brien, S. J., M. E. Roelke, L. Marker, A. Newman, C. A. Winkler, D. Meltzer, L. Colly, J. F. Evermann, M. Bush & D. E. Wildt. 1985. Genetic Basis for Species Vulnerability in the Cheetah. *Science*, vol. 227, pp. 1428–1434.

Owen, Richard. 1834/1835. On the Anatomy of the *Cheetah*, *Felis jubata*, Schreb. *Transactions of the Zoological Society of London*, vol. 1, pp. 129–137.

Peters, G. 2002. Purring and similar vocalizations in mammals. *Mammal Review*, vol. 32, no. 4, pp. 245–271.

Peters, Gustav. 1981. Das Schnurren der Katzen (Felidae). *Säugetierkundliche Mitteilungen*, vol. 29, pp. 30–37.

Pocock, R. I. 1916. On the Hyoidean Apparatus of the Lion (*F. leo*) and Related Species of *Felidæ*. *The Annals and Magazine of Natural History, Including Zoology, Botany, and Geology*, vol. 28, series 8, pp. 222–229.

Pontier, Dominique, Nathalie Rioux & Annie Heizmann. 1995. Evidence of selection on the orange allele in the domestic cat Felis catus: the role of social structure. *Oikos*, vol. 73, no. 3, pp. 299–308.

Sharp, N. C. C. 1997. Timed running speed of a cheetah (*Acinonyx jubatus*). *Journal of Zoology*, vol. 241, issue 3, pp. 493–494.

Sunquist, Mel & Fiona Sunquist. 2002. *Wild Cats of the World*. Chicago: University of Chicago Press.

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*, vol. 201, pp. 195–209.

Wilson, Don E. & DeeAnn M. Reeder (eds.). 2005. *Mammal Species of the World. A Taxonomic and Geographic Reference* (3rd ed.). Baltimore, Maryland: Johns Hopkins University Press.

Effects of open and directed prompts on filled pauses and utterance production

Robert Eklund[1,2,3] *& Mats Wirén*[4]
[1] *Karolinska Institute / Stockholm Brain Institute, Stockholm, Sweden*
[2] *Voice Provider Sweden, Stockholm, Sweden*
[3] *Linköping University, Linköping, Sweden*
[4] *Department of Linguistics, Stockholm University, Stockholm, Sweden*

Abstract

This paper describes an experiment where open and directed prompts were alternated when collecting speech data for the deployment of a call-routing application. The experiment tested whether open and directed prompts resulted in any differences with respect to the filled pauses exhibited by the callers, which is interesting in the light of the "many-options" hypothesis of filled pause production. The experiment also investigated the effects of the prompts on utterance form and meaning of the callers.

Introduction

Spontaneous speech differs from (most) printed text in that it includes *disfluency* (to use the most common term), i.e. pauses (unfilled/silent and filled), repetitions, segment prolongations, repetitions, truncated words and so on, with a reported average frequency of around 6% of all "words" uttered (Fox Tree, 1995; Oviatt, 1995; Brennan & Schober, 2001; Bortfeld et al., 2001; Eklund, 2004). From an automatic speech recognition perspective this poses a problem in the design of automated services, since disfluency is not always easy to detect and recognize, and consequently difficult to either "recognize-and-disregard" or to interpret and exploit. In this paper we analyse filled pause incidence in a Wizard-of-Oz (WOZ) data collection, using real customer care agents and real customers with authentic problems, the latter being unaware of their calls being recorded and analysed. More specifically, the phenomenon analysed in this paper is the incidence of filled pauses in customer utterances following either *directed* or *open* system prompts, asking them to describe their reason for calling. Furthermore, we investigated the effects of the prompts on utterance form and meaning of the callers.

Filled Pause hypotheses

Except unfilled pauses (UPs), filled pauses (FPs) constitute the most common form of hesitation in spontaneous speech, and Eklund (2004) reported that approximately 25% of all disfluencies were filled pauses.

Already in the 1950s it was shown that FPs exhibit different distribution and behavior as compared to all other types of disfluency (Mahl, 1958; Christenfeld & Creager, 1996).

Over the years, FPs have been explained according to a number of different hypotheses as to their function(s) in speech, and some (not all) of these will be summarised in the following paragraphs. Note that we will use our own names for the presented hypotheses.

Floor-holding hypothesis. Maclay & Osgood, (1959) were probably the first to suggest that FPs can be used to maintain the floor in conversation, i.e. as a means to prevent interlocutors from breaking in. This view was also forwarded by Livant (1963).

Help-me-out hypothesis. That FPs can be used as a signal asking for interlocutor help was suggested by Clark & Wilkes-Gibbs (1986), or that FPs simply signal to the listener that the speaker is encountering slight timing problems in the production of speech was proposed by Clark (2002). When a speaker is looking for a word or term which is not available to them, uttering "uh" signals to the listener that some help is desired.

Self-monitoring/error detection hypothesis. Levelt (1989) suggested that FPs are a sign of internal error detection, a thread that was extended by Christenfeld & Creager (1996) who were of the opinion that anything that halted speech production could result in emitted FPs, making FPs adhere to Baumeister's (1984) notion of "choking under pressure".

Many-options hypothesis. Lounsbury (1954) proposed that FPs "correspond to the points of highest statistical uncertainty in the sequencing of units in any given order" (ibid., p. 99), i.e. at the beginning of clauses, before the speaker has "committed" to anything, and where speech planning consequently is most difficult. This has been repeatedly confirmed by e.g. Beattie & Barnard (1979) who observed that 55.3% of all FPs produced by customers in telephone conversations (directory enquiries) occurred at the beginning of utterances. Along the same lines, Cook (1971) observed that FPs tended to occur before the first, second or third word of a clause. Shriberg (1994) and Eklund & Shriberg (1998) reported that speakers used FPs at the beginning of utterances more often than in any other position of an utterance. Eklund (2004) reported that 45.3% of FPs were utterance-initial in a large set of corpora, while Boomer (1965) observed that the most frequent position for hesitations were after the first word of phonemic clauses. Perhaps the most striking confirmation of the many-options hypothesis is found in a study by Schachter et al. (1991) who, in order to test the many-options-hypothesis studied hesitations in lectures within three disciplines with varying degrees of inherent optionality: (1) natural science, with very few options (there are very few options to describe the orbit of a planet or the outcome of a chemical reaction); (2) social science (with an intermediate degree of available options); and (3) humanities (with an high number of ways to describe, for example, what Shakespeare really meant with a certain passage). They found that lecturers within the humanities used more FPs than lecturers within social sciences, who, in turn, used more FPs than did lecturers within natural sciences. To rule out individual differences, the same set of lecturers also gave talks on a common subject, in which case they all produced an equal number of FPs.

Attention-getting signal. Lalljee & Cook (1974) reported a number of experiments aimed at testing the floor-holding hypothesis, all of which failed to provide support for the floor-holding function of FPs. Instead, they suggested that FPs might simply fill an attention-getting function, which could also explain the oft-reported high incidence of FPs in utterance-initial positions. However, they also suggested that filled pauses might serve several different functions in conversation, and that any experiment designed to test only one particular hypothesis may not produce significant results because it fails to account for other functions.

Summing up, filled pauses have been assigned several different functions, and several of the hypotheses have been supported by experimental data. One thing to stress, as we have already mentioned was pointed out by Lalljee & Cook (1974) is, of course, that the hypotheses described above are not mutually exclusive, and that FPs might serve several functions, possibly even more than just one function at the same time.

However, there is strong support for the many-options hypothesis, or as Christenfeld (1994) summarizes his study: "more options did produce more filled pauses" (ibid., p. 192).

Semantic categories

Caller utterances were analyzed both with respect to linguistic form and meaning. To represent the meaning of utterances, we used the same tripartite semantic categories (*family*, *intent*, *object*) as in the system that was later deployed (Boye & Wirén, 2007). The first of these elements, *family*, corresponds to the general product family which the call concerns (e.g. fixed telephony, mobile telephony, broadband, etc.), whereas *intent* represents the action associated with the request (e.g. order, want-info, change-info, activate, want-support, report-error, etc.), and *object* represents the specific product or entity (e.g. particular names of products, or concepts like "telephone number", "SIM card", or "password"). For the purposes of our analysis, there were 10 families, around 30 intents, and about 170 objects.

Each slot in a semantic triplet can take the value "unknown", representing the absence of information. For instance, the most accurate semantic category for the common fragmental utterance "broadband" is (*broadband*, *unknown*, *unknown*), since this request conveys nothing about either the *intent* or *object*.

The present study

The aim of this paper is two-fold: First, to study filled pause production in the speech of customers in a customer care entrance, following either a directed system prompt or an open prompt. The hypothesis is that if the many-options hypothesis is true, then FP frequency should be higher in the open-prompt settings. Second, to study whether the prompts have any effect on the semantic triplets.

Data collection

The data analysed this paper were collected during a pilot project carried out at TeliaSonera between December 2004 and February 2005 at the TeliaSonera Customer Service Call Center in Sundbyberg (Sweden).

The aim of the project was to prepare the ground for the launching of speech-based call routing in the Telia residential customer care, a service reached at the number 90200, handling 14 million calls annually.

Call routing is the task of directing callers to a service agent or a self-service that can provide the required assistance.

The data were collected using a novel variant of the Wizard-of-Oz (WOZ) technique (for historical descriptions of WOZ, see Fraser & Gilbert,1991; Dahlbäck, Jönsson & Ahrenberg, 1993; Eklund, 2004), using authentic agents as wizards and authentic customers who were not aware of the fact that the calls were being recorded. Consequently, the quality of the data collected can be assumed to be even better than that of traditional WOZ collections, where neither agents/wizards nor customers are authentic, but are acting out roles given to them by researchers, a critique often raised against WOZ (Allwood & Haglund, 1992; von Hahn, 1986). A detailed description of the present data collection is given in Wirén et al. (2006).

The general structure of the dialogue is as follows: First the (simulated) system plays an initial open prompt, containing a welcome message and an invitation to the caller to describe their reason for call.

If the utterance contains sufficient information to route the call, no more dialogue is needed. If, on the other hand, the utterance contains some but not all information necessary to route the call, the system asks a *disambiguation question* to try to obtain the information required to route the call.

Directed vs open prompts

The experiment described here examined how customers reacted linguistically when asked to express their business, comparing two disambiguation prompts: one **directed prompt**, giving some hints as to possible ways to describe themselves, and one **open prompt**, giving no hints on how to formulate their business.

The two prompts were:

(1) Directed prompt:

Jag behöver veta lite mer om ditt ärende. Gäller det till exempel beställning, prisinformation eller support?

(‘I need some additional information about the reason for your call. Is it for example about an order, price information or support?’)

(2) Open prompt:

Kan du säga lite mer om vad du vill ha hjälp med?

(‘Could you please tell me some more about the reason for your call?’)

The dialogs between the wizards (authentic agents) and the (authentic) customers were transcribed by an independent consulting company, STTS (**www.stts.se**), following the Nuance Guidelines. Although transcription did not focus on disfluency labelling, one type of disfluency was labeled, i.e. the filled pause, which was indicated by the item **@hes@** in the transcriptions. All instances of these hesitation labels were located and listened to (by the first author) to confirm that they were in fact cases of (Swedish) filled pauses, most often transcribed as “eh”.

Comments on data collection

As explained above, the data collection and the experiment described above allow us to take a look at some of the hypotheses proposed concerning the role of the filled pause. It could be argued that the data set is fairly limited, but it has the advantage of being entirely naturalistic (to the point that “experimental” is almost a misnomer) and that it effectively pits directed prompts against open prompts in an otherwise natural setting, with no “roles” assigned, and where all speakers were completely unaware of their speech being recorded for analysis.

Results

The collected data are summarised in *Table 1*.

Utterance form

As can be seen in *Table 1*, the use of an open prompt has dramatic effects on the syntactic-categorical behavior of the customers’ utterances. Following the directed prompt, 72% of the utterances are telegraphic noun-only utterances, and sentences (that contain a finite verb) constitute less than 10% of the utterances. Following the open prompt, more than 40% of the utterances are clauses (including a finite verb) and (one) noun-only utterance are down to less than 20%.

Table 1. Summary Statistics for the directed prompt and the open prompt, and a syntactic-categorical analysis of the customers' utterances and ratios for all categories divided by number of utterances and words. ***Legend: S*** *= clause containing (at least one) finite verb;* ***Noun*** *= single noun;* ***NP*** *= noun phrase;* ***VP*** *= verb phrase;* ***AdvP*** *= adverbial phrase;* ***AP*** *=*

Table 1. Summary Statistics for the directed prompt and the open prompt, and a syntactic-categorical analysis of the customers' utterances and ratios for all categories divided by number of utterances and words. ***Legend: S*** *= clause containing (at least one) finite verb;* ***Noun*** *= single noun;* ***NP*** *= noun phrase;* ***VP*** *= verb phrase;* ***AdvP*** *= adverbial phrase;* ***AP*** *=*

adjective phrase; ***Y/N*** *= "yes" or "no";* ***Interj*** *= interjection;* **–** *= no response given.*

Prompt	Utts	Words	Syntax	% Utts	% Words
Directed	118	216	N = 85	72.0	39.4
			S = 11	9.3	5.1
			Y/N = 8	6.8	3.7
			NP = 6	5.1	2.8
			– = 3	2.5	1.4
			Y/N, Noun = 2	1.7	0.9
			VP = 1	0.8	0.5
			AdvP = 1	0.8	0.5
			AP = 1	0.8	0.5
Open	121	791	S = 49	40.5	6.2
			NP = 26	21.5	3.2
			Noun = 24	19.9	3.0
			VP = 11	9.1	1.4
			AP = 5	4.1	0.6
			AdvP = 2	1.6	0.2
			– = 2	1.6	0.2
			Y/N = 1	0.8	0.1
			Interj = 1	0.8	0.1

Also, utterances following the open prompt are on average three times longer than utterances following the directed prompt. All this clearly shows that the use of an open prompt has clear effects on the linguistic behavior of the callers.

Utterance meaning

Following Boye & Wirén (2007), we can regard every element in the semantic triple as one "concept". We can then obtain a measure of how information increases in the dialogue by computing the difference between the triples representing each user utterance, where "difference" means that the values of two corresponding elements are not equal. The results for semantic concepts are shown in *Table 2*.

As can be seen in *Table 2*, although there is a gain in the number of semantic concepts retrieved from the customers' utterances, the gain is marginal and not statistically significant, either using a *t* test (two-sampled, two-tailed: *p*=0.16 with equal variances assumed; and *p*=0.158 with equal variances not assumed) or Mann-Whitney *U* test (two-tailed, *p*=0.288).

Table 2. Summary Statistics for semantic concept triplets following the directed and the open prompt. Ratios are given for number of concepts compared to number of utterances and words, as well as totals and ratios for the differences (**DIFFs**) *between concepts in and concepts out, i.e., how many concepts you "win" by asking the disambiguation prompt.*

Prompt	Concepts In	Concepts Out	DIFFs Total	DIFFs Change	DIFFs /Utts	DIFFs /Words
Directed	136	244	108	0	0.9	0.5
Open	144	248	122	18	1.01	0.15

As was pointed out in Wirén et al. (2006), however, two other observed differences were that there were no instances following the directed prompt where an already instantiated concept (e.g. `fixedTelephony`) was changed to something else (e.g. `broadband`), while this happens 18 times following the open prompt. Furthermore, following the directed prompt, one never "gains" more than one new concept, while there are 26 instances following the open prompt where the gain is two concepts, and even two cases where the gain is three concepts (which also means that one concept is changed).

Filled pause frequency

FP frequency is shown in *Table 3*.

Table 3. Summary Statistics for utterances, words and filled pauses, and ratios for FPs/Utts and FPs/Words.

Prompt	Utterances & Words			Filled Pauses		
	Utts	Words	Words /Utts	FPs	FPs /Utts	FPs /Words
Directed	118	216	1.8	16	0.14	0.074
Open	121	791	6.5	60	0.50	0.076
Σ	239	1007	4.2	76	0.32	0.075

Needless to say, there is a striking stability across the two settings from a FPs/number-of-words point of view. While number of words per utterance increased by a factor three following the open prompt (with an ensuing difference in number of FPs per utterance), FP occurrence divided by number of words is almost exactly the same following the two prompts. However, the figures, 7.4% and 7.6%, respectively, are considerably higher than the approximately 3.5% reported in Eklund (2004, p. 235) for Swedish in a similar setting, which could possibly indicate that real problems and genuine planning (as was the case here) leads to a higher FP rate than what is observed in a more traditional WOZ data collection with the planning of "pretend" tasks (which was used in Eklund, 2004), which in and by itself is of interest, but needs independent corroboration.

Filled pause distribution

As was previously mentioned it has been shown repeatedly that FPs tend to occupy initial positions in utterances. FP distribution in the present study is shown in *Table 4*.

Table 4. Summary Statistics for FP position, either utterance-initial or in other position.

Prompt	FP position		Σ
	Initial	Other	
Directed	14	2	16
Open	36	24	60
Σ	50	26	76
Proportion	65.8%	34,2%	100%

As is shown in *Table 4*, not only do the majority of FPs occur in utterance-initial position, they do so markedly more so than was reported in Eklund (2004, p. 239), where 1178 out of 2601 (45.3%) of FPs were utterance-initial. The difference is statistically significant given a *Z*-test for two proportions (two-tailed, $p<0.01$).

However, once again there is no statistically significant difference between the two prompt settings. This seemingly repeats the results for FP frequency that subjects with real-world problems in a real-world setting produce more FPs than subjects in a WOZ collection, even when great care is taken to make the WOZ collection as authentic as is possible.

Discussion

Although it could, quite reasonably, be argued that the data set studied in this paper is too small to allow any far-reaching conclusions to be drawn, a counterargument would be that the data are as ecologically valid as is possible, which makes the results interesting, especially when compared to more traditional WOZ data collections, e.g. Eklund (2004).

The first, obvious, result is that it that there is no difference in FP production with respect to the two prompts—the greater number of FPs following the open prompt is most likely a result of the longer utterance following the open prompt. It would seem that to the extent that the many-options hypothesis is valid, the "real-world-authenticity" has already defined and delimited the planning problems the customers might have, and that at least these two prompts in no serious way address that particular problem. From this follows that if FP production is indeed a "planning metric" – and there is much support for that hypothesis, as we have seen – then it would seem that the use of directed prompts in a call center does not help customers in their general speech planning, or at least that this assumption receives no support in the present study.

The different prompts do, however, have other effects on customer utterances, as we have seen above, in that open prompts lead to longer and less telegraphic utterances. It could be argued that the two types of prompts mainly address the *form* of the customers' utterances, rather than the *content*, since there is no significant difference with respect to the "gain" in meaning following open prompts as compared to utterances following directed prompts.

However, different utterances following the open prompt still exhibit a greater *variation* with respect to the increase in meaning: As mentioned in earlier, there were no instances following the directed prompt where an already instantiated concept was changed to something else, while this happens 18 times following the open prompt. Furthermore, following the directed prompt, one never "gains" more than one new concept, while there are several instances following the open prompt where the gain is two or even three concepts.

Finally, the fact that more FPs were produced in this corpus than in the similar WOZ collection reported in Eklund (2004), both as such and in utterance-initial position, could indicate that authentic data possibly can reveal processes that remain hidden in WOZ collections, even if these are well-designed, and also be taken as support for the many-options hypothesis, even if no observable difference was found between the two prompt settings.

Conclusions

In conclusion, we found that FP occurrence is almost exacly the same following the two prompts. However, FP incidence in general was considerably higher than that reported in Eklund (2004), and the majority of FPs occurred in utterance-intial position – markedly more so than was reported in Eklund (2004, p. 239).

Taken together, this might indicate that an authentic setting differs from a WOZ collection, however well-designed.

Concerning utterance form, the two prompt settings gave dramatic differences: Caller utterances following the open prompt were much longer, and also much more conversational in the sense that the utterances more often constituted full clauses, including a finite verb. On the other hand, it was somewhat surprising that there was no significant difference of the gains in meaning with respect to utterances following the two prompts, although the variation was much larger after the open prompt.

Acknowledgements

The data analysed in this paper are covered by a right of use agreement ("nyttjanderättsavtal") between TeliaSonera and Karolinska Institute (637/10), signed on 8 October 2009. Thanks to Jan Lindgren for help with terminology.

References

Allwood, Jens & Björn Haglund. 1992. *Communicative Activity Analysis of a Wizard of Oz Experiment*. Internal Report, PLUS ESPRIT project P5254.

Baumeister, Roy F. 1984. Choking Under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance. *Journal of Personality and Social Psychology*, vol. 46, no. 3, pp. 610–620.

Bortfeld, Heather, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober & Susan E. Brennan. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech,* vol. 44, no. 2, pp. 123–147.

Boomer, Donald S. 1965. Hesitation and grammatical encoding. *Language and Speech*, vol. 8, no. 3, pp. 148–158.

Boye, Johan & Mats Wirén. 2007. Multi-slot semantics for natural-language call routing systems. *Proc. Bridging the Gap: Academic and Industrial Research in Dialog Technologies*. NAACL Workshop, 26 April 2007, Rochester, New York, USA.

Brennan, Susan E. & Michael F. Schober. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, vol. 44, pp. 274–296.

Christenfeld, Nicholas. 1994. Options and Ums. *Journal of Language and Social Psychology*, vol. 13, no. 2, pp. 192–199.

Christenfeld, Nicholas & Beth Creager. 1996. Anxiety, Alcohol, Aphasia, and *Ums*. *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 451–460.

Clark, Herbert H. 2002. Speaking in time. *Speech Communication*, vol. 36, pp. 5–13.

Clark, Herbert H. & Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, vol. 22, pp. 1–39.

Cook, Mark. 1971. The incidence of filled pauses in relation to part of speech. *Language and Speech*, vol. 14, part 2, pp. 135–139.

Dahlbäck, Nils, Arne Jönsson & Lars Ahrenberg, Wizard of Oz Studies — Why and How. 1993. *Knowledge-Based Systems*, vol. 6, no. 4, pp. 258–266. Also in: Mark Maybury & Wolfgang Wahlster (eds.). 1998. *Readings in Intelligent User Interfaces*, San Francisco, CA: Morgan Kaufmann.

Eklund, Robert. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Dept. of Computer and Information Science, Linköping University.

Eklund, Robert & Elizabeth Shriberg. 1998. Cross-linguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. *Proceedings of ICSLP 98*, Sydney, 30 Nov–5 Dec 1998, vol. 6, pp. 2631–2634.

Fox Tree, Jean E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, vol. 34, pp. 709–728.

Fraser, Norman M. & G. Nigel Gilbert. Simulating speech systems. 1991. *Computer Speech and Language*, vol. 5, pp. 81–99.

von Hahn, Walther. 1986. Pragmatic considerations in man–machine discourse. *Proc. COLING*, 25–29 August 1986, Bonn, Germany, pp. 520–526.

Lalljee, Mansur & Mark Cook. 1974. Filled Pauses and Floor-Holding: The Final Test? *Semiotica*, vol. 12, no. 3, pp. 219–225.

Levelt, Willem J. M. 1989. *Speaking. From Intention to Articulation*. Cambridge, MA: MIT Press.

Livant, William Paul. 1963. Antagonistic functions of verbal pauses: filled and unfilled pauses in the solution of additions. *Language and Speech*, vol. 6, part 1, pp. 1–4.

Maclay, Howard & Charles E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word*, vol. 5, pp. 19–44.

Mahl, George F. 1958. On the use of "ah" in spontaneous speech: Quantitative, developmental, characterological, situational, and linguistic aspects. *American Psychologist*, vol. 13, p. 349.

Oviatt, Sharon. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Language*, vol. 9, pp. 19–35.

Schachter, Stanley, Nicholas Christenfeld, Bernard Ravina & Frances Bilous. 1991. Speech Disfluency and the Structure of Knowledge. *Journal of Personality and Social Psychology*, vol. 60, no. 3, pp. 362–367.

Shriberg, Elizabeth Ellen. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California, Berkeley.

Wirén, Mats, Robert Eklund, Fredrik Engberg & Johan Westermark. Experiences of an In-Service Wizard-of-Oz Data Collection for the Deployment of a Call-Routing Application. *Proc. Bridging the Gap: Academic and Industrial Research in Dialog Technologies*. NAACL Workshop, 26 April 2007, Rochester, New York, USA.

Dynamic Vocal Tract Length Normalization in Speech Recognition

Daniel Elenius, Mats Blomberg
Department of Speech Music and Hearing, CSC, KTH, Stockholm

Abstract

A novel method to account for dynamic speaker characteristic properties in a speech recognition system is presented. The estimated trajectory of a property can be constrained to be constant or to have a limited rate-of-change within a phone or a sub-phone state. The constraints are implemented by extending each state in the trained Hidden Markov Model by a number of property-value-specific sub-states transformed from the original model. The connections in the transition matrix of the extended model define possible slopes of the trajectory. Constraints on its dynamic range during an utterance are implemented by decomposing the trajectory into a static and a dynamic component. Results are presented on vocal tract length normalization in connected-digit recognition of children's speech using models trained on male adult speech. The word error rate was reduced compared with the conventional utterance-specific warping factor by 10% relative.

Introduction

Mismatch between training and test conditions is a major cause of performance degradation in automatic speech recognition. Much effort has been invested into reducing this mismatch using adaptation and normalization techniques. A special adaptation category is based on predictive modelling (Gales, 1998). In this approach, explicit knowledge of how specific properties of the speaker or the acoustic environment affect the speech signal is applied to adapt the model to such conditions without the use of separate adaptation data. The technique has been used to compensate for mismatch in background noise (Gales, 1998), vocal tract length (VTL) (Lee and Rose, 1996) and voice source quality (Blomberg and Elenius 2009).

Adaptation based on predictive modelling typically uses the unknown test utterance itself as adaptation data. A transform, which operates on all frames of the utterance or all the trained models, is adjusted to maximize the likelihood of the decoding process. This procedure has been successfully applied in Vocal Tract Length Normalization (VTLN) on both adult and children's speech (Potamianos and Narayanan, 2003; Giuliani et al., 2006; Blomberg and Elenius, 2008). The transform in VTLN is a warping function which expands or compresses the frequency axis of the input signal or the trained model before matching.

However, there are arguments against the use of a time- and, accordingly, phoneme-invariant VTL transform. The effective vocal tract length of a speaker is dynamically increased/decreased by protrusion/spreading of the lips and by lowering/raising of the larynx (Fant, 1960; Dusan, 2007). Especially the larynx height can, to a large degree, be changed without shifting the perceived phonetic identity. This may give rise to intra- and inter-speaker variability for repeated pronunciations of the same word sequence. Another argument is that the length difference between two vocal tracts is in general not evenly distributed. For this reason, the frequency mapping function between their transfer functions is phoneme-specific. For example, phonemes with their main resonance frequencies belonging to the mouth cavity are expected to be quite insensitive to difference in pharynx length. The potential of phoneme-dependent warping has been demonstrated both in terms of formant frequencies (Fant, 1975) and cepstral deviation (Potamianos and Narayanan, 2003).

In recent years, increased interest has been directed towards a time-varying warp factor in VTLN to account for the above effects. Miguel et al. (2005) estimated a frame-specific warp factor by a three-dimensional Viterbi decoding process. Blomberg and Elenius (2007) searched for the best combination of warp factor specific phone models for the test utterance. Maragakis and Potamianos (2008) used a two-pass method, where spectrally similar regions of the test utterance were transformed by the same warp factor. Elenius and Blomberg (2009) computed phoneme-specific warp factors for a group of children using adult models. Although they achieved systematic differences between the factor values, the use of these on another child group with the same age distribution improved the recognition accuracy only marginally.

The studies report little or moderate improvement from using phoneme-dependent and time-varying warp factors. However, the results should not be seen as a final assessment of the idea, since there are still approaches to the problem which have not yet been explored. In this paper, we propose and evaluate new methods for the implementation of dynamic warp factors. In short, a standard HMM is modified by extending each state by a number of warp factor specific sub-states. Constraints on the warp factor trajectory are implemented in the transition matrix. The trajectory can be specified to be constant or to have a constrained rate-of-change within realizations of phones or phone states. We have studied frame-wise adjustment in either of two ways: by unconstrained rate-of-change of the warp factor and by limiting the change to ±1 quantization step. Comparison is made with a phone-model-dependent warp factor, which gives an identical value for all instantiations in the utterance of a phone model, and with the conventional case, a time-invariant warp factor value for the whole utterance.

Although the speaker property used in this report is vocal tract length, the approach should be applicable also to other speaker and environment characteristics with time-varying behaviour, such as speech loudness, voice source quality, speech rate and fluctuating background noise.

Method

Dynamic modelling is accomplished by extending the HMMs of the acoustic model with a speaker characteristic dimension. New states are added in order to model property values deviating from those in the training data. The probability density functions of the new states are derived from the original state distribution by a speaker characteristic property transform. Thus, a probability density function for a new speaker characteristic property value is predicted based on the original pdf and a parametric transform. The rate-of-change constraints are implemented by means of changing the state transition probabilities in the extended model, which we will refer to as a Speaker Characteristic Augmented HMM (SCA-HMM). Further details of the method are given in (Elenius, 2010).

Phone model specific warping

Phone model specific warping does not require the SCA-HMM representation, and can be implemented as a standard HMM. The warp factors are estimated in a combined recognition-estimation step, in which a full recognition procedure is performed for each combination of factors considered. The search objective is to find the set of individually warped phone models, which maximizes the likelihood of the utterance. An exhaustive search is computationally very heavy and a reduction of the search space is required to make the search feasible. In this report, the warp factor is determined separately for each phone model, while the other models have an initial default warp factor value, as in Blomberg and Elenius (2007). Although the separate search has been shown to be sub-optimal, it is included here for comparison with other estimation methods.

Phone and sub-phone instance specific warping

The spoken realization of a phoneme is influenced by several variability sources, such as phonetic context, its position in the utterance, within a stressed or unstressed word, etc. It is thus unlikely that two realizations in the same utterance would have identical warp factors. This reduces the efficiency of phone model specific values. To account for this effect, phone-instance-specific warp factors can be used.

There is also motivation for changing the warp factor within a phone, i.e., between sub-phone segments. One example is the occlusion and the release phase of unvoiced plosives, which have different characteristics and should be modelled differently. The occlusion phase is likely to consist mostly of background sounds which are speaker-independent, while the release has speaker characteristic features.

These segments need to be compensated differently.

Variability in the movements of the articulators may cause a need to change the warp factor on a frame-by-frame basis. In general, the rate-of-change is expected to be slow due to the limited speed of the articulators, but more rapid changes might be required at phoneme boundaries and in transitional regions. To model this effect, different constraints should be used within and between phone-instantiations. We have chosen unconstrained rate-of-change at transitions between models. This is also motivated from a computational point of view, since constraining the trajectory across model boundaries would require an extensive increase in complexity using the current approach.

As mentioned above, the expansion of a standard HMM into an SCA-HMM adds sub-states to each original HMM state, where each of the new states represents a transformed version of its source state. Transitions between states are then added to model the different types of factor dynamics. The structure of the new SCA-HMM is indicated in Figure 1. We will refer to the states of the original HMM as main-states in the SCA-HMM. The new warp-specific states will be described as sub-states of the original main-states.

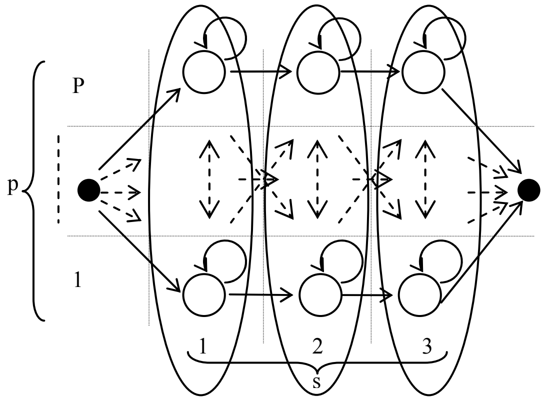


Figure 1. An SCA-HMM with a speaker-property value index, p, mixture, m, and main-state index, s. The original model is a three-state left-to-right HMM.

Constraints on the rate-of-change are implemented by selecting a subset of the possible transitions. In cases of more than one connection from a sub-state to other sub-states, uniform transition probability distribution is used.

Phone-instance-specific warping is realized by only allowing transitions between sub-states with the same warp factor. This will force the warp factor to remain fixed throughout the SCA-HMM. In Figure 1, this corresponds to having only horizontal connections between the main-states.

Sub-phone-specific warping is realized by adding transitions between sub-states of different warp factors between the main states of the phone-instance-specific model. In Figure 1, these are represented by non-horizontal connections between the main-states.

Frame-specific warp factors are implemented by adding intra-main-state transitions to the sub-phone-specific case. These correspond to vertical transitions within a main-state in Figure 1. Two types of possible transitions are used; one that enables a change to any other value and one that limits the factor change rate to one quantisation step. In the latter case, the same constraint is also applied to the inter-main-state transitions.

Static-dynamic factor decomposition

A straightforward candidate rule for constraining the allowed warp factor range during an utterance is to make certain that the range is sufficient for most speakers. However, this will permit the warp factor trajectory to vary between values corresponding to the shortest and the longest allowed vocal tracts in the same utterance. This is likely to limit recognition accuracy. We have approached this problem by decomposing the trajectory into a static, utterance specific, and a dynamic component, which are jointly optimized. The static component is estimated in a grid search while the dynamics, being modelled in the transition matrix, are determined in the Viterbi process. In this way, the dynamic range can be reduced by excluding inter-speaker variability of the static component. To reduce the increased computational demand of the joint two-fold search, existing speed-up techniques, such as a two-pass method (Lee and Rose, 1996) or a tree-based procedure (Blomberg and Elenius, 2009), can be applied.

Experiments

An experiment has been performed on the connected-digit corpus TIDIGITS. A speech recognition system was trained on the adult-male part of the full training set. Evaluation was performed on the child part of the full TIDIGITS test set. The child subset consists of more than 3800 strings with more than 12600 digits.

In an experiment using a single component representation of the warp factor, the trained phone models were extended to SCA-HMMs by frequency warping with a warp factor in the range 1.0 to 1.7 with a step of 0.02. During recognition, warp factors were estimated specific to the utterance, a phone model, a phone instance, a phone state instance, or a speech frame. For the frame-wise estimation, an unconstrained rate-of-change of the warp factor was compared to one limited to ±0.02 between frames within a phone instance. In both cases, the inter-phone-instance change of warp factor was unconstrained.

Experiments with a static and dynamic component were performed using an SCA-HMM of 9 sub-states per main-state, spanning a dynamic interval of 0.16. Its centre represented the static component, which was determined by a grid search in steps of 0.02. The range of the combined warp method was limited to the same interval as above.

The experiments were performed using a connected-digit recognition system with triphone HMMs implemented in HTK. Each acoustic model consisted of 3 states with a GMM consisting of 16 mixtures and diagonal covariance matrices. A 39-dimensional acoustic feature vector was composed by 12 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 (Pitz and Ney, 2005). To avoid erroneous warping due to cepstral smoothing effects during transformation, training was performed using 18 cepstral coefficients. After warping the trained models, their mean and variance vectors were reduced to contain the 12 lower static, delta and acceleration coefficients, like in (Blomberg and Elenius, 2008).

Results and Discussion

The word error rates of the investigated methods are shown in Table 1. The original error rate using adult models was considerably decreased by all VTLN methods. Using single-component representation, none of the estimation units was significantly better than a standard utterance-specific warp factor. In contrast, the decomposed static and dynamic representation demonstrated clear superiority over the utterance-specific factor and all single-valued estimation units, with a minimum WER of 3.47% for state-instance-specific factor. Similar relations between the different estimation units were observed as for the non-decomposed case.

The superior result of the state-instance specific factors compared with that of the phone-instance condition supports the initial argument that the warp factor should be allowed to differ between sub-phonemic sound-events.

The error rate of the phone-model-specific factors is substantially higher than the other techniques. One reason for this might be found in the suboptimal search algorithm.

Table 1. WER of VTLN methods. The baseline result for original male models is 47.55%.

	Estimation technique	
VTLN estimation unit	Single component	Static & dynamic
Utterance	3.85	-
Phone model	6.47	-
Phone instance	4.19	3.69
State instance	3.84	3.47
Frame, Δ unconstrained	3.87	3.67
Frame, |Δ| <= 0.02	4.13	3.71

An example of chosen warp factor trajectories for a child’s utterance (from the TIDIGIT training set) is shown in Figure 2. The unconstrained frame-specific warp factor exhibits a substantial variation from frame to frame. A high rate-of-change is used for acquiring the best match. Still, the repetitions of identical digits exhibit similar, regular, patterns.

The heavily rate-constrained frame-based trajectory is evidently too smooth to adjust to rapid intra-phone effects. It is probable that an intermediate rate limit would give a better result. A few instances of abrupt change in the trajectory are explained by the fact that the rate was unconstrained at phone boundaries.

The sub-phone instance specific method results in a warp factor, which is constant within main-states and changes instantly at state transitions. It partly includes the abrupt changes observed in the frame-based method but removes within-state variation.

Static-dynamic decomposition can be observed to correct excessive warping of single component warping at several positions.

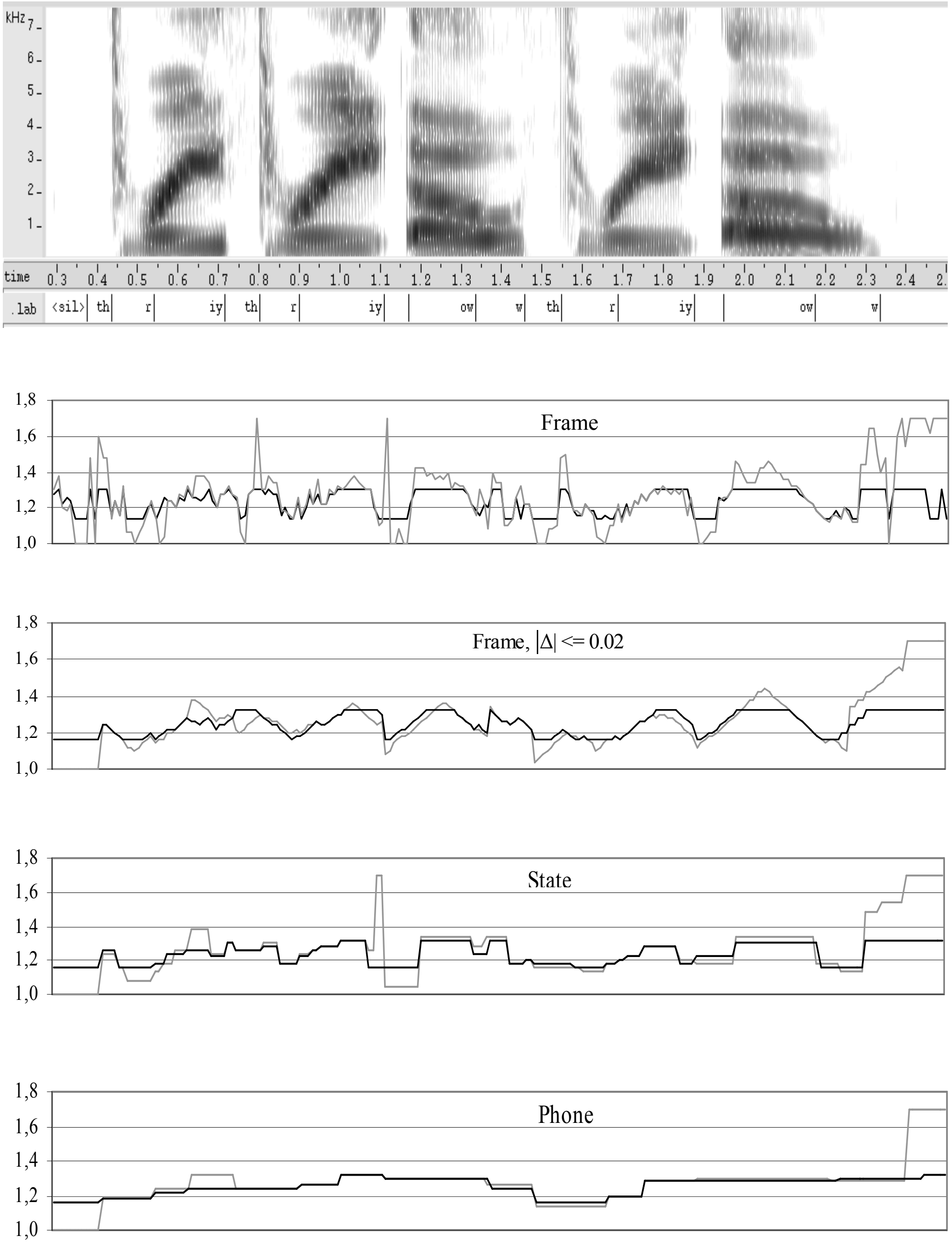


Figure 2. Warp-factor as a function of time for frame, state, and phone instance-specific warping for a boy's utterance "3 3 oh 3 oh". Gray and black curves represent the use of a single-component and a static-dynamic decomposed warp factor, respectively, during decoding. An utterance-specific factor was estimated to 1.24.

Conclusion

A novel method to incorporate time-varying speaker characteristic properties into the acoustic model was presented and applied to vocal tract length normalization. The main performance improvement compared with the conventional utterance-specific warping is due to the factor decomposition into a static and a dynamic component. In this way, the dynamic range during an utterance is limited to what can be expected for an individual speaker. Of the tested conditions, the best position for implementing the dynamic component was between phone sub-states, keeping the warp factor fixed within each state. The method lowered the standard utterance-specific warping error rate by 10% relative.

We are optimistic regarding its potential for further improvement. The proposed technique is flexible and was straightforwardly implemented in a conventional phone-based HMM system. The rate-of-change constraints of the warp factor are efficiently implemented in the transition matrix. This approach makes it possible to train model-specific dynamic properties of the warp factor using conventional training procedures. Further work will be directed to such training.

The proposed approach is not limited to frequency warping, but can be applied to other time-varying speaker characteristic and environmental properties, such as voice source quality, speech rate, non-stationary background noise, microphone distance, etc.

Acknowledgement

The work was funded by the Swedish Research Council.

References

Akhil T, Rath P, Umesh S and Sanand D R (2008). A Computationally Efficient Approach to Warp Factor Estimation in VTLN Using EM Algorithm and Sufficient Statistics. *Proceedings of Interspeech*, 1713-1716.

Blomberg M and Elenius D (2007). Vocal tract length compensation in the signal and model domains in child speech recognition. *Proceedings of Fonetik-2007*. TMH-QPSR, 50(1), KTH, Stockholm.

Blomberg M and Elenius, D (2008). Investigating Explicit Model Transformations for Speaker Normalization. *Proceedings of ISCA-ITRW Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark.

Blomberg M and Elenius, D (2009). Tree-based estimation of speaker characteristics for speech recognition. *Proceedings of Interspeech*, 580-583.

Dusan S (2007). Vocal Tract Length during Speech Production. *Proceedings of Interspeech*, 1366-1369.

Elenius D (2010). *Accounting for Individual Speaker Properties in Automatic Speech Recognition*. Lic. Thesis, CSC/TMH, KTH, Sweden.

Elenius D and Blomberg M (2009). On Extending VTLN to Phoneme-specific Warping in Automatic Speech Recognition. *Proceedings of Fonetik 2009*. Dept. of Linguistics, Stockholm University.

Fant G (1960). *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton.

Fant G (1975). Non-uniform vowel normalization. *Quarterly Progress and Status Report*, Department for Speech Music and Hearing, Stockholm, Sweden.

Gales M J F (1998). Predictive model-based compensation schemes for rubust speech recognition. *Speech Communication*, 25: 49-74.

Giuliani D, Gerosa M and Brugnara F (2006). Improved Automatic Speech Recognition Through Speaker Normalization. *Computer Speech & Language*, 20(1): 107-123.

Lee L and Rose R C (1996). Speaker Normalization Using Efficient Frequency Warping Procedures. *Proceedings of ICASSP*, 353-356.

Maragakis G and Potamianos A (2008). Region-Based Vocal Tract Length Normalization for ASR. *Proceedings of Interspeech*, 1365-1368.

Miguel A, Lleida E, Rose R C, Buera L and Ortega A (2005). Augmented state space acoustic decoding for modeling local variability in speech. Proceedings of ICSLP, 3009-3012.

Pitz M and Ney H (2005). Vocal Tract Normalization Equals Linear Transform in Cepstral Space. *IEEE Transactions on Speech and Audio Processing*, 13(5): 930-944.

Potamianos A and Narayanan S (2003). Robust Recognition of Children's Speech. *IEEE Transactions on Speech and Audio Processing*, 11(6): 603-616.

Vowel Dependence for Electroglottography and Audio Spectral Tilt

Laura Enflo
Dept. of Speech, Music and Hearing, Royal Institute of Technology (KTH)

Abstract

The spectral tilt has been calculated for the vowels from audio signals and from the derivative of electroglottography (EGG) signals in a Swedish one-speaker corpus of 5277 sentences. Vowel dependence has been found for the audio spectral tilt but not for the EGG spectral tilt, which also gets steeper as the sound pressure level increases. The EGG spectral tilt values had an average standard deviation of 1.3 dB/octave and the corresponding standard deviation value for the audio spectral tilt was 2 dB/octave.

Introduction

Spectral tilt, or spectral slope, is an important parameter in voice synthesis and voice perception. One way of defining it is 'the least squares linear fit to the harmonic peaks in the spectrum of the glottal waveform calculated from an audio speech signal'. A less steep spectral tilt corresponds to a louder voice and when loudness decreases, the spectral slope has been confirmed to increase (Fant and Lin, 1988 & Hanson, 1997). Spectral tilt has also been linked to different voice qualities. For example, steeper spectral tilt values have been found for female voices perceived as breathy by speech therapists (Karlsson, 1988 & 1992). In automatic detection of prominence (emphasis) in speech, on the other hand, the spectral tilt parameter has been found to be a weak indicator (Al Moubayed et. al., 2010).

The spectral tilt is, in consequence, dependent on the placement of formant peaks. Formants are known to provide the necessary acoustic information for vowel identification. For each vowel type, however, the formant frequencies have a high variability across speakers and phonetic contexts (e.g. Hillenbrand et. al., 1995). Independent manipulations of the formant frequency and spectral tilt parameters have been carried out in a more recent experiment. Each of the parameters had on their own a strong impact on the perception of vowel quality. Nevertheless, the perceptual effect of changes in spectral tilt was lessened in stimuli in which formant frequencies, as in natural-sounding speech, changed over time. Therefore, spectral tilt might be of little relevance to the process of identifying vowels correctly (Kiefte and Kluender, 2005).

Electroglottography, henceforth EGG, is a widely used technique for the assessment of vocal-fold contact during phonation (e.g. Orlikoff, 1998). An electroglottograph is provided with two electrode plates, which are placed on each side of the larynx. The idea with EGG is to send an electrical current from one electrode to the other and record the amplitude of this signal. When the vocal folds are closed, the signal can pass, such that the current is higher, and when the glottis is open, the current is lower. Several studies have shown that EGG is a more robust technique than audio for fundamental frequency (f0) determination thanks to its simpler waveform (e.g. Vieira et. al., 1996).

Many research projects have aimed to acquire information about the glottal voice source. A widely popular model is the Fant theory of a separated voice source and filter (Fant, 1960). In Fant's model, the source can be modeled with the derivative of the glottal flow volume velocity. Before glottal source parameters can be obtained from the audio speech signal, however, vocal tract resonances need to be removed with the help of inverse filtering (Miller, 1959 & Rothenberg, 1973). Although the inverse filtering procedure can be done digitally, it often needs time-consuming modifications afterwards which make it harder to obtain the amount of data needed for a satisfactory statistical evaluation. Therefore, a considerable amount of research has been carried out on how to use the speech spectrum instead of time-domain estimations. One discovery has been that H1*-A3* (the amplitude of the third formant relative to that of the first harmonic) is correlated with the source spectral tilt, except when the first harmonic is weak (Hanson, 1997). A recent study has shown that H1*-A3* is vowel dependent (Iseli et. al., 2007).

Although the EGG signal is not a measure of glottal flow, it can give information about the glottal voice source without the use of the audio signal, which according to Fant's theory is strongly influenced by the linear filter that can be modeled from the vocal tract. In brief, audio spectral tilt is determined by both the source and the filter, whereas the EGG spectral tilt is affected only by the source. If EGG spectral tilt shall be used in speech synthesis and analysis in the future, it would be valuable to know whether it is vowel dependent or not. Since vowel identities are mainly determined by the shape of the vocal tract, it is likely that the EGG has no impact on this matter. However, since the result in the study by Iseli et. al. (2007) is contradictory to this idea, it is worth while to make an investigation.

MF Corpus

The MF corpus consists of 5277 sentences from newspaper texts and literature read by a professional male Swedish actor and was recorded in 2002 in a studio. The sampling frequency of the data is 16 kHz. Each audio signal has a corresponding signal with the EGG derivative.

Annotation was made with Sjölander's speech aligner (Sjölander, 2003). Occasional problems in the annotation of unusual words or names resulted in exclusion of some sentences, so that the speech material in this study added up to 5114 sentences in total. From this set of data all vowels with duration of at least 31 ms were picked (i.e. 500 data points), with the consequence that less than 7 % of all the vowels were discarded for the purpose of obtaining more reliable data. The average duration of a vowel was 117 ms.

The following vowels are represented in the corpus: [a, ɑː, e, eː, ɪ, iː, u, uː, ɵ, ʉː, ʏ, yː, ɔ, oː, ɛ, ɛː, æ, æː, ø, øː, œ, œː]. The total number of picked vowels was 35990 from the EGG files and 36585 from the audio files, with the distribution between vowel types as seen in Table 1.

Table 1. Number of picked vowels from the MF corpus, sorted in order of frequency when calculating the EGG spectral tilt (EGG ST). Number of vowels picked for calculating the audio spectral tilt (Audio ST) are also included.

Vowel	**EGG ST**	Audio ST
a	**3866**	3945
ɪ	**3269**	3346
ɔ	**2634**	2665
ɛ	**2633**	2749
e	**2616**	2693
ɑ :	**2594**	2649
e:	**2490**	2514
o:	**2148**	2066
i:	**2003**	2059
ʉ :	**1628**	1631
u:	**1593**	1554
u	**1436**	1448
ɵ	**1374**	1401
æ	**1038**	1089
ø:	**886**	911
œ	**836**	827
ʏ	**678**	698
y:	**597**	594
ɛ :	**541**	555
œ:	**443**	473
ø	**376**	394
æ:	**311**	324
Total	**35990**	36585

Parameters

For each of the vowels, the respective sound pressure level, loudness, duration, audio spectral tilt and EGG spectral tilt were calculated. A more thorough description of these features follows below:

Sound pressure level: the logarithmic value of the sound pressure of the speech signal relative to the human hearing threshold p_{ref}=20e-6 Pa (Timoney et. al., 2004).

Loudness was calculated according to the standard ITU-R BS 1770-1 (ITU Original) (ITU-R, 2006 & Nygren, 2009).

Duration: The start time was subtracted from the end time for each respective vowel and the results were displayed in seconds.

Audio spectral tilt and *EGG spectral tilt*: The spectral tilt values were obtained in a similar way for audio and for EGG. Three methods were tried out and evaluated on the EGG files. The best method was then chosen for both audio and EGG spectral tilt calculations. In the first method investigated, all separate spectrum peaks were found (using the findpeaks function by O'Haver, 1995) and a linear line which best fitted the data was calculated in a least-squares sense. The slope of the line provided the spectral tilt value. For the second method explored, the calculations were made in the same way, but with a cut-off frequency threshold corresponding to 6000 Hz on a logarithmic scale, in order to erase the impact of the peaks in the frequency range above the threshold, which was thought to be a source of error. Although the median values from both of these two methods showed stability (with standard deviations of 0.13 versus 0.15, respectively), the standard deviation of the average values was 40% higher for the first method in comparison to that of the second method. In addition, the average values differed greatly from the median values (on average 3.8 dB/octave) in the first method. In the second method, this difference was only 0.4 dB/octave on average. In the third method the calculations were made in the same way as for the first two methods, but with a cut-off frequency threshold corresponding to 3000 Hz on a log scale. In addition, spectral tilt data suffering from one of the three following problems were elimited from the analysis: 1) measured point was singular, 2) value was measured on too high frequencies, at least corresponding to 200 Hz, which is about twice as high as the average frequency for the speaker's voice, 3) signal was noisy in the middle frequency range, resulting in positive values. 10 % of the data had to be discarded due to these three problems.

The average value per vowel had a standard deviation of 10.5 dB/octave in the second method. For the third method, the corresponding standard deviation was 1.3 dB/octave. Consequently, the third method was used in this experiment.

Analysis

All of the parameters mentioned above were implemented in and the analysis was performed with Matlab 7.3.0 (release 2006b). The calculations were made on the 500 middle data points of each vowel, except for the vowel duration parameter. Statistical analysis was carried out in SPSS.

Results

The sound pressure level is 77 dB ± 3.4 dB for the audio spectral tilt values (Figure 1).

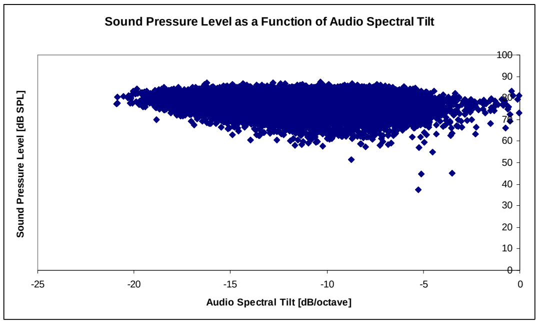


Figure 1. Sound pressure level as a function of audio spectral tilt.

For the EGG spectral tilt, there is a noticeable decrease of the sound pressure level as the spectral slope is decreased (Figure 2).

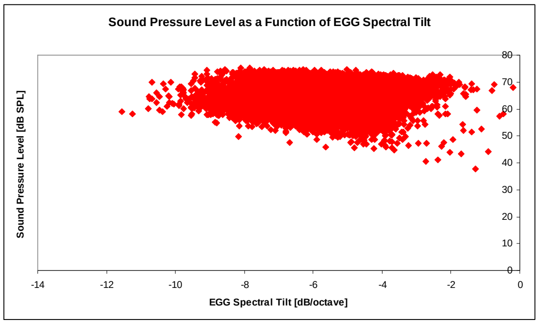


Figure 2. Sound pressure level as a function of EGG spectral tilt.

Statistically significant differences for the audio spectral tilt values were found between most vowels. The median values (Figure 3) have the same values on standard deviations as the average audio spectral tilt (Figure 4); 2 dB/octave.

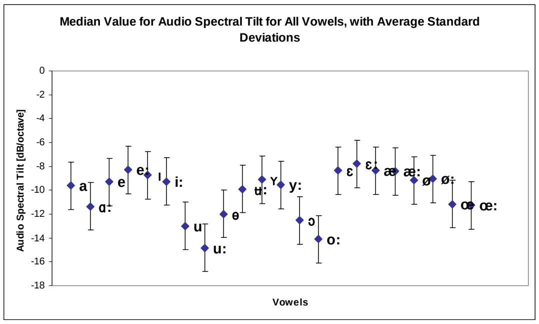


Figure 3. Median values for audio spectral tilt for all vowels, with the average standard deviations.

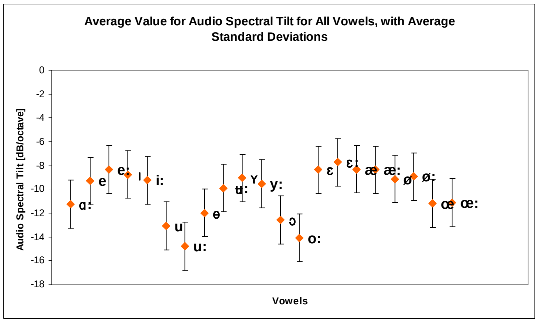


Figure 4. Average values for audio spectral tilt for all vowels, with the average standard deviations.

For EGG spectral tilt, the differences between the vowels are not significant and the small existing variations are shown in the graph with the median values (Figure 5). The average EGG spectral tilt values show a similar pattern (Figure 6).

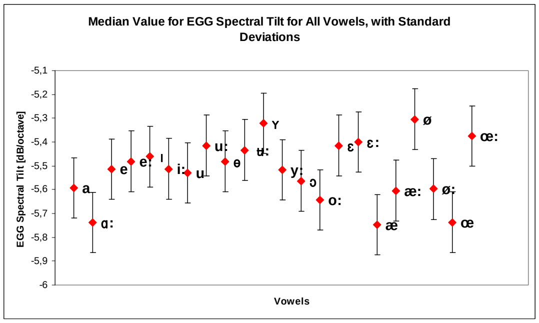


Figure 5. Median values for EGG spectral tilt for all vowels, with the average standard deviations.

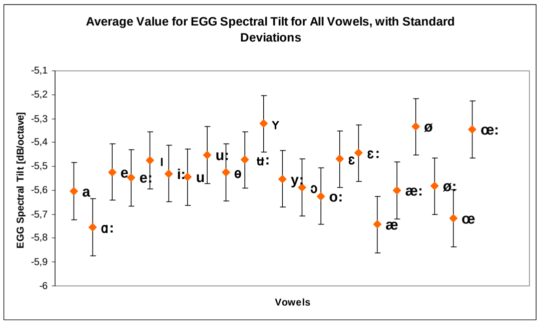


Figure 6. Average values for EGG spectral tilt for all vowels, with the average standard deviations.

The average loudness (Figure 7) and sound pressure level (Figure 8) for each vowel corresponded to each other, with / ʏ / and / œ:/ reaching the highest values.

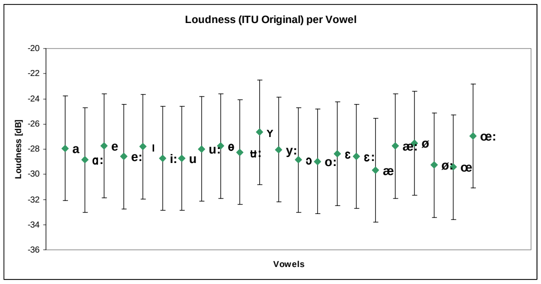


Figure 7. Average values for the ITU Original loudness for all vowels with the average standard deviations.

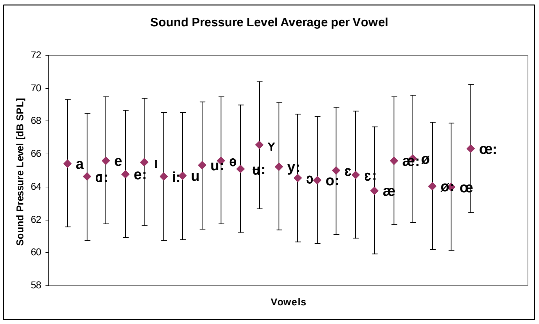


Figure 8. Average values for the sound pressure levels for all vowels with the average standard deviations.

Discussion

This study confirms previous findings of vowel dependence in parameters related to audio spectral tilt and also supports the conclusion that this feature, due to its variability, is insufficient for correct vowel identification. In the light of the source-filter-theory introduced by Fant (1960), it is not surprising that the EGG spectral tilt has no vowel dependence; a vowel is determined by formants which are formed in the vocal tract (the filter) whereas the EGG is related to the source only. This conclusion can, however, be blinded by the fact that we only used one speaker and to problems in EGG measurements, which, although they can be significant, we know nothing about in this case. When using EGG spectral tilt in speech synthesis applications, it is useful to know that vowel dependence is likely to be of little importance. Hence, all sorts of recorded speech, both spontaneous and read, where the vowel distribution is likely to be skewed, can be used without further sorting for EGG spectral tilt.

In this investigation, the sound pressure levels cannot be reconfirmed as being decreased with a steeper audio spectral tilt, also when looking at each vowel separately. This could possibly be explained by a low variability in sound pressure level for the speaker. The EGG spectral tilt, on the other hand, shows a decrease with a lowered sound pressure level.

Further research on the use and nature of the EGG spectral tilt would be valuable, especially since it seems to be a comparatively stable parameter.

Conclusions

The EGG spectral tilt parameter does not vary significantly with vowel type and the values typically get steeper as the sound pressure level increases. The audio spectral tilt values are vowel dependent.

References

Al Moubayed, S, Ananthakrishnan, G & Enflo, L (2010). Automatic Prominence Classification in Swedish. To be published in *Proceedings of Speech Prosody 2010, Workshop on Prosodic Prominence*. Chicago, USA.

Fant, G (1960). Acoustic Theory of Speech Production. Mouton, The Hague, Paris.

Fant, G & Lin, Q (1988). Frequency domain interpretation and derivation of glottal flow parameters. *Speech Transmission Laboratory Quarterly Progress Scientific Report* 1988(2-3): 1-21.

Hanson, H M (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America* 101(1): 466-481.

Hillenbrand, J, Getty, L J, Clark, M J, and Weeler, K (1995). Acoustic characteristics of American English vowels, *J. Acoust. Soc. Am.* 97: 3099-3111.

Iseli, M, Shue, Y-L & Alwan, A (2007). Age, sex and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am.* 121 (4): 2283-2295.

ITU-R (2006). Rec, ITU-R BS. 1770-1, Algorithms to measure audio programme loudness and true-peak audio level. International Telecommunication Union.

Karlsson, I (1988). Glottal waveform parameters for different speaker types. *Speech '88: Proceedings 7th FASE Symposium*, edited by W.A. Ainsworth and J.N. Holmes (Institute of Acoustics, Edinburgh), 225-231.

Karlsson, I (1992). Modelling voice variations in female speech synthesis. *Speech Communication* 11, 1-5.

Kiefte, M & Kluender, K (2005). The relative importance of spectral tilt in monophthongs and diphtongs. *J. Acoust. Soc. Am.* 117: 1395-1404.

Miller, R L (1959). Nature of the vocal cord wave. *J. Acoust. Soc. Am.* 31: 667-677.

Nygren P (2009). Achieving equal loudness between audio files - Evaluation and improvements of loudness algorithms. Master's thesis, Dept. of Speech, Music and Hearing, KTH, Stockholm, Sweden.

O'Haver, T C (1995). Version 2 Last revised Oct 27, 2006: http://terpconnect.umd.edu/~toh/spectrum/findpeaks.m

Orlikoff, R F (1998). The uses and abuses of electro-glottography, *Phonoscope* 1: 37–53.

Rothenberg, M (1973). A new inverse-filtering technique for deriving the glottal airflow during voicing. *J. Acoust. Soc. Am.* 53: 1632-1645.

Sjölander, K (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik 2003*: 93-96.

Timoney J, Lysaght T, Schoenwiesner M, McManus (2004). Implementing loudness models in Matlab. In *Proceedings of the 7th Int. Conference on Digital Audio Effects (DAFX-04)*. Naples, Italy.

Vieira, M N, McInnes, F R & Jack, M A (1996). Robust F0 and jitter estimation in pathological voices. In *ICSLP-1996*: 745-748.

Automatic estimation of pitch range through distribution fitting

Johan Frid and Gilbert Ambrazaitis
SOL, Lund University

Abstract

In this paper we present a method for automatic selection of 'Floor' and 'Ceiling' values for pitch analysis. The method involves fitting a log-normal tied mixture model to distributions of pitch values and using the resulting parameter values. The method's ability to deal with pitch estimation errors in spontaneous speech is demonstrated.

Introduction

When performing pitch analysis, it is common to limit the range of possible values by providing the analysis method with a minimum ('Floor') and a maximum ('Ceiling') value. Setting these parameters so that they are close to the speaker's actual pitch range will greatly improve the analysis as many types of pitch estimation errors will be minimized.

It is, however, somewhat troublesome to select good values. For an unknown speaker or sound file, one really has no idea of what the actual range of the speaker is. If one knows the gender of the speaker, it may be possible to make a rough guess at which parameter values to use (e.g., female: 100 Hz– 500 Hz, male: 75 Hz– 300 Hz), but this is still rather error prone, and for fully automatic analysis, information about gender may not be available.

Our approach to this task will be the method of bootstrapping; we first perform an initial pitch analysis with a very wide range (allowing for both male and female voice ranges; children's voices were ignored for the context of this paper), then we analyse the results, which will result in a guess at the parameters values, which we then feed back into the pitch analysis in order to get the final pitch values.

Earlier work

Sönmez et al. (1997) develop a pitch analysis model that allows estimation of pitch statistics from pitch tracks which contain 'errors' such as doubling and/or halving. The model is a log-normal tied mixture (LTM) with three log-normal distributions with tied means and variances and they estimate the parameters of the model using the Expectation-Maximization algorithm.

Their system is part of a larger speaker recognition system and they unfortunately present very little experimental results of the LTM itself, only some plots of one female speaker are shown. Furthermore, they do neither give any 'Floor' and 'Ceiling' values for their pitch tracker nor present any information on how to guess the starting values for the EM algorithm. On the other hand it is an interesting way to treat the problem; pitch estimation errors are handled not by filtering away the errors, but rather by incorporating them in the model. The model itself is described in enough detail for us to use it in this study.

Edlund & Heldner (2007) analyse percentiles of pitch distributions and compare them with means and standard deviations. They also use a log (the 'semitone') scale. They conclude that percentiles indeed are useful for estimating a speaker's relative pitch.

Pitch estimation errors are handled by looking at the intensity and the modality of the pitch distribution. They state that a filtering procedure removes frames in the lower modes of the intensity and (in case of a bimodal F0 distribution) F0 distributions but further details like intensity thresholds and their method of detecting bi-modality are left out. Furthermore, they do not state which initial 'Floor' and 'Ceiling' levels were used. As these details are not provided we have not tried to replicate their method in this study.

Furthermore, whereas it is clear that filtering away intensity frames with low values should lead to 'better' pitch ranges and also might be justified from production and perception perspectives, we think that there might be situations where it is advantageous to have a method that relies on one parameter – F0 – only.

De Looze & Hirst (2008) use a somewhat similar procedure. They also compare the percentiles (an alternative term for that is '100-quantiles' and the authors refer to them simply as 'quantiles') with manually estimated levels of maximum and minimum pitch.

They start off from an initial pitch estimate where Floor=60 Hz and Ceiling=750 Hz. This wide pitch range setting is necessitated by the need to initially accommodate both male and female pitch ranges (children may be another issue, but we ignore that in the context of this paper). They then present some empirical justification for setting Floor=0.75*q25 (25th percentile) and Ceiling=1.5*q75 (75th percentile) as minimum and maximum pitch values.

They do not seem to use the log or semitone scale and do not mention how to deal with bimodal F0 distributions and/or pitch halving/doubling. Later papers (De Looze & Hirst 2010) have suggested other variants of the formulae, for instance Floor=0.83*q15 and Ceiling=1.92*q65, but it is not clear whether this works better or not.

Their method may appear somewhat simpler than the other models. On the other hand, it is described in enough detail for us to replicate it.

Pitch distributions

Our strategy for estimating suitable Floor and Ceiling parameters involves looking at the distribution of pitch values for a given speaker. The distribution, roughly, provides us with a profile of a speaker's most used pitch levels and enables us to make a reasonable estimate of the speaker's actual pitch range.

The distribution of any variable may be analysed by constructing a histogram of the data. This assesses the probability distribution of a given variable by depicting the frequencies of observations occurring in certain ranges of values. The histogram is sometimes criticized (Wilkinson 1992) for being strongly affected by the choice of the number of bins (or sub-ranges) and also of being ineffective at depicting the true shape of the distribution. An alternative which often works better is to use the kernel density instead. Figure 1 shows the histogram and the density estimates of the distribution of log(F0) of one male speaker. F0 was estimated using Praat's AC method (Boersma 1993) with Floor=60 Hz and Ceiling=750 Hz. Histogram and density estimates were calculated using R (R Development Core Team 2009); the 'truehist' function in the MASS package (Venables & Ripley 2002) and the 'density' function in the (standard) stats package. The higher values (in the right hand side of the figure, about > 6.0) correspond to pitch levels above 400 Hz and are most likely pitch estimation errors.

It may be worthwhile here to point out that it of course would be easy to get rid of these errors by setting an arbitrary 'Ceiling' at, say, 250 Hz – but – the idea in this paper is exactly NOT to use arbitrary 'Floor' and 'Ceiling' values, but rather to estimate them automatically based on an analysis of all obtained pitch values.

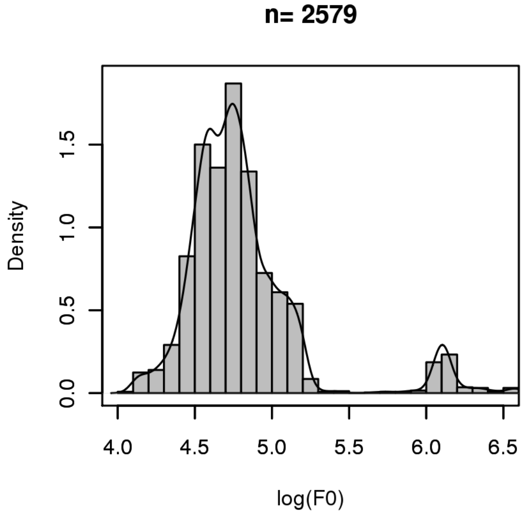


Figure 1. Histogram (vertical bars) and density (solid line) estimates of the distribution of log(F0) of 2579 voiced frames (one speaker).

Distribution fitting

The distribution estimate obtained with the kernel density described above is examined using the method of distribution fitting. Data (in this case density values) may be fitted by proposing a model that provides a description of the data. Some parameters of the model are allowed to vary (they are 'free') and an algorithm then tries to minimize the difference between the predictions of the model and the observed data by choosing suitable values. One common example of a model is the Gaussian (or normal) distribution, that has two free parameters, location and scale, whose starting values sometimes are estimated by the arithmetic mean and standard deviation. Since the distribution of log(F0) often is more similar to a normal distribution than the distribution of F0 and since we want to deal with pitch halving/doubling we will adopt the log-normal tied mixture of Sönmez et al (1997).

Starting values

Minimization algorithms need reasonable starting values. Since we use a wide range, the result it is likely to contain pitch estimation errors. Therefore the arithmetic mean and the standard deviation will sometimes be quite bad starting values.

Robust estimators

We therefore employ some more robust measures of location and scale of the data that can ignore some of the outliers in the data. Some common robust techniques include an 'M-estimator' for location and the 'Bi-weight mid-variance' for scale (Herrington 2002). Figure 2 shows the difference between model predictions of a log-normal model with arithmetically versus robustly calculated values.

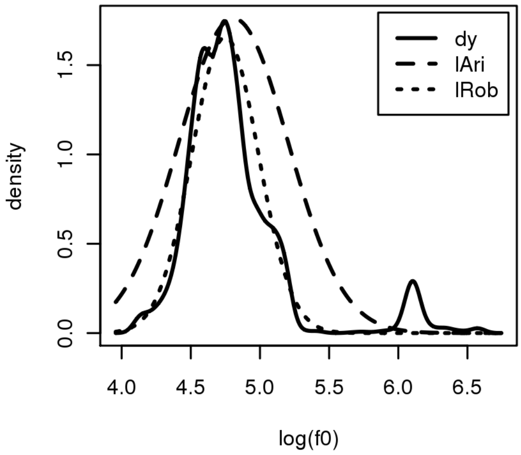


Figure 2. Density (dy) and predictions made by log-normal models with arithmetically (lAri) and robustly (lRob) calculated values.

Note how the robust ('lRob') model follows the log(F0) density values that we deem 'correct' much more closely than the arithmetic ('lAri') model. In this respect, we can say that the robust values are better since the model based on them provides a better description of the data than the model based on arithmetic values. In addition, note that the log-normal models appear normal in Figure 2, since the x-scale is logarithmic.

Mode and half-width-at-half-max

In some cases of bimodal pitch distributions, neither the arithmetic nor the robust method will give good starting values. We therefore also use the mode (the value that occurs the most frequently in a data set or a probability distribution) for location and the half-width-at-half-max (HWHM) for spread.

Since a bimodal distribution may arise either from halving doubling of pitch, any one of the peaks in such a distribution may be the 'correct' one. So we also need use the mode and the HWHM of the second-highest peak.

Outline of the method

Our proposed 'Distribution Fitting' model uses Praat for pitch analysis and R for models and distribution fitting. Here follows an outline:

- Perform an initial pitch analysis with a wide pitch range, accommodating for the pitch ranges of adult men and women. Following De Looze and Hirst (2008), we currently use Floor=60 Hz and Ceiling=750 Hz.
- Get a robust estimate of location and scale of log(F0).
- Estimate the mode and half-width-at-half-max of log(F0) (both raw data and first residual)
- Fit log-normal tied mixture models to the log(F0) distribution using non-linear least-squares with:
  1. robust location and scale
  2. mode and hwhm
  3. mode and hwhm of residuals of 2

  as starting values and 'plinear' and 'port' algorithms (provided by R)
- Reject 2 and/or 3 if their estimated 'location' is:
  - ○ too far away from the robust estimate of location (currently location +/- 0.9*scale)
  - ○ < 120 or > 375
- Return the model with the lowest residual sum-of-squares and calculate a new log-normal distribution with its location and scale parameters
- Calculate Floor and Ceiling from this new model: currently Floor=0.75 * 1st quantile, Ceiling=1.5 * 3rd quantile. These calculations are similar to De Looze and Hirst (2008).

Testing the method

We performed a comparison of 1) the method of De Looze & Hirst (2008), 2) a model based on the robust estimators and 3) the 'Distribution fitting' model described above. Method 2 differs from Method 1 in that instead of using actual quantiles, we estimate the 1st and 3rd quantiles of a log-normal distribution with the parameters values that the robust estimations result in. Below, we will refer to them as the 'lAri', 'lRob' and 'DF' models.

Estimation of 'Floor' and 'Ceiling'

We estimated Floor and Ceiling values with all three methods for 429 speakers (both male and female) from the public part of the Swedia 2000 (Aasa et al. 2000) database. Speech files contain roughly 30s of spontaneous speech. It is somewhat problematic to assess the performance of the methods as we do not have any notion of what the 'correct' values would be for these speakers (no 'Gold standard'). Figures 3-6 show some examples where the DF model makes a good choice but seem problematic for the non-fitting models. In each figure, the vertical lines show the 'Floor' and 'Ceiling' levels, the observed distribution is drawn with a solid line and the predicted distributions are shown using different line types.

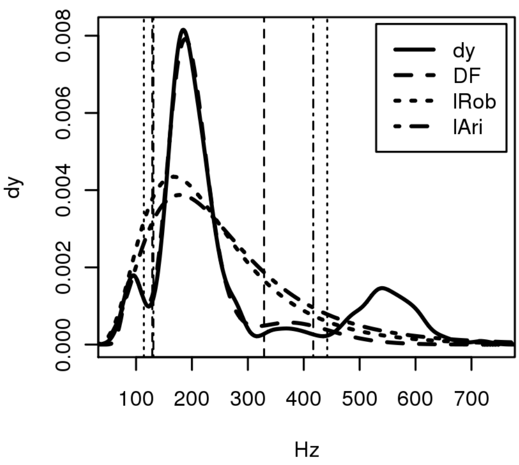


Figure 3. Female speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method.

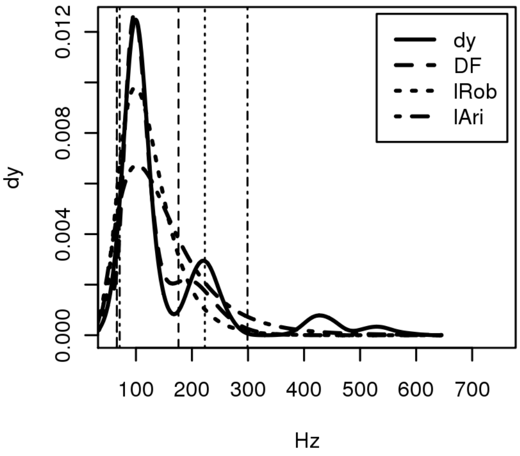


Figure 4. Male speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method. Distribution is influenced by pitch levels of a female interviewer.

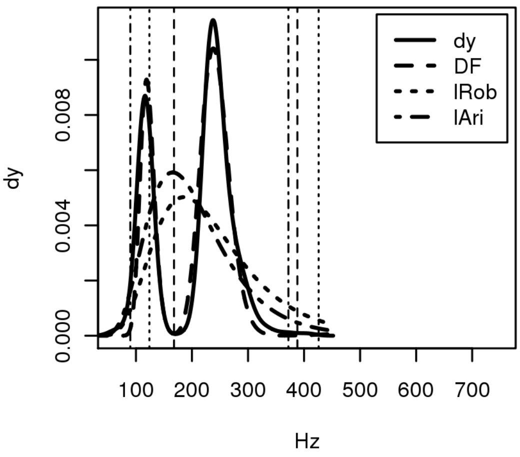


Figure 5. Female speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method.

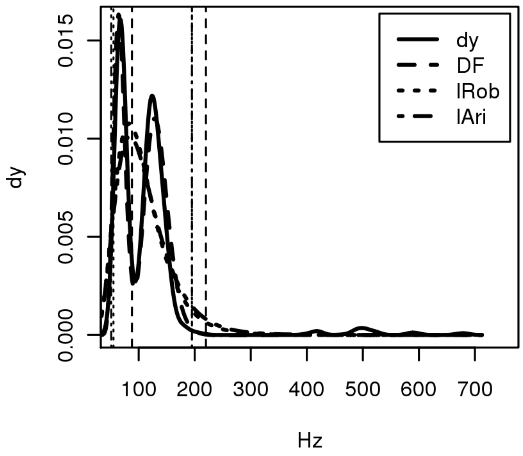


Figure 6. Male speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method.

Figure 3 shows a case where background noise increases the 'Ceiling' estimate for the non-fitted models. In Figure 4, the main speaker is male, but there is also a female interviewer. This, of course, adds frames in the female pitch range. The DF model handles this well. Figures 5 and 6 are examples where pitch halving lowers the 'Floor' estimates.

Visual inspection of plots like Figures 3-6 for all speakers indicate that the DF model always makes a reasonable 'Floor' and 'Ceiling' estimate.

Log-normality

We will show another aspect of the DF model: it truncates the data so that it becomes closer to a normal or log-normal distribution, which may be advantageous for further statistical analysis, comparisons, standardisations and grouping of data from different speakers.

In this sub-study we used the Swedia material described above, but also material from TIMIT (Garofolo et al. 1990). This database consists of much more controlled speech material, has 'cleaner' recordings and consequently contains much less pitch estimation errors. We estimate Floor and Ceiling levels by the DF and the lAri models for each speaker in each database. Then we truncate the data and compare the distribution of the remaining values with the log-normal distribution. This is done by comparing the quantiles of the data with the quantiles of the log-normal distribution (this is the same procedure as doing a 'QQ-plot') and computing the correlation coefficients between these quantiles. The method is described in Johnson (2008). Then we compare the correlation coefficient in scatterplots, see Figures 7 and 8.

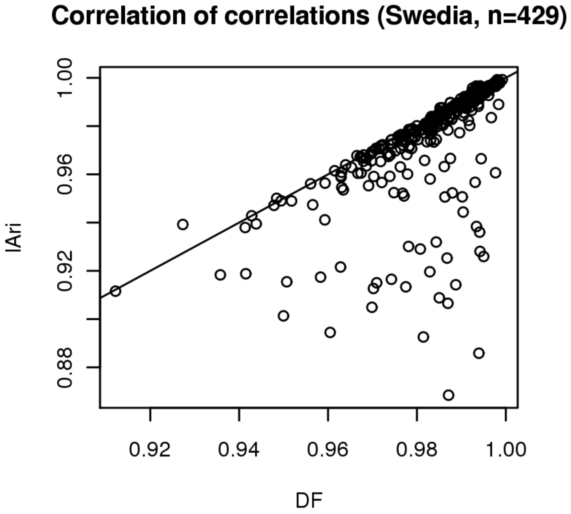


Figure 7. Scatterplot of correlation coefficients for the DF and lAri models with a log-normal distribution. Speakers from the Swedia database.

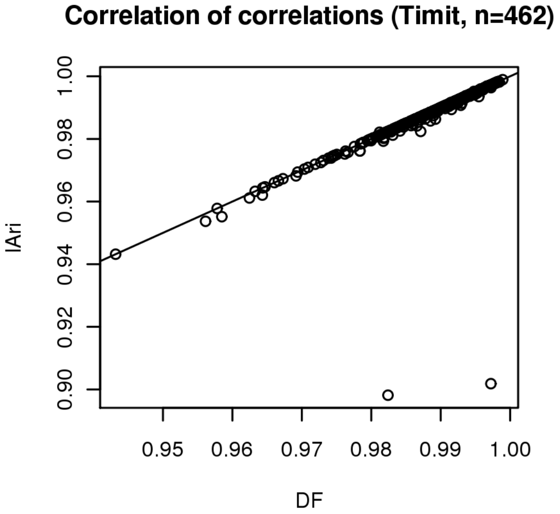


Figure 8. Scatterplot of correlation coefficients for the DF and lAri models with a log-normal distribution. Speakers from the Timit database.

In Figure 7 we see that in the majority of cases the two models show similar performace, more or less, but there is a also a noticeable amount of cases (around 8% of the data) where the correlation is higher with the DF model.

Figure 8 shows the result for TIMIT. Here we see much less difference, as might have been expected since pitch halving and doubling errors are scarcer. Yet, we see two cases of creaky speakers (in the lower-right hand corner of the plot), which are made to correlate well with the log-normal distribution with the DF model, but not with the lAri model.

Summary and conclusion

In this paper we have presented a procedure that determines reasonable upper and lower limits of F0 automatically from an initial F0 estimate with a very generous range. It is insensitive to the individual speaker's actual pitch range and accommodates both for men and women (we have not tested on children).

We show that compared to quantile-based models the DF model performs at least as well for clean, controlled speech and – based on a qualitative analysis – better for spontaneous speech in less controlled set-ups.

The method can be used for re-analysis of pitch with reasonable floor and ceiling values. It may also be used as a speaker's 'pitch range' to which one can 'normalize' or 'standardize' or determine a 'base-value' in the sense of Traunmüller & Eriksson (1995) (see also Traunmüller 1994 and Lindh & Eriksson 2007).

Further studies may include how to model the truncated values. A possible investigation is to see whether they are best modelled by a log-normal model or if there might be other models and/or transformations that describe the data better.

Acknowledgements

This work is supported by a grant from the Swedish Research Council (VR) project 2009-1566.

References

Aasa A, Bruce G, Engstrand O, Eriksson A, Segerup M, Strangert E, Thelander I & Wretling P (2000). Collecting dialect data and making use of them: an interim report from Swedia 2000. *Proceedings of Fonetik 2000*, University of Skövde, pp. 17-20.

Boersma P (1993) Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings* 17, pp. 97-110.

De Looze C & Hirst D (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *Proceedings of International Conference on Speech Prosody* (4 : 2008 avril 6-9 : Campinas, Brazil). 2008, pp. 135-138.

De Looze C & Hirst D (2010). Integrating changes of register into automatic intonation analysis. *Proceedings of Speech Prosody* (2010 : Chicago, USA) [Forthcoming].

Edlund J & Heldner M (2007). Underpinning /nailon/: automatic estimation of pitch range and speaker relative pitch. In C. Müller (Ed.), *Speaker Classification* II (Vol. LNAI 4441, pp. 229-242). Berlin, Germany: Springer-Verlag.

Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D & Dahlgren N (1990). DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology. NTIS Order No. PB91-505065.

Herrington R (2002). *Using Robust Mean and Robust Variance Estimates to Calculate Robust Effect Size* [WWW document]. URL http://www.unt.edu/benchmarks/archives/2002/july02/rss.htm

Johnson K (2008) *Quantitative Methods in Linguistics*. Oxford: Blackwell.

Lindh J & Eriksson A (2007). Robustness of Long Time Measures of Fundamental Frequency. *Proc. Interspeech* 2007, pp. 2025-2028.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sönmez M K, Heck L, Weintraub M & Shriberg E (1997). A Lognormal Tied Mixture Model of Pitch for Prosody-based Speaker Recognition. *Proc. EUROSPEECH 97*, Rhodes, Greece, September 1997, Volume 3, pp. 1391 – 1394.

Traunmüller H (1994). Conventional, biological, and environmental factors in speech communication: A modulation theory. *Phonetica* 51, pp. 170-183

Traunmüller H & Eriksson A (1995). *The frequency range of the voice fundamental in the speech of male and female adults*. Unpublished Manuscript (can be retrieved from http://www.ling.su.se/staff/hartmut/aktupub.htm).

Venables W N & Ripley B D (2002). *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Wilkinson L (1992). Graphical displays. *Statistical Methods in Medical Research*, 1, pp. 3–25.

Quantity in Central Standard Swedish and Fenno-Swedish

Pétur Helgason[a], Catherine Ringen[b], Kari Suomi[c]
[a]Department of Linguistics & Philology, Uppsala University, [b]Department of Linguistics, University of Iowa, [c]Faculty of Humanities, Oulu University
Authors' names in alphabetical order.

Abstract

Main findings and conclusions are summarised from a comparison of the durational realisation of quantity in Central Standard Swedish (CSS) and Fenno-Swedish (F-S) based on exactly the same word materials. Both varieties have the same basic quantity pattern, with a complementary durational relationship between the stressed-syllable vowel and the following consonant. But F-S has, in addition, quantity patterns that do not exist in CSS. These patterns are reviewed and the conclusion is drawn that they are similar to some Finnish quantity patterns. Altogether, the results and the review strongly suggest that the F-S quantity patterns have been influenced by Finnish.

Introduction

Swedish is a quantity language in which quantity is contrastive only in syllables with primary stress. In most varieties, a phonetically short vowel in the stressed syllable must be followed by a consonant that is phonetically long (or by two different consonants). In addition, if a phonetically long vowel in the stressed syllable is followed by a consonant, the consonant is always phonetically short. This entails that there is a complementary durational relationship between the stressed-syllable vowel and the following consonant (Elert 1964). For example, in *väg* [veːg] 'road', the vowel is long and the following consonant is short (V:C), whereas in *vägg* [vegː] 'wall' the vowel is short and the following consonant is long (VC:). These restrictions obtain for CSS. This is also the dominant pattern in F-S, but there are also deviations in F-S that will be discussed below.

F-S has been spoken in Finland for centuries. Since the 19th century a phonetically Finnish-based standard F-S has been attaining a prestige status. Consequently, there are many differences in pronunciation between F-S and Sweden-Swedish. Kuronen & Leinonen (1999) note that "even though the segmental differences between Finland-Swedish and Sweden-Swedish are clearly audible the main and for the listener most obvious difference is in the prosodic character of these dialects". One reason for this circumstance, among many others concerning e.g. the rules of accentuation and its phonetic realisation, is that certain F-S durational patterns are very much like Finnish durational patterns.

In Finnish, quantity is contrastive for both vowels and consonants, independently of each other and of stress (which is fixed to the initial syllable). Contrastively long segments are best analysed as sequences of two identical phonemes, and there are words like *ta.ka* [t̪ɑkɑˑ] /t̪ɑkɑ/, *taa.ka* [t̪ɑːkɑ̆] /t̪ɑɑkɑ/, *tak.ka* [t̪ɑkːɑ̆] /t̪ɑkkɑ/, *taak.ka* [t̪ɑːkːɑ̆] /t̪ɑɑkkɑ/, *ta.kaa* [t̪ɑkɑː] /t̪ɑkɑɑ/, *taa.kaa* [t̪ɑːkɑː] /t̪ɑɑkɑɑ/, *taak.kaa* [t̪ɑːkːɑː] /t̪ɑɑkkɑɑ/ (with syllable boundaries indicated in the orthographic forms). Consequently, e.g. *takka* and *taakka* can be given as CVCCV and CVVCCV, respectively; for more information on the phonological interpretation of Finnish quantity see Suomi, Toivanen & Ylitalo (2008: 39-42) and the references therein.

The CSS materials of this comparison come from Helgason & Ringen (2008) and the F-S materials come from Ringen & Suomi (in preparation). The primary objective in these studies was to investigate the phonetic realisation of the fortis ~ lenis distinction in the respective varieties. The recording procedures and segmentation criteria were highly similar across the two experiments (cf. Helgason & Ringen 2008). The CSS native speakers were three female and three male native speakers who have lived in Stockholm most or all of their lives. For the 12 F-S speakers, six female and six male, F-S was the first language and the language of education, as it was for both of their parents. The F-S speakers, who are all fluent in Finnish, come from all three areas in mainland Finland in which F-S is spoken: Uusimaa/Nyland, Turunmaa/Åboland and Pohjanmaa/Österbotten. The ages of the speakers in both varieties ranged from early twenties to late forties.

The speakers read a word list twice. The list contained both fortis and lenis stops in initial, medial and final positions. A word list was chosen instead of words embedded in carrier sentences in order to get utterance-initial stops. The initial stops were followed and the medial and final stops were preceded by short and long vowels in the stressed syllable; all target words had stress on the initial syllable. The present comparison includes only the 34 test words that were shared between the F-S and CSS data sets and excludes other test words.

A clear difference between CSS and F-S is that, unlike F-S speakers, speakers of CSS frequently preaspirate postvocalic fortis stops. Of the six CSS speakers, two female speakers had a mean preaspiration duration exceeding 55 ms, two male speakers had a mean preaspiration duration less than 35 ms, and the remaining two speakers (a female and a male) had intermediate preaspiration durations. Here we have included the preaspiration in the vowel duration. Post-aspiration of initial stops, however, was not included in the duration of the following vowel.

Main findings and conclusions

Eight word types (monosyllabic vs. disyllabic, short vs. long vowel, fortis vs. lenis postvocalic stop) were examined. In each word type, in CSS, the stressed-syllable vowel was either short or long, and the following fortis or lenis stop was then either long or short. The word types were monosyllabic with a short vowel, e.g. *däck*, *dagg*; monosyllabic with a long vowel, e.g. *tak*, *lag*; disyllabic with a short vowel, e.g. *bytte*, *ledde*, and disyllabic with a long vowel, e.g. *baka*, *väga*. In the *baka* words, the F-S speakers differed in behaviour. Eight speakers systematically pronounced [bɑɑkɑ] (with a short medial consonant) and four systematically pronounced [bɑɑkkɑ] (with a long medial consonant). We discuss the *baka* words separately below.

Several consistent differences between CSS and F-S were observed. In both monosyllabic and disyllabic words, short consonants had shorter and long consonants had longer durations in F-S than in CSS. The only exception to this generalisation is that in the [bɑɑkkɑ] rendition the long consonant did not have a longer duration than the corresponding consonant in CSS. Another consistent finding was that in both monosyllabic and disyllabic words, short vowels always had shorter durations in F-S than in CSS. All these observations can be summarised by stating: in comparison to the CSS speakers, the F-S speakers exaggerate the syntagmatic quantity difference between the stressed-syllable vowel and the following consonant, the structure within which the Swedish quantity contrast is realised.

Direct comparison of monosyllabic words between Finnish and Swedish is difficult as monosyllabic words are not common in Finnish and have hardly been investigated. Nevertheless, the overall differences between the quantity systems of Finnish and Swedish seem to account for the syntagmatic quantity exaggeration observed in F-S. Bilingual F-S speakers also hear and speak Finnish in which a phonetically short vowel in the stressed syllable can be followed by a short consonant or a long consonant, and similarly a phonetically long vowel in the stressed syllable can be followed by a short consonant or a long consonant, and all these possibilities are mutually contrastive. In Finnish, then, speakers of F-S are used to a system in which duration matters more than in Swedish, in both the stressed-syllable vowel and in the following consonant (as well as later in the word). They have learned that in Finnish, segments that are intended to have short duration must be clearly short, and that segments that are intended to have long duration must be clearly long, because there is less predictability in Finnish than in Swedish in this respect.

Direct comparison of disyllabic words in F-S, CSS and Finnish is not possible in all cases because historically Finnish lacks lenis stops, and studies involving recent loanwords in which lenis stops occur for some speakers have not been undertaken. This also means that, as far as lenis stops are concerned, speakers of Fenno-Swedish can be little influenced by Finnish. But disyllabic words with fortis medial stops can be directly compared.

There are two disyllabic word types in the present material that have, in segmental and gross durational terms, lexically numerous equivalents in Finnish. One of these types is represented by the word *backa* (short vowel + long consonant). Spoken in either CSS or F-S, this type sounds durationally very much like Finnish CVCCV words, e.g. *pakko*. The other type is represented by the word *baka* (in CSS: long vowel + short consonant). If spoken in CSS, the medial consonant sounds "half-long" to Finnish ears.

Recall from above that our F-S speakers pronounced the words of this type, exemplified by *baka*, either with the durational pattern [bɑɑkɑ] or with the pattern [bɑɑkkɑ]. To Finnish ears, these pronunciations sound durationally very much like Finnish CVVCV (e.g. *raaka*) and CVVCCV (e.g. *taakka*) words, respectively. Table 1 shows the mean V/C durational ratios for the *baka* word type computed from our CSS speakers' productions as well as from the [bɑɑkɑ] and [bɑɑkkɑ] renditions by the F-S speakers. For comparison, Table 1 also gives ranges of corresponding mean ratios for Finnish reported in Lehtonen (1970), Suomi, Toivanen & Ylitalo (2003), Suomi & Ylitalo (2004) and Ylitalo (2009), excluding the contrastively accented words in the last reference.

Table 1. Vowel/consonant duration ratio in baka *type words in CSS and two groups of F-S speakers compared with the range of ratios observed for Finnish CVVCV and CVVCCV words.*

	Sw	Fi
CSS [bɑːkɑ]	1.22	
F-S [bɑɑkɑ]	2.39	1.95-2.74 (CVVCV)
F-S [bɑɑkkɑ]	1.00	0.80-0.94 (CVVCCV)

Table 1 shows that the V/C durational ratio is larger in CSS than it is in F-S [bɑɑkkɑ], and very much smaller than it is in F-S [bɑɑkɑ]. It can also be seen that these ratios in F-S are more similar to Finnish ratios in CVVCCV and CVVCV words, respectively, than they are to the CSS ratio. However, the major difference here between CSS and F-S is that while CSS has only one durational pattern for the *baka* words, F-S has two patterns.

The variation between the pronunciations [bɑɑkɑ] and [bɑɑkkɑ] is determined by both regional and sociolinguistic factors. The usage of [bɑɑkɑ] is considered to be "provincial", and often a speaker switches to [bɑɑkkɑ] after moving to the capital area. According to Reuter (2008), the longer the medial consonant in such words, the higher the socioeconomic status of the speaker. The Swedish spoken in and around Helsinki is usually considered Standard Fenno-Swedish. Our speakers who produced [bɑɑkkɑ] come from Helsinki and Turku and the region between these cities, but other speakers from roughly the same areas produced [bɑɑkɑ] instead. A generalisation without exceptions is that the five speakers who are *not* from the broad Helsinki/Turku area always produced [bɑɑkɑ] (four of them from Österbotten).

A further consistent finding was that, in both monosyllabic and disyllabic words, long vowels always had shorter durations in F-S than in CSS; the only exception was the [bɑɑkɑ] rendition in which the long vowel did not have a shorter duration than the corresponding vowel in CSS. Even so, the shorter C durations in F-S compensated for this and consequently the V:/C durational ratios were larger in F-S than in CSS for these words as well.

A final consistent inter-variety difference was that, in monosyllabic words, the fortis ~ lenis differences were greater, for both vowel and stop durations, in CSS than in F-S; in disyllabic words the comparison could not be made as the fortis stops exhibited two distinct patterns in F-S. Even here the explanation may be Finnish: Since Finnish practically lacks the fortis ~ lenis distinction, its salience may therefore have been reduced in F-S.

In CSS the observed fortis/lenis closure duration ratios were 1.68 in long final stops in monosyllabic words, 1.94 in short final stops in monosyllabic words, 1.44 in medial long stops in disyllabic words, and 2.31 in medial short stops in disyllabic words. These are very large ratios. It is as if CSS strives for an extremely clear perceptual contrast between its fortis and lenis stops: utterance-initially (where occlusion duration cannot be estimated) the former are aspirated and the latter are prevoiced (Helgason & Ringen, 2008), a rare combination in the languages of the world, in other positions the contrast is supported, in addition to voicing parameters, by a large durational difference.

Discussion

Our major finding was that there are several systematic durational differences between CSS and F-S. These are just an addition to the long list of segmental and prosodic differences between the two varieties given and exemplified by Kuronen & Leinonen (1999, 2008). We wish to argue that the prosodic differences are not the result of a *sua sponte* development of F-S. It cannot be an accident that to the extent F-S differs from CSS, the differences are almost always in the direction of Finnish.

It was observed that in those word types in which the CSS and F-S quantity systems are phonologically identical (i.e. all word types except the *baka* words), there were nevertheless consistent phonetic differences in that F-S speakers exaggerated the syntagmatic quantity difference between the stressed-syllable vowel and the following consonant. In addition to these phonetic differences, the repertoire of phonological quantity patterns is larger in F-S than in CSS. CSS only has the two patterns VC: and V:C as in *backa* and *baka*. Even recent loanwords must have one of these two quantity patterns; a given word may have alternative pronunciations with respect to quantity, but the alternation is between the two native patterns, e.g. *foto* may be pronounced as either [fuːtu] or [futːu]. F-S in contrast has four quantity patterns.

Where CSS has *baka* [bɑːkɑ], F-S has the two alternative patterns [bɑɑkkɑ] and [bɑɑkɑ]. The *backa* [bɑkːɑ] type is very similar in both varieties. The fourth quantity pattern in F-S is VC, i.e. a short stressed-syllable vowel followed by a short consonant. This pattern is called *kortstavighet,* 'shortsyllabicity' in Swedish. In some rural F-S dialects, at least for older speakers, the VC pattern has survived from Proto-Nordic, and e.g. *baka* is pronounced [bɑkɑ]. In non-rural standard varieties of F-S, the VC pattern occurs in a large number of both fully native and borrowed words including *foto* [foto]; Kuronen & Leinonen (2010: 14-15) list seven groups of such words, and in some words there is vacillation between pronunciations with a short and a long vowel. An additional dimension in this variation is that the long-vowel pronunciation of these words usually only occurs in a formal register.

Thus while CSS has two possible quantity patterns in initially stressed disyllabic words Finnish has eight, and F-S uses four patterns. Although there may be no minimal pairs yet involving the *foto* and *baka* types of words (CVCV against CV:CV ~ CV:C:V), and although the [bɑɑkɑ] ~ [bɑɑkkɑ] variation is so far only sociolinguistic, the door is open for four distinct quantity patterns.

References

Elert, C (1964) *Phonologic Studies of Quantity in Swedish*. Sweden: Almqvist & Wiksell.

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

Kuronen, M & Leinonen, K (1999). Fonetiska skillnader mellan finlandssvenska och rikssvenska. *Svenskans beskrivning 24*. Linköping Electronic Conference Proceedings.
URL: http://www.ep.liu.se/ecp/006/011/.

Kuronen, M & Leinonen, K (2008). Prosodiska särdrag i finlandssvenska. In: M Nordman, S Björklund, Ch Laurèn, K Mård-Miettinen & N Pilke, eds, *Svenskans beskrivning 29*. Finland: Skrifter utgivna av Svensk-Österbottniska samfundet 70, 161-169.

Kuronen, M & Leinonen, K (2010). *Svenskt uttal för finskspråkiga*. Finland: Juvenes Print – Tampereen yliopistopaino Oy.

Lehtonen, J (1970). *Aspects of quantity in standard Finnish.* Studia philologica Jyväskyläensia VI. Finland: Jyväskylä University Press.

Reuter, M (2008). Mikael Reuter vägrar vara språkpolis. Newspaper article in *Vasabladet*, June 6th 2008. URL:
http://www.vasabladet.fi/story.aspx?storyID=21513.

Ringen, C & Suomi, K (in preparation). The voicing contrast in Fenno-Swedish.

Suomi, K, Toivanen J & Ylitalo, R (2003). Durational and tonal correlates of accent in Finnish. *Journal of Phonetics,* 31, 113-138.

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

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

Ylitalo, R. (2009). *The realisation of prominence in three varieties of Standard Spoken Finnish*. Ph.D. dissertation. Acta Universitatis Ouluensis B Humaniora 88. URL: http://herkules.oulu.fi/isbn9789514291142.

C. Ringen's research was supported by a National Science Foundation (grant BCS-0742338).

Fo curves – smooth, seamless yet pulsed?

Björn Lindblom
Department of Linguistics, Stockholm University, 10691 Stockholm, Sweden

Abstract

This paper presents pilot data and analyses suggesting that Central Swedish F0 contours can be described with good numerical accuracy as sequences of syllabic overlapping quasi-Gaussian pulses. Each such building block is temporally aligned with a given syllable. It is specified by its 'amplitude' (peak frequency) and the temporal extent of its rise and fall slopes. According to the proposed account an F0 curve is seen as the sum of the bell-shaped pulses and a declination ramp.

In revisiting classic publications on respiratory and phonatory speech processes one notes that, although Stetson's 'chest pulse' notion is rejected by most phoneticians today, one nonetheless finds a great deal of evidence for increased P_s and EMG activity in stressed syllables (language examples: English and Swedish). Curiously, some of that evidence comes from the work by Ladefoged, one of Stetson's fiercest critics.

Are the results of the present modeling merely a fluke outcome of the curve fitting exercise, or do they tell us something significant about the motor control of prosodic phenomena in Swedish?. The following facts lend preliminary support to the latter possibility. ***1****. Syllables with main and secondary lexical stress show larger peaks than syllables with weak or no lexical stress.* ***2****. Pulses overlap in time thereby often losing their original discrete nature and hiding a possible pulsatile origin of F0 characteristics.* ***3****. The bell-shaped pulses contribute a rise-fall component to the output F0 trace and seem well suited to handling the correlates of the tonal accents. As expected the main grave-acute difference was found to be one of timing.* ***4.*** *One pulse per syllable was sufficient. It occurred within the time frame of the syllable. However the surface location of an F0 peak was sometimes temporally shifted away from its associate syllable (as seen in 4, 40, 400 and 32, 320, 3200 sequences). This phenomenon turned out to be successfully predicted by the syllable-aligned superposition model.*

The notion of 'chest pulse'

Stetson and Ladefoged

In 1951 Stetson proposed that there is an identifiable ballistic '*chest pulse*' for each syllable. He based this claim on measurements of rib cage movements, air pressure in the trachea and the lungs and on some EMG records. Partly, we may presume, he also used his intuition.

Ladefoged's *Three areas of experimental phonetics* (1967) contains a summary of research that refutes Stetson's hypothesis. Ladefoged found Stetson's work technically unreliable and was unable to identify chest pulses in his own experimental records: "*As might be expected from our electromyographic studies, we could find no correlation between subglottal pressure and syllables. It is clear, not only from the records we have been discussing here, but also from hundreds of others, that Stetson (1951) is wrong in claiming that there is a relation between respiratory activity and the syllable …*" (Ladefoged 1967, p 46-47).

He continues: "*Thus figures 17 and 18, which show even the small variations in subglottal pressure due to each opening and closing of the vocal cords, do* ***not show peaks*** *of subglottal pressure* ***which can be correlated with the individual unstressed syllables****.*

We can only presume that, in so far as Stetson's records sometimes indicate different results, it must have been because his subjects were not talking in a normal conversational tone. It seems probable that they were talking more loudly, slowly and distinctly than is customary.

In our opinion ***there is certainly insufficient basis for a chest pulse theory of the syllable in normal speech****.*"

Subsequently the field of phonetics has accepted Ladefoged's criticism (cf Lehiste

1970:109). Currently it would no doubt take a dim view of any attempt to revive the chest pulse notion.

Let us turn to Ladefoged’s Figure 18 which compares subglottal pressure for statements and questions and for disyllabic noun-verb pairs distinguished by stress placement: *‘digest’*, ‘*torment’*, ‘*pervert’*, ‘*survey’*. We here focus on the strong-weak vs weak-strong patterns of the nouns and verbs. Ladefoged notes that there is only one P_s peak per word. Moreover, in nouns the peak occurs on the first vowel; in verbs on the second. “*It is apparent that* ***every stress is accompanied by an extra increase of subglottal pressure***” (p 46).

We should also examine more closely some of his EMG findings, for instance his Figure 8 which shows the activity of the (expiratory) internal intercostals during ‘*The old man doddered along the road*’. A single motor unit was captured and its instantaneous frequency was calculated in impulses per second. This parameter is plotted in Figure 1

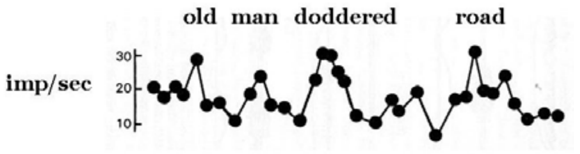


Figure 1. Firing rate of single internal intercostal motor unit as a function of time [adapted from Figure 8 in Ladefoged 1967].

The stressed words are indicated at the top of Figure 1. It can be seen that the stresses are clearly associated with greater activity in the respiratory muscle under observation.

In two more recent publications Ladefoged resumed his EMG studies to replicate his findings (Ladefoged & Loeb 2002, Ladefoged 2005). Single unit measurements were made in the internal intercostals during *‘The old man doddered along the road’*. A closely similar result was obtained. His overall conclusion (boldface ours): “ ***stressed syllables*** *may or may not have a greater intensity or a higher pitch, and they may or may not have bursts of internal intercostals activity. However, they* ***always use greater respiratory energy***.” (Ladefoged 2005:20).

Singers and actors

Sundberg et al (1995) contains a diagram of F0, sound level and esophageal pressure recorded as a professional bartone sings a descending scale with three tones on each scale tone and with emphasis on the first beat in each bar. The subglottal pressure trace reflects this prominence by a clear increase.

Another example is the oesophageal pressure variations of a speaker who recites “*Kung Karl, den unge hjälte, Han stod i rök och damm…...*” in a loud, clear and overarticulated style (‘stage speech’). The listener perceives a strong prominence on *‘Karl’*, *‘ung’* and *‘hjält’* and again the oesophageal pressure trace shows high values for those syllables.

From chest pulse to stress pulse?

Ladefoged’s criticism of Stetson’s methodology is well taken. And we have no reason to doubt the absence of measureable correlates of ‘pulses’ in Ladefoged’s data, e g, no peaks of subglottal pressure which can be correlated with the individual unstressed syllables.

The curious thing though is that his own EMG and P_s measurements do demonstrate that some sort of ’pulse’ (extra respiratory activity) is associated with the syllable. Would it be fair to say that Ladefoged’s data show the existence of a ‘chest pulse’ (“greater respiratory energy”) at least in stressed syllables in English? In other words a ‘stress pulse’?

I believe that to be a reasonable interpretation. Below its implications for F0 analysis are explored.

A pilot experiment: Procedures

The data of this report come from a male (phonetically trained) speaker of Central Swedish (= yours truly). He was recorded producing words and short phrases selected as follows.

Speech samples and recordings

Utterance length: From monosyllabic up to five syllables. Prosodic patterns: Grammatically legal permutations of four prosodic syllable types: (a) *main stress* (‘huvudtryck’) and acute accent (in SAOB notation a ‘4’ syllable); (b) main stress (‘huvudtryck’) and grave accent (a ‘3’ syllable); (c) *secondary stress* (‘starkt bitryck’, a ‘2’ syllable); (d) *weak or no stress* (‘svagtryck’, here denoted by ‘0’). Words with the grave accent were consistently compounds with their obligatory secondary stress.

The subject read pairs of words/phrases: a meaningful word or phrase (examples in Table 1) followed by the corresponding test word as a ‘*re-iterant*’ form - a (nonsense) word that has the same prosodic pattern as the model word and that uses segmental content in a controlled way.

4	sup
40	fjäder
04	pannå
400	kritiker
040	kaninen
004	logistik
4000	serierna
0400	botaniker
0040	intressera
0004	professionell
44	röd bil
444	Bengt går fort
440	Ulf läser
404	bra musik
044	Katrin vet
4004	Ted muckar gräl
40004	Per leker polis

32	kråkspark
320	brevlåda
302	motorbåt
032	miljövård
3200	busgrabbarna
3020	Londongatan
3002	väderprognos
0320	rabattkortet
0302	polispiket
0032	pyramidform

Table 1. Examples of prosodic patterns investigated. The numerical notation of stress and tonal accents is similar to that of SAOB (Elert 1970).

At least five tokens of each word or phrase were recorded. A later check of the recordings indicated that the subject had successfully maintained constant normal vocal effort and a pattern of uniform rhythm and intonation for the majority of speech samples.

kritiker	[ˈlɑ:nanan]
MacMillan	[laˈmɑ:nan]
aladåb	[lalaˈnɑ:n]

Table 2. Examples of model words and their 're-iterant' variants.

Re-iterant items were constructed by replacing single consonants and consonant sequences by [l]. [m] or [n] and by substituting [ɑ:] for vowels with main or secondary stress and [a] for other vowels. The use of sonorant, less constricted consonants reduces microprosodic perturbations of the F0 curve. Examples of *re-iterant* forms are given in Table 2.

F0 tracking & deriving average contours

Waveform files (sampling rate = 44100 Hz) were opened in Wavesurfer 1.7.5 for F0 analysis using the ESPS autocorrelation option. The output table contained F0 sampled every 10 ms. Time locations of vowel onsets and offsets were also tabulated.

For each test item repetitions were lined up in EXCEL side by side for synchronization and derivation of an average contour. As line up points vowel onsets were used. The averaging was performed across tokens once a syllable-by-syllable synchronization had been achieved. The alignment procedure was to locally shift the time scale of each syllable so as to match its vowel onset to the corresponding vowel onset of a reference token. This meant erasing or interpolating between values. These time shifts were generally small and were preferably made at points of minimum rate of F0 change. The F0 curves of this report are all derived by means of this procedure.

Results

Declination component

The F0 traces consistently showed a declination effect. To fit pulses to the Fo contours more conveniently I looked for a method to remove the declination component.

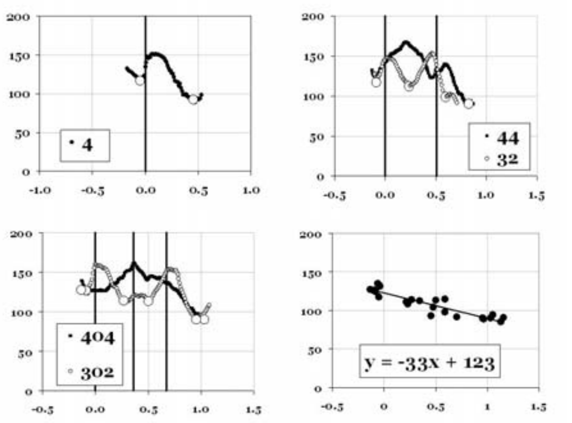


Figure 2. The declination ramp. Average F0 curves for selected test items. Vowel onsets are indicated by the vertical lines. Large open circles show points that were measured to quantify the declination effect. Time zero occurs at the first vowel onset. Bottom right: When these points are plotted on the same diagram they form a linear cluster with a falling slope of 33 Hz per second.

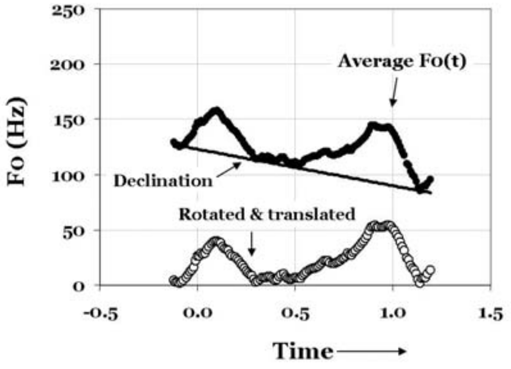


Figure 3. Data preparation. Removal of declination for curve fitting.

In the further processing of the data the declination component was removed by rotating the pattern so as to make it horizontal and translating it so as to place its minimum points on the zero line. Figure 3 summarizes the procedure.

The pulse

For a mathematical description of the bumps of the rotated and translated curves (cf Figure 3) I chose the Gauss curve:

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

letting x equal time and redefining μ as the time location of the pulse peak and using σ as a measure of the temporal extent of the rise and fall segments. In most cases the σ values for the rise and the fall were not equal implying asymmetrical pulse shapes.

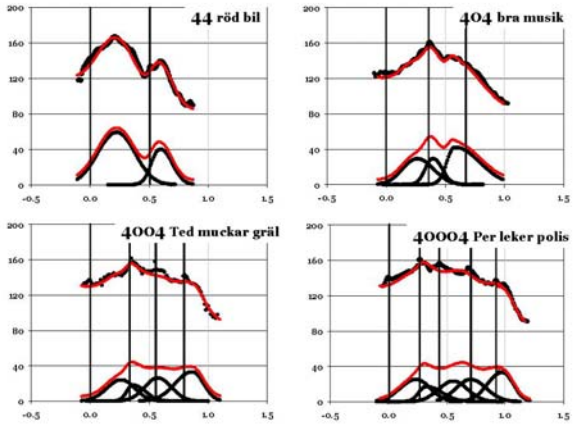


Figure 4. Phrases containing syllables with accent I. At the top of each panel: F0 contour (black) and the calculated curve (red). Below: Pulses individually (black) and summed (red).

Modeling the tonal accents

Figure 4 presents four phrases whose initial and final syllables have main stress and the acute accent. The top of the panels compares the observed F0 contour with a calculated curve. As can be seen there is a rather close match.

The bottom part shows the individual pulses (black) and a curve representing their sum (red). This curve was translated and rotated (=putting the declination back in) to derive the calculated trace.

These phrases were spoken in a 'legato' fashion. Note for the 44 pattern (top left panel) that there is no juncture between 'röd' and 'bil'. Nonetheless the F0 trace exhibits a valley in between the peaks which the pulse model handles satisfactorily.

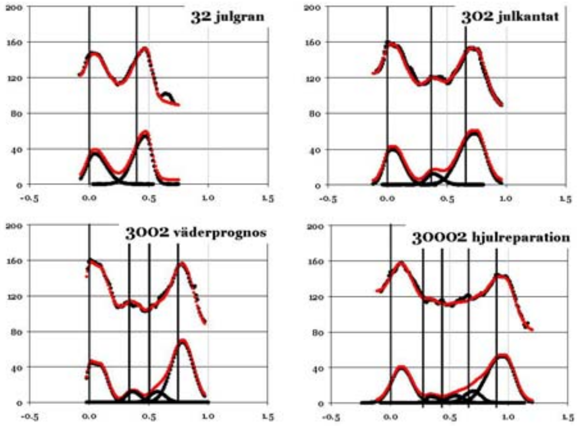


Figure 5. Phrases containing syllables with accent II.

Figure 5 is analogous to Figure 4. It presents four words with the compound 3(-)2 pattern. Initial syllables have main stress and the grave accent. Final syllables have secondary stress. Again we see a close match.

Figures 4 and 5 are representative of all test items analyzed.

Figure 6 applies the pulse analysis of two phrases differing in the placement of contrastive stress.

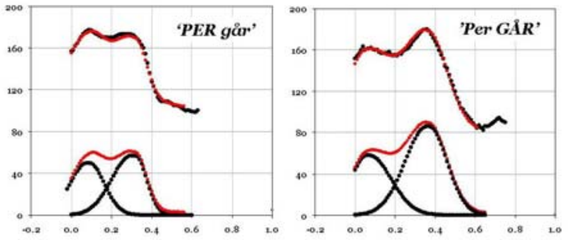


Figure 6. Two '44' sequences with emphatic stress on the first and the second syllable respectively.

Comparing grave and acute patterns

Figure 7 presents a comparison of the phrase '*grå ben*' (two adjacent acute accents) and the 32 compund '*gråben*'. Again it is noteworthy that the 44 pattern was spoken 'legato' without a juncture. Nonetheless there is a dip in F0 between the peaks - a fact which is compatible with the present pulse analysis.

The grave/acute contrast is here mainly linked to the timing of the first stress pulse and to the faster fall of the grave pulse.

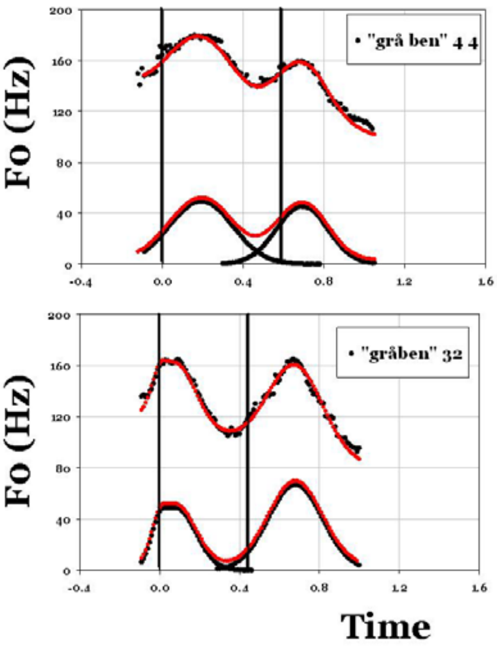


Figure 7. A ‘minimal pair’ illustrating the similarity between a 44 phrase and a 32 pattern.

Riad (1998) has proposed that historically the emergence of Scandinavian tonal accents is linked to a situation with two adjacent stresses and a need to resolve this ‘stress clash’ (‘betonginskrock’). It is interesting to place his discussion of tone accent origins in the context of the pulse model. In a sense the top 44 phrase in Figure 7 represents a stress clash. The reorganization needed to go from the 44 to the 32 pattern is a moderate one: a matter of a small timing shift and possibly a rate change in the fall of the first peak.

Peak delay

In the curve fitting results I find that the temporal alignment of pulse peaks shows some variation relative to its associated syllable and corresponding vowel onset. but the effect is small and so far nothing systematic about it has emerged.

Surface peaks on the other hand do behave in a fairly regular fashion. Figure 7 compares three words with stress and acute accent on the first syllable differing in the number of syllables per word: 4 , 40 , 400 as in ‘*sup*’, ‘*fjäd*er’ och ‘*kritiker’*. As more syllables is appended to the first stressed one the observed peak is progressively delayed. The second peak of words with the grave accent under goes a similar shift, e.g. ‘*anmäl’*. ‘*anmäla’’*, ‘*anmälare’’’*. This phenomenon can be given an explanation in terms of the pulse analysis.

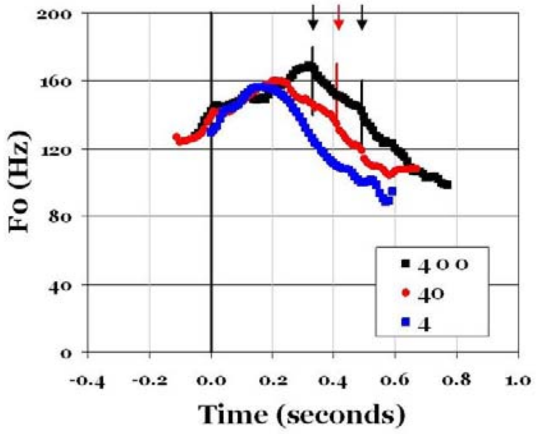


Figure 8. F0 peak delay as a function of the number of syllables following. Zero time occurs at the vowel onset in the first syllable. The vertical arrows indicate later vowel onsets..

Figure 8 reveals that, as the pulses of the following unstressed syllable(s) are added to the falling slope of the stressed pulse, the surface peak is delayed relative to its underlying pulse peak,

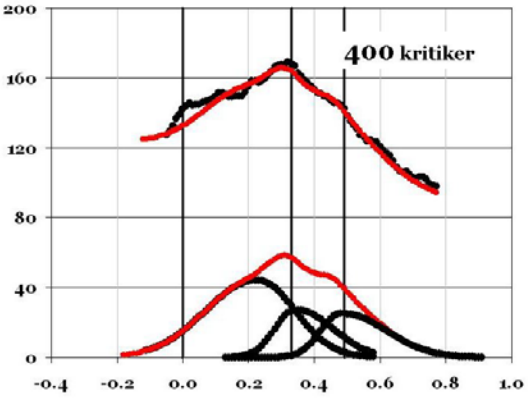


Figure 9. Pulse analysis of a ‘400’ pattern, ‘kritiker’ demonstrating how adding unstressed pulses on the falling slope of the stressed pulse delays the surface peak relative to its underlying pulse peak.

Discussion

Independent motivations for pulses

Fitting the pulse model to the present F0 data was easy and could done with reasonable accuracy. It may be remarked that this is not too surprising given the number of free parameters available. Still we need to ask if the model may not also have a more interesting, independent motivation. Perhaps it works so well because it is compatible with general facts about the nature of stress and syllables?

Nature of stress

Let us begin even further back than Stetson looking at Jespersen's view of stress. He writes (1926:119): *"Akzent (Druck) ist Energie, intensive Muskeltätigkeit, die nicht an ein einzelnes Organ gebunden ist, sondern der gesamten Artikulation ihr Gepräge gibt."* In other words he views *"Druck als Gesamtenergie"* (p 120). Let me elaborate on this a bit.

It appears clear that the motor organization of a syllable involves simultaneously coordinating articulatory, phonatory and respiratory processes. Activating this three-part system inevitably requires physiological effort implying that also unstressed syllables have some degree of 'stress'. It is also evident that the stress parameter is not a continuous function of time but is discretely updated once every syllable.

According to Jespersen greater stress radiates more energy into all subsystems. As a result articulatory movements become more extensive, vocal fold tension increases and traces of a higher level of respiratory performance can be detected in EMG and P_s records. This scenario is summarized in Figure 10.

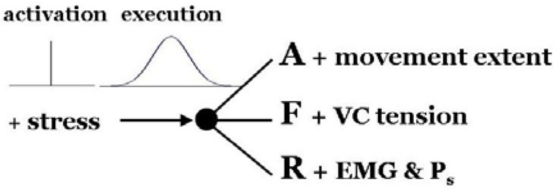


Figure 10. An attempt to summarize Jespersen's theory of stress and to identify a possible origin of the F0 pulses postulated in the present study.

Speculations

Where does the F0 pulse come from?

A possible hypothesis: From the stress pulse which controls the activity level not only of the respiratory system but also of phonatory and articulatory processes. I claim that the rise-fall charateristics of the F0 pulse derive from a stress-induced increase in vocal fold tension (and a slightly higher P_s).

Where did the tonal accents come from?

A second hypothesis: From the stress (and F0) pulses (assumed to be present in Old Norse) which in main stress syllables split into two variants: one syllabically early (grave) pulse and one syllabically late (acute) pulse The distinction may have been in place soon after the first millenium (Riad 2005, 2006).

Postscript

The present project owes a great deal to Öhman's work on word and sentence intonation (1967). His is the first attempt to model F0 contours quantitatively using step function control commands and smoothing filters. It also proposed physiological interpretations and discussed the model's implications for historical and dialectal developments of Scandinavian tone accents. A truly pioneering effort.

References

Abrahamsson M & Sundberg J (2007): "Subglottal pressure variation in actors'stage speech", in Rees M (ed): *Voice and Gender*, 343-47.

Bruce G (1977): *Swedish word accents in sentence perspective* Travaux de l'institut de linguistique de Lund 12. Lund: CWK Gleerup.

Elert C-C (1970): *Ljud och ord i svenskan*, Stockholm:Almqvist&Wiksell.

Jespersen O (1926): *Lehrbuch der Phonetik*, Teubner:Leipzig.

Ladefoged (1967): *Three Areas of Experimental Phonetics*, London:Oxford University Press.

Ladefoged P & Loeb G (2002): "Preliminary experiments on respiratory activity in speech", http://www.linguistics.ucla.edu/people/ladefoge/LadefogedAndLoebRespiration.pdf

Ladefoged P (2005): "Speculations on the control of speech", 3-21 in Hardcastle W J & Mackenzie Beck (eds): *A figure of speech*, LEA : New Jersey.

Lehiste I (1970): *Suprsegmentals*, Cambridge:MIT Press.

Riad T (1998). "The origin of Scandinavian tone accents", *Diachronica XV:1*, 63-98.

Riad T (2005). "Historien om tonaccenten", 1-27 in Falk C & Delsing L-O (eds): *Studier i svensk språkhistoria 8*, Lund: Studentlitteratur.

Riad T (2006): "Scandinavian accent typology", *Sprachtypol Univ Forsch (STUF)*, Berlin 59:36-55.

SAOB = *Svenska Akademiens Ordbok*, Lund 1898.

Stetson R H (1951): *Motor Phonetics*, Amsterdam:North-Holland.

Sundberg J, Iwarsson J & Hagegård H (1995): "A singer's expression of emotions in sung performance",. 217–229 in Fujimura O & Hirano M (eds): *Vocal Fold Physiology: Voice Quality Control*, Singular, San Diego.

Öhman S (1967): " Word and sentence intonation: A quantitative model".*STL/QPSR 8*(*2-3):20-54.*

Coarticulation as incomplete interpolation

Björn Lindblom[1], Diana Krull[1] & Harvey M Sussman[2]
[1]Dept of Linguistics, Stockholm University, Stockholm Se 10691 Sweden.
[2]Dept of Linguistics, Austin Texas 78712 USA.

Abstract

The main result of this project is that we successfully replicated Öhman´s (1966) study and were able to describe both his and our own data using 2-D locus equations. Restating this finding we can say that, for V_1CV_2 with C = [b], [d] or [g], F2 transition offsets and onsets could be accurately predicted from information on the F2 values of V1 and V2 and the identity of the consonant. These results are fully compatible with an account that describes speech movements as unfolding phoneme-by-phoneme and with de-activation of articulatory structures not recruited by the current phoneme. According to this view vowel-dependent coarticulation of consonants in V_1CV_2 utterances can be characterized as a process of incomplete interpolation between articulatorily specified goals.

Background and goals

Studies of coarticulation usually take the classic work of the sixties as a point of departure. The present report is no exception.

Öhman (1966) investigated formant transitions for [b], [d] and [g] in V_1CV_2 sequences. He found patterns incompatible with a context-independent ***'locus'*** value for each place of articulation as proposed by Delattre et al (1955). Öhman's work shaped how phoneticians subsequently came to look at the relationship between phonetic segments and their acoustic correlates. Liberman & Mattingly (1985) captured that view in an often quoted remark:

"*…… there is simply no way to define a phonetic category in purely acoustic terms*".

Later Sussman and colleagues redefined 'locus' as the observable onset of a F2 transition (Sussman et al 1991). Analyzing CV syllables they used 'locus equations' (LE:s) - F2 transition onset plotted against F2 midpoint of the following vowel - to demonstrate robust linear relationships.

$$F2_{onset}(V2)=a+b*F2_{mid}(V2) \qquad (1)$$

In a number of publications, Sussman and colleagues have shown that the slopes and the intercepts of the LE metric provide distinct representations of place of articulation in stops and, as suggested by Krull (1987), can serve as useful indices of degree of coarticulation (Sussman (in prep) for a summary).

In recent work (Agwuele et al 2008, Lindblom et al 2007) the metric has been modified to handle the effect of both preceding and following context on F2 onsets:

$$F2_{onset}(V2)=a+b*F2_{mid}(V2)+c*F2_{mid}(V1) \qquad (2)$$

The formula says that, in V_1CV_2 sequences, the $F2_{onset}$ of the second vowel is a linear weighted combination of the $F2_{mid}$ values of V1 and V2. As in Eq (1) *'a'* and *'b'* are coefficients representing intercept and slope; *'c'* is also a slope measure.

One goal of the present project is to evaluate these formulas with the aid of Öhman's (1966) measurements and some new data of our own.
A second aim is to show how the new results throw new light on the process of coarticulation.

Öhman's data revisited

The (1966) speech samples consisted of all combinations of /b/ /d/ and /g/ with /yː/ /øː/ /ɑː/ /oː/ and /uː/ in initial and final position. Each word occurred five times. Accordingly there were 125 (5*5*5) items for each place. The subject was a male Swedish speaker who was instructed to pronounce the speech samples on a monotone pitch and with equal stress on the initial and final vowels.

Öhman used data of the kind shown in Figure 1 as counter-evidence against the 'locus'. Originally this notion was defined as an invariant underlying frequency value representing the starting point of all CV transitions in practice found by extrapolating the transitions backwards in time for different vowels to a common point of intersection. Figure 1 illustrates a case for which such a common point cannot be identified.

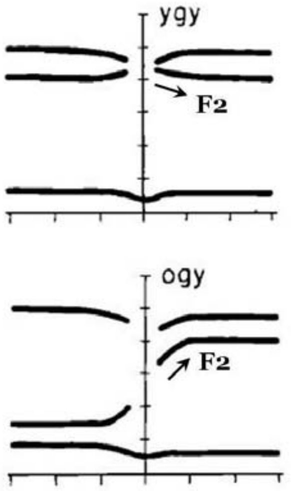


Figure 1. Stylized formant patterns for two test words (from Öhman's Fig 9). In /ygy/ F2$_{onset}$ is higher than the following F2$_{mid}$ and the transition is falling. In /ogy/ F2$_{onset}$ is lower than F2$_{mid}$ and the F2 transition is rising.

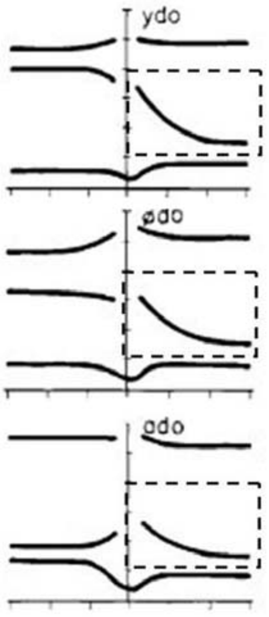
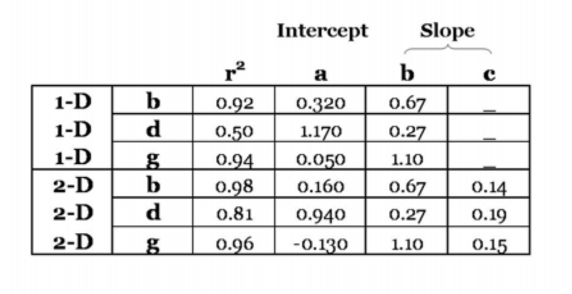


Figure 2. Stylized spectrograms of /y:do:/ /ø:do:/ and /ɑ:do:/. The dashed rectangles draw attention to the effect of V$_1$ on the onset of F2 in /do:/.

Looking at the stylized formant patterns (Öhman's Figs 7-9), one detects certain regularities in the data. Although, in Figure 2, the second syllable is kept the same, the F2 onsets are seen to depend on F2 in the first vowel. It was observations of this type that suggested that the 1966 data might be described in a simpler quantitative manner by taking both V$_1$ and V$_2$ into account.

To explore this possibility question we used the numbers of Tables II and IV (Öhman 1966) and made F2onset(V2)-versus-F2mid(V2) diagrams. We fitted both Eq (1) and Eq(2) to the data using a multiple regression method. Our aim was to establish standard LE slopes and intercepts using Eq (1) and to examine a possible effect of V1 on the F2 onsets of V2 with the aid of Eq (2). The result of the analyses is presented in Table I.



			Intercept	Slope	
		r^2	a	b	c
1-D	b	0.92	0.320	0.67	_
1-D	d	0.50	1.170	0.27	_
1-D	g	0.94	0.050	1.10	_
2-D	b	0.98	0.160	0.67	0.14
2-D	d	0.81	0.940	0.27	0.19
2-D	g	0.96	-0.130	1.10	0.15

Table 1. Top half: Results of fitting Eq(1) to the data in Öhman (1966). Lower part: Fitting Eq(2) to the same data. The intercept "a" is specified in kHz; "b" quantifies the effect of V; "c" reflects the V$_1$ dependence. All slopes and intercepts reached at least the 1% significance level.

Both equations provide adequate descriptions but it is clear that the 2-D model significantly improves the fit to the data. Comparing *c* slopes indicates that as expected V$_1$ plays a significant role in determining the F2onset(V2) values.

An indication of the numerical accuracy of the 2-D predictions is obtained by looking at the average absolute differences between of the model and the measurements. In Hz they were found to be 40 ([b]), 65 ([d]) and 60 Hz ([g]) which amounts to 4-5%.

A replication

To examine the generality of these findings we recorded 4 male speakers of Swedish producing the (1966) test words in accordance with Öhman's experimental procedures. The speech samples consisted of all combinations of /b/ /d/ and /g/ with /y:/ /ø:/ /ɑ:/ /o:/ and /u:/ in initial and final position. Each word occurred five times. The test items were pronounced on a monotone pitch and with equal stress on the initial and final vowels. Formant frequency measurements were made using Swell software (Ternström 2000).

The top panel of Figure 3 shows the [b]-words, the middle the [d]-words and the bottom the [g]-words. The data form quasi-linear patterns but the points cover a rather broader range along the ordinate than what is normally seen. Our statistical analyses indicate that this is mainly due to the effect of V_1.

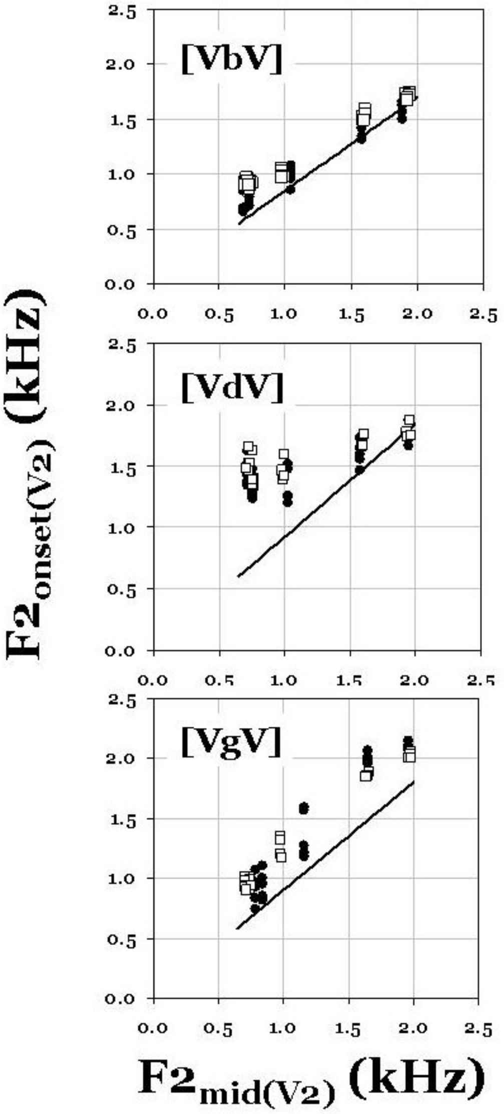
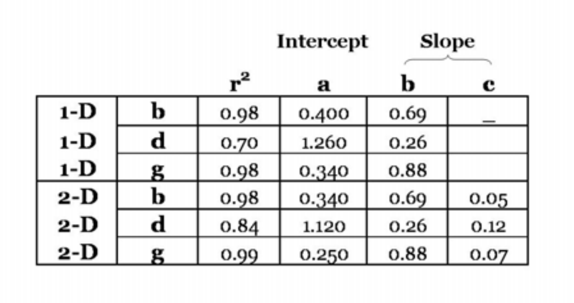


Figure 3. The data from Öhman (solid circles) compared with results from new measurements (unfilled squares). The straight lines are simulated ‘locus lines’ to be discussed later in the interpretation of the results.

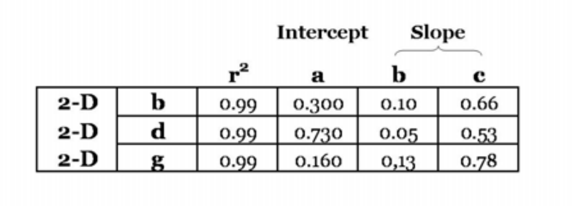
The statistical results for the new data are presented in Table 2. As we examine Figure 3 and compare Tables 1 and 2, we observe strong parallels. Slopes, intercepts and r^2 scores, pattern in a similar fashion. As expected, V2 has a stronger effect than V1 on the onsets of the F2 transitions.

Predictive accuracy - again as measured in terms of the average absolute difference between observed and calculated values comes out to be - 35 ([b]), 50 ([d]) and 45 Hz ([g]) which corresponds to an average error of 3%.



			Intercept	Slope	
		r^2	a	b	c
1-D	b	0.98	0.400	0.69	_
1-D	d	0.70	1.260	0.26	
1-D	g	0.98	0.340	0.88	
2-D	b	0.98	0.340	0.69	0.05
2-D	d	0.84	1.120	0.26	0.12
2-D	g	0.99	0.250	0.88	0.07

Table 2. Results for speech samples analyzed to replicate Öhman (1966). The organization of the table is identical with that of Table 1. “b slopes” are linked to V_2 and “c slopes” to V_1.



			Intercept	Slope	
		r^2	a	b	c
2-D	b	0.99	0.300	0.10	0.66
2-D	d	0.99	0.730	0.05	0.53
2-D	g	0.99	0.160	0,13	0.78

Table 3. 2-D regression analyses for $F2_{offset(V1)}$. The numbers should be compared with those in the lower part of Table 2.

2-D analyses were also made for the F2 transitions in V1 (Table 3). Comparing the results with the correspoinding numbers in Table 2 we note that they indicate strong linearity for all three places. As can be expected the dependence on V2 is reduced whereas the effect of V1 is strengthened.

Discussion

The 2-D model summarizes the two sets of observations with a fair degree of accuracy. The findings imply that, if we know the place of the C and the F2 values of V1 and V2, we can recreate the formant transitions in a meaningful quantitative manner. Conceptually the predictions would seem to represent a process of ‘assimilation’ in the sense that transition offsets and onsets are displaced in the direction of the surrounding vowels. Degree of context-dependence is captured by the slopes. The numbers indicate stronger anticipatory than regressive effects. V2 has a larger weight than V1 in modifying F2onset(V2). Conversely V1 dominates F2offset(V1).

Inferring articulatory mechanisms

What do these results tell us about the underlying articulatory mechanisms? How do the patterns arise? For bilabial closure the tongue body is free to anticipate the following V. For [d] the tongue blade/tip is the primary articulator. The tongue body can be shaped by surrounding vowels although its mobility is somewhat limited by the anterior closure. Coarticulation in [g] differs from that in [b] and [d] in that the same articulator for C and V is used: the tongue body. [b] uses lips and tongue separately. [d] uses blade and body of tongue in a semi-independent manner.

Accordingly the three places present three different pictures of coarticulatory coordination. Thus each place will be considered separately.

APEX simulations

As a baseline for our interpretations we simulated LE's for [b] [d] and [g] using our articulatory model APEX. We used it to generate formant patterns and area functions for a representative sample of vowel articulations (top left panel Figure 4). As a method of simulating stop closures "coarticulated" with the 16 vowels selected we proceeded as follows. The jaw specifications of the vowels were changed to a value typical of the stops (=7 mm). New area functions were derived for the modfied vowels. Three additional series of area functions were then produced by introduing narrow constrictions (~ "stop closure") at three points. For [b] the final section (lips) was replaced; For [d] the constriction was placed just behind the teeth. For [g] the 'closure' occurred 3.5 cm posterior to the teeth. APEX calculated the F patterns for these 48 (3*16) peudo-stops and their F2's were plotted against the F2 of the corresponding vowels. The results are presented in Figure 4.

For all places data points form linear rather tight clusters. The three places produce similar slopes that all come close to the 45 degree line.

Although the simulations of "place" are physiologically crude they enable us to make comparisons with the real data and help us understand why observed LE's come out the way they do. For instance they tell us that introducing a local perturbation (narrow constriction 0.5 cm long) somewhere along the VT does not change the formants appreciably. Hence "F2onsets" plotted against "F2vowel" should approximate straight lines with slope=1.0. In other words, LE 'linearity' is to some extent a consequence of the articulation-to-acoustics mapping.

Coarticulation in [VbV] sequences

The least unrealistic APEX simulations are those for [b]. It is interesting that observed LE's for symmetrical [VbV] often show slopes significantly smaller than 1.0.

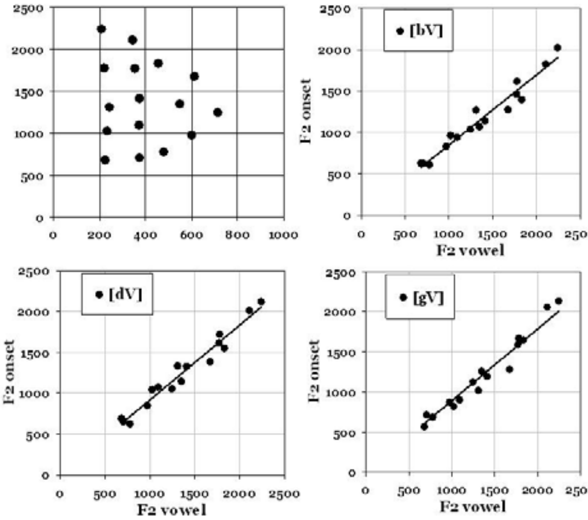


Figure 4 Top row: Vowel formant patterns (left); Simulated LE for "[bV]" (right); Bottom panels: Simulated LE:s for "[dV]" (left) and "[gV]" (right).

One hypothesis is that speech is produced in a phoneme-by-phoneme fashion (Joos 1948) and that the vowel activity during the [VbV] is *de-activated* during the [b] closure. Articulatory tongue traces of tongue height in [ibi] tend to show a trough effect (Lindblom et al 2002) which may be stronger in aspirated stops (Engstrand et al 1997). The latter observation suggest the possibility that the the trough shape of the tongue is also related to aerodynamic factors (Hoole et al 1998).

Producing [d]

The de-activation hypothesis can also be applied to the production of [d] as suggested by the following investigation.

Lindblom (2003) reports data from a 20-second X-ray film of a Swedish male speaker It contains information on tongue shapes and a Principal Components Analysis. The speech samples of relevance here are six test words: [ɛ:'di:], [ɛ:'de:], [ɛ:'da(s:)], [ɛ:'dɑ:(l)], [ɛ:'dɔ(lk)], [ɛ:'du:(s)]. The tracings for these words form a subset of a total of 400 frames analyzed. The images taken at 50 frames/sec portray a midsagittal articulatory profile. Tracings of acoustically relevant structures were made. Contours were converted into tables with x- and y-coordinates, calibrated in mm and corrected for head movements. The tongue contours were redefined in a jaw-based coordinate system and resampled at 25 equidistant ‘fleshpoints’ in preparation for a Principle Components Analysis (PCA).

The input to the PCA was a matrix whose rows specifed the indivual images and whose columns contained the x and y coordinates. The output consisted of a set of basic tongue shapes (the PC’s) and weights to be applied to the various PC’s to recreate observed contours.

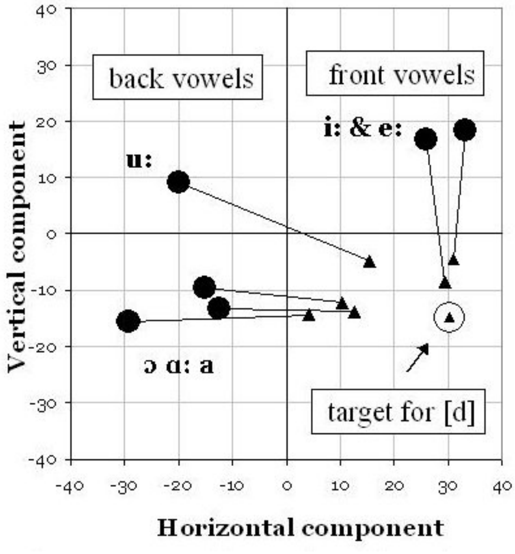


Figure 5. PCA representation of dV transitions. The vertical component of PC1 along the ordinate, and its horizontal component along the x-axis.

Using the weights for a given tongue shape we obtain the location of that shape in the articulatory space defined by the PCA. Figure 5 is such a diagram for the most important component (PC1). The vertical position (ordinate) is plotted against the horizontal position (abscissa) for tongue contours observed at the release of [d] (triangles) and vowel targets (solid circles). The straight lines stylize the [dV] transitions.

The lines show a systematic fanlike pattern. Interestingly it is possible to derive a mean point of intersection (encircled triangle). By definition it is context-free (cf ‘locus’). It has been labeled ‘[d] target’ but the reality of this point is unlikely to be a shape that the speaker “aims at”. More plausibly it should be interpreted as a shape for *complete de-activation* of vowel context. This account of [d] presents a neat parallel with evidence for de-activation in [b]. Since the de-activated tongue in [b] and [d] has a mid-range F2, the de-activation hypothesis has the further advantage of explaining where the LE slopes for [b] and [d] come from. The reason why they are < 1.0 ([d] more so than [b]) is that, in the limit (complete de-activation). they become horizontal lines.

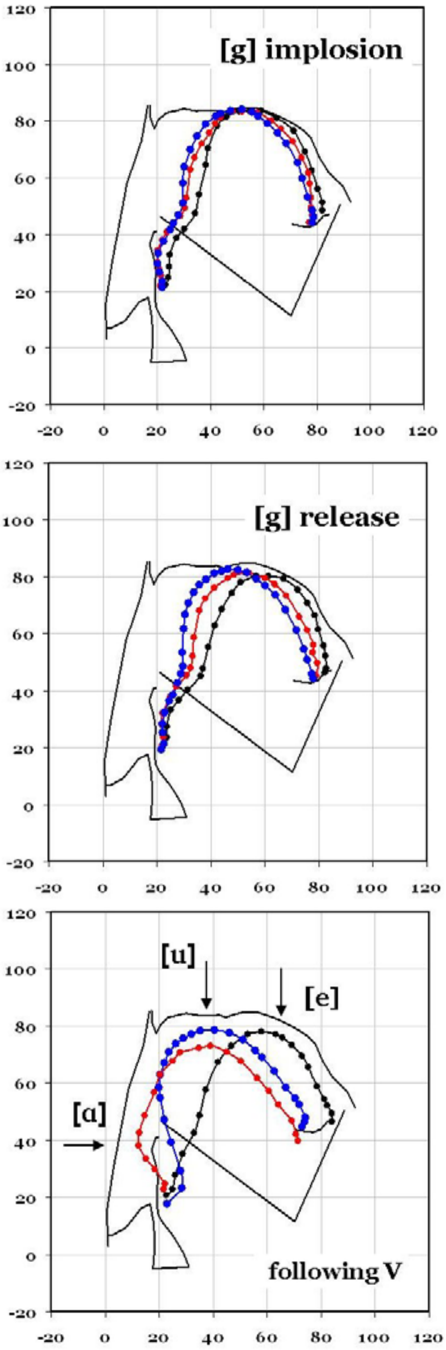


Figure 6. Tracings of articulatory profiles during [ɛ :'ge:] (black contour), [ɛ :'gɑ :(l)] (red), and [ɛ :'gu:(s)] (blue).

Articulation and acoustics of [g]

Since the same articulator (tongue body) is used both for [g] and the adjacent vowels the de-activation idea cannot explain coarticulation in [VgV]. What we need to account for is why data points for [g] with back vowels form a cluster whose slope often *exceeds* 1.0 (Figure 3). To think about that problem it is helpful to look at a few articulatory profiles. Figure 6 presents tracings from [ɛː'geː], [ɛː'gɑː(l)] and [ɛː'guː(s)].

The first diagram of the vertical series in Figure 6 compares the tongue contours at the moment of [g] implosion. Note how minimal the front-back differences are. The next picture shows the situation at the release. There is greater separation between the contours presumably in anticipation of the upcoming vowels. Nonetheless the three occlusions form a pretty tight group.

The third set presents profiles at the vowel targets. The significance of the arrows is that they indicate where these vowels have their maximal constrictions. If, using place of constriction as our criterion, we were to arrange standard vowels along a front-back continumm we would obtain the following series: [i] [u] [o] [ɔ] [ɑ] [a]. As shown by Figure 6 the profiles for [g] closure resemble [i] and [u] rather closely but their articulatory 'distance' to the other vowel increases as we move along the series. Interestingly so do the extents of F2 transitions: [gu] has a small transition whereas [ga] has a large one.

We can conclude with a refinement of a popular rule. Users of LE have often used LE slope as an index of degree of coarticulation (Krull 1987). For [b] and [d] this rule says that an *increase in slope* implies greater coarticulation (greater resemblance between F2onset and F2vowel). In the case of [g] data with slopes > 1.0 the rule must be changed into "a *decrease in slope* implies greater coarticulation (greater resemblance between F2onset and F2vowel).

References

Agwuele A, Sussman H M & Lindblom B (2008): "The effect of speaking rate on consonant vowel coarticulation", *Phonetica 65*:194–209

Delattre P C, Liberman A M & Cooper F S (1955): "Acoustic loci and transitional cues for consonants", *J Acoust Soc Am 27*:769-773.

Engstrand O & Lindblom B (1997): "The locus line: does aspiration affect its steepness?", Fonetik 97, *Phonum 4*:101-104, Umeå Universitet.

Hoole P, Munhall K & Mooshammer C (1998): "Do air-stream mechanisms influence tongue movement paths?", *Phonetica 55*:131-146.

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

Krull D (1987): "Second formant locus patterns as a measure of consonant-vowel coarticulation", *PERILUS V, Stockholm University*, 43-61.

Liberman A & Mattingly I (1985): ''The motor theory of speech perception revised'', *Cognition 21*:1–36.

Lindblom B (2003): "A numerical model of coarticulation based on a Principal Components analysis of tongue shapes", *XVth International Congress of Phonetic Sciences*, Barcelona, Spain.

Lindblom B, Agwuele A, Sussman H M, Cortes E E (2007): The effect of emphatic stress on consonant vowel coarticulation", *J Acoust Soc Am 121 (6)*:3802–3813.

Sussman H M, McCaffrey H A & Matthews S A (1991): ''An investigation of locus equations as a source of relational invariance for stop place categorization,'' *J Acoust Soc Am 90*:1309–1325.

Sussman H M S (in press): "A Locus Equation perspective on coarticulation", in Embarki M. & Marchal A (eds): *Coarticulation: between physics and metaphysics*, Cambridge (UK): Cambridge University Press.

Ternström, S. (2000). Soundswell-signal workstation. Manual version 4.0 Hitech Development AB.

Öhman S (1966): "Coarticulation in VCV utterances: Spectrographic measurements", *J Acoust Soc Am 39*:151–168.

Voice Similarity - a Comparison Between Judgements by Human Listeners and Automatic Voice Comparison

Jonas Lindh and Anders Eriksson
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg

Abstract

Comparison between the way human listeners judge voice similarity and how state-of-the art GMM-UBM systems for voice recognition compare voices is a little explored area of research. In this study groups of informants judged the similarity between voice samples taken from a set of fairly similar male voices that had previously been used in a voice line-up experiment. The result from the listening tests was then compared to the scores from a UBM-GMM automatic voice comparison system, built on the Mistral LIA_RAL open source platform. The results show a correlation between scores obtained from the automatic system and the judgements by the listeners. Listeners are, however, more sensitive to language dependent parameters or idiosyncratic phonetic features such as speaking tempo, while the system only bases its likelihood ratios on spectral similarities, i.e. timbre.

Introduction

Automatic methods, often referred to as Automatic Speaker Recognition systems, are increasingly being used in forensic phonetic casework, but most often in combination with aural/acoustic methods. It is therefore important to get a better understanding of how the two systems compare. However, a text independent system does (in most cases) not use information on how anything is spoken, so we will here refer to such a system as an Automatic Voice Comparison (AVC) system. Most importantly, we must know if and under what circumstances the outcome of the two types of analyses (automatic vs acoustic/auditory) may come into conflict and how to resolve the problem if they do. The present study is an attempt to shed some light on how human auditory voice similarity judgements compare with voice similarity scores obtained by automatic systems. We have only found a few studies where human perceptual evaluation of voice similarity and automatic methods have been directly compared. In her master's thesis, Kahn (2008) approached the problem in a manner very similar to the method applied in our study. Perception data was acquired by having subjects judge voice similarity in a pairwise comparison test using a five point scale. Likelihood ratio scores were obtained by a GMM-UBM system built on the Mistral/Alizé toolkit (http://mistral.univ-avignon.fr/). Speakers were related to some extent and read speech was recorded. Some classic acoustic measurements like mean F0 and formants were also done. The results show no significant correlation between system scores and perceptual evaluations or acoustic parameters. However there are correlations between listeners' judgments and acoustic speech parameters. The conclusion is that listeners base their judgements on acoustic information of the type represented by factors like F0 and formant values but also on speech style. Zetterholm et al. (2004) used an imitator as impostor to test a text dependant speaker verification system (Melin et al., 1998). The imitator was first trained by testing the system and receiving feedback both through listening to the target voice and being informed about the similarity (log likelihood ratio) scores. In an ABX perception test, 22 listeners rated voice similarity for all voices used in the system. The results of the listening test were then compared to the scores obtained by the system. Other studies have been conducted comparing aural similarity judgements with acoustic differences. Cleary et al. (2005) performed a study on voice similarity assessments by children with and without cochlear implants using a discrimination test. F0 needed to differ by at least 2–2.5 semitones for normal-hearing children to perceive the voices as belonging to different talkers. Several others have also studied speaker recognition correlation between aural judgements and classic acoustic parameters such as F0 and formants (Brown, 1981; Murry and Singh, 1980) or solely aural degree of similarity as a factor to be used to choose voices for a line-up (Rietveld et al., 1991).

The speech material used in the present study was originally produced for an ear witness study where 7 speaker line-ups were used to test voice recognition reliability in ear witnesses. The speakers in that study were matched for general speaker characteristics like sex, age and dialect. Some of the results obtained in the earwitness study served as inspiration for the present study. It was found, for example, that the occurrence of false identifications was not randomly distributed but systematically biased towards certain speakers. Such results raise obvious questions like: Why were these particular speakers chosen? Are their speaker characteristics particularly similar to those of the intended target? Would an aural voice comparison test single out the same speakers? And how would these voices be ranked by an automatic recognition system?

In the present study we have approached these questions by combining two experiments. In one of the experiments, listeners are asked to judge voice similarity in a pairwise comparison test. In another experiment the same stimuli are analyzed using a state-of-the-art GMM-UBM system. And in the final analysis we compare the outcomes of the two experiments and see to what extend they produce similar predictions and compare these predictions with some of the results obtained in the ear witness experiment mentioned above.

Method

To correlate the two different kinds of measures, perceptual judgements on a five point scale and raw (not normalised) likelihood ratios, we first needed to choose compatible scales to represent both types of data. In the present case we decided that using ordinal scales (in this case rank order) for both results would suffice. There are possibilities to calculate a distance between models in the automatic system, for example cross likelihood ratio, suggested by Reynolds (1995) or normalised cross likelihood ratios (Le et al., 2007). But we choose instead to consider the informants' judgements as rankings of most to least alike since we cannot be sure that the subjects have judged the similarities on a more precise scale than a rank ordering.

To be able to collect sufficiently large amounts of data, two different web tests were designed. One of the web based forms was only released to people that could insure a controlled environment in which the test was to take place. Such a controlled environment could for example be a student lab or equivalent. A second form was created and published to as many people as possible throughout the web, a so-called uncontrolled test group. The two groups' results were treated separately and later correlated to see whether the data turned out to be similar enough for the results to be pooled.

Mistral LIA_RAL - an open source toolkit for building a voice comparison system

The NIST speaker recognition evaluation campaign started already 1996 with the purpose of driving the technology of text-independent voice recognition forward as well as test the performance of the state-of-the-art approach and to discover the most promising algorithms and new technological advances (from http://www.nist.gov/speech/tests/sre/ Jan 12, 2009). The aim is to have an evaluation at least every second year and some tools are provided to facilitate the presentation of the results and handling the data (Martin and Przybocki, 1999).

A few labs have been evaluating their developments since the very start with increasing performances over the years. These labs generally have always performed best in the evaluation. However, an evaluation is a rather tedious task for a single lab and the question about some kind of coordination came up. This coordination could be just to share information, system scores or other to be able to improve the results. On the other hand, the more natural choice to be able to share and interpret results is open source. On the basis of this Mistral and more specifically the Alizé SpkDet packages were developed and released as open source software under a so-called LGPL licence (Bonastre et al., 2005; Bonastre et al., 2008).

The very foundation of Mistral is Alizé, which is the umbrella for all developed packages you might include in your own application or framework. The feature extraction is handled by SPro, an open source signal processing toolkit (Guillaume, 2004). Using a background model trained using Maximum Likelihood Criterion and individual models trained to maximise the a posteriori probability that the claimed identity is the true identity given the data (MAP training) is called the GMM-UBM approach (Reynolds et al., 2000).

Description of the AVC system used for this study

For the set up used in this study the so-called state-of-the-art GMM-UBM approach was adopted. Frame selection was made based on simple energy detection and the removal of silences longer than 100 milliseconds from each recording. 19 MFCCs were extracted together with delta and acceleration coefficients. 512 Gaussian mixture models were applied. The UBM was trained on 2 minutes of spontaneous speech (after frame selection) from 628 male speakers in the Swedia dialect database (Eriksson, 2004). The state-of-the-art performance of this kind of system for band limited (phone speech) is given in Fauve and Matrouf (2007). The recordings used here were sampled at 16 kHz/16 bits. The test recordings were between 13–15 seconds in duration.

The web based listening tests

The listening tests had to be made interactive and with the results for the geographically dispersed listeners gathered in an automatic manner. Google docs provide a form to create web based question sheets collecting answers in a spreadsheet as you submit them and that was the form of data collection we chose to use for the perception part of the study. However, if one cannot provide a controlled environment, the results cannot be trusted completely. As an answer to this problem two equal web based listening tests were created, one intended for a guaranteed controlled environment and one openly published test, here referred to as uncontrolled. The two test groups are here treated separately and correlated before being merged in a final analysis.

The listening test material

In the ear witness project mentioned above, the aim is to gain a better understanding of earwitness reliability. One study was designed in which children aged 7–9 and 11–13 and adults served as informants. A total of 240 participants, equally distributed between the three age groups, were exposed to an unfamiliar voice (the planning of a crime, PoC). After two weeks, the witnesses were asked to identify the target-voice in a line-up (7 voices). Half of the witnesses were exposed to a target-present line-up (TP), and the other half to a target-absent line-up (TA). The recordings used for the line-ups consisted of spontaneous speech elicited by asking the speakers to describe a walk through the centre of Gothenburg based on a series of photos presented to them. The 9 (7 plus 1 in TA + target) speakers were all selected as a very homogeneous group, with the same dialectal background (Gothenburg area), age group (between 28–35). The speakers were also selected from a larger set of 24 speakers on the basis of a speaker similarity perception test using two groups of undergraduate students as subjects. The subjects had to make similarity judgments in a pairwise comparison test where the first item was always the target speaker intended for the line-up test. Subjects were also asked to estimate the age of the speakers. The recordings used for these tests were 16 kHz /16 bit wave files.

In the perception test for the present study, 9 voices were presented pair-wise on a web page and listeners were asked to judge the similarity on a scale from 1 to 5, where 1 was said to represent “Extremely similar or same” and 5 “Not very similar”. Since we wanted to minimize the influence of any particular language or speaking style influence the speech samples were played backwards. The listeners were also asked to submit information about their age, first language and dialectal background (if Swedish was their first language). There was also a space where they could leave comments after the completion of test and some participants used this opportunity. The speech samples used in the perception test were the first half of the 25 second samples used in the earwitness line-ups, except for the pairs where both samples were from the same speaker. In these cases the other item was the second half of the 25 second samples. Each test consisted of 45 comparisons and took approximately 25 minutes to complete. 32 (7 male, 25 female) listeners performed the controlled listening test and 20 (6 male, 14 female) the uncontrolled test.

Results

The results are first outlined separately and then compared in the final subsection.

System scores

Comparing all voices was done by training models for each voice as a target before testing. Models were also tested against themselves. The scores are presented as raw (not normalised) likelihood ratios.

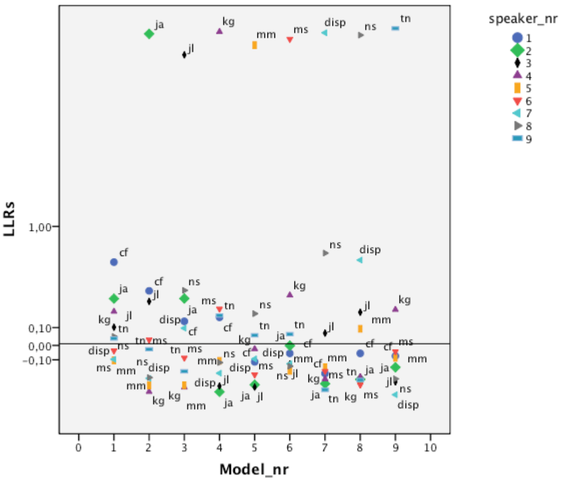


Figure 1. Distribution of likelihood ratio scores on a logarithmic likelihood ratio scale (y-axis) and model number (x-axis).(The planning of a crime (PoC)=disp).

For all subsequent comparisons we have converted the raw (not normalised) likelihood ratios to ranks. For easy comparison of the totality of the results we present the data as a rank matrix

Table 1. The table shows speaker ranks when tested against models of each speaker including themselves. Mean values and standard deviation for each speaker's rankings are also shown.

Model	JA	JL	KG	MM	MS	PoC	NS	TN	CF
JA	1	3	9	8	4	7	6	5	2
JL	3	1	9	8	6	5	2	7	4
KG	9	8	1	5	2	7	6	3	4
MM	8	9	4	1	7	5	2	3	6
MS	4	7	2	9	1	6	8	3	5
PoC	8	3	7	4	5	1	2	9	6
NS	7	3	6	4	9	2	1	8	5
TN	6	8	2	5	3	9	7	1	4
CF	2	4	3	9	7	8	5	6	1
Mean rank	5.33	5.10	4.77	5.88	4.88	5.55	4.33	5.00	4.10
Std dev	1.7	2.9	2.9	3.1	2.8	2.6	2.7	2.6	2.7

The results show that some speakers are generally higher ranked than others. For example speaker CF has a mean rank of 4.1, indicating that he is somewhat of a wolf. From the system scores (figure 1) it is also clear that the CF model suffers from a high degree of false acceptance, which indicates that the model is a lamb (Campbell, 1997; Doddington, 1985; Melin, 2006).

Listening test result

Both listening tests separately (controlled and uncontrolled) show significant inter-rater agreement (Cronbach's alpha = 0.98 for the controlled and 0.959 for the uncontrolled test). When both datasets are pooled the inter-rater agreement remains at the same high level (alpha = 0.975) indicating that listeners in both subgroups have judged the voices the same way. This justifies using the pooled data from both groups (52 subjects altogether) for the further analysis of the perception test results. The results of the perception test are presented in Table 2. The rankings are based on the means of the similarity judgments.

Table 2. The table shows speaker ranks based on mean similarity judgement for both listener groups pooled.

Speaker	JA	JL	KG	MM	MS	PoC	NS	TN	CF
JA	1	4	5	3	6	8	9	7	2
JL	3	1	8	5	7	4	2	9	6
KG	5	9	1	2	3	7	8	6	4
MM	4	5	2	1	3	8	9	7	6
MS	7	8	6	5	2	9	3	1	4
PoC	5	3	6	4	9	1	7	8	2
NS	6	2	8	5	3	7	1	9	4
TN	6	9	5	4	1	7	8	2	3
CF	2	9	6	7	3	5	8	4	1
Mean rank	4.3	5.6	5.2	4.0	4.1	6.2	6.1	5.9	3.6
Std dev	2.0	3.2	2.4	1.8	2.6	2.5	3.2	2.9	1.7

Also in the perception test, speaker CF receives the highest mean rank with low variation. This indicates that speaker CF is also the most likely to be picked as the target if uncertain in a closed set line-up. This was indeed also the case in the earwitness study where speaker CF was the speaker most often confused with the target speaker resulting in a large number of false acceptances in both the TA and the TP conditions.

Comparison between system scores and listening tests

In order to visualize the results presented in the matrices above we used Multidimensional Scaling to produce 2-dimensional Euclidean distance models (similar to Kahn, 2008).

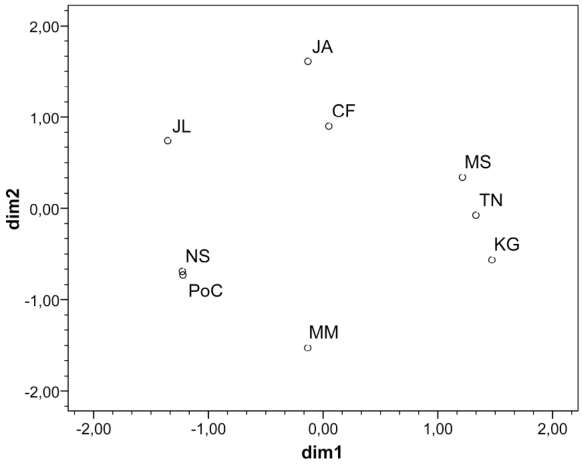


Figure 2. Euclidean distance model representing the data rankings based on the log likelihood ratios from the ASC system.

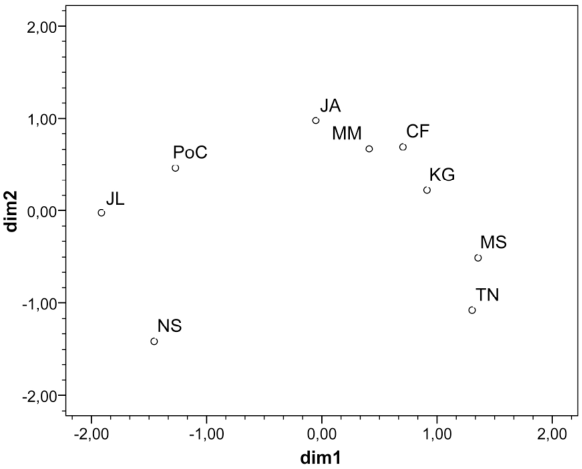


Figure 3. Euclidean distance model representing the data rankings based on the perception results.

We may observe several similarities between the representations of the AVC and the perception scores. Two groups of speakers cluster together in both diagrams. Speakers JL and NS group together with target voice in the earwitness study (PoC). Speakers MS, TN and KG who form a group in the AVC analysis also group together in the perception results. The latter speakers are also among those who are least often subject to false acceptance in the line-up experiment.

Discussion and Conclusions

Even though it is difficult to draw any precise conclusions on whether humans and the system perform the same kind of comparisons, some conclusions may be drawn regarding the influence of linguistic factors. It seems as if humans include what is being said and how it is said in voice similarity judgements. This probably made the similarity judgment a rather difficult task using backward speech (also reflected in their comments). The system score ranking gives a clue to why the voice CF can be considered a 'wolf'. It is the voice with highest mean rank in both Table 1 and 2.

We also mean that many differences can be explained by the use of linguistic/phonetic cues still present in backward speech. Not all factors are eliminated this way, for example pausing and articulation rate. A separate analysis using rankings based on articulation rate shows MM, CF and PoC clustering together. This may explain why MM and CF cluster in Figure 2, in spite of the fact that they are quite dissimilar in the AVC analysis. These 2 voices are also subject to a high degree of the false acceptances made by participants in the voice line-up study. The false acceptances are most biased towards CF, which we suspect is a combination of speech feature similarities and his ‘wolfness’.

Speakers JL, NS and PoC also form a group in both analyses, but here we may see that whereas NS and PoC are regarded as identical in the ASC analysis they are quite widely separated in the perception data. This difference contains important information about the influence of speaking style for the perception results. Speaker NS is namely the speaker who was also used for the mock incriminating call. The voice is thus the same in both samples which is detected by the AVC system. The speaking style is, however, quite radically different. And as has been pointed out above such factors as articulation rate and pausing are at least partly present even if the speech samples are played backwards. It seems reasonable to suggest that the listeners observe this difference and therefore judge the speech samples as quite different in spite of the fact that the voice characteristics are very similar.

Generally we can conclude that identifying a speaker in the voice line-up study was also a difficult task. We suspect that listeners use at least two different strategies. They may pay a lot of attention to voice quality or concentrate on speaking parameters such as articulation rate and maybe to some extent pronunciation. In the first case they are much more likely to make correct identifications in a line-up task or judge the voices more in agreement with the automatic system in a voice similarity task.

One of the aims of the present study was, as pointed out in the introduction, to look for similarities between automatic and perceptual analyses, but also to detect possible conflicting differences. The present study does not contain any conflicting results, but several examples of how human listeners integrate factors which depend on speaking style even when the task is explicitly to judge voice similarity.

References

Bonastre, J-F, Wils, F. & Meigner, S. (2005) ALIZE, a free toolkit for speaker recognition, in Proceedings of ICASSP, 2005, pp. 737–740.

Bonastre, J-F, Scheffer, N., Matrouf, C., Fredouille, A., Larcher, A., Preti, A., Pouchoulin, G., Evans, B., Fauve, B. & Mason, J.S. (2008) ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In Odyssey 2008 - The Speaker and Language Recognition Workshop, 2008.

Brown, B. (1981). An experimental study of the relative importance of acoustic parameters for auditory speaker recognition. Language and Speech 24: 295–310.

Campbell, J. (1997) Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9):1437–1462.

Cleary, M., Pisoni, D. & Kirk, K. (2005) Influence of Voice Similarity on Talker Discrimination in Children With Normal Hearing and Children With Cochlear Implants. Journal of Speech, Language, and Hearing Research 48(2005/015): 204–223.

Doddington, G. (1985) Speaker recognition - identifying people by their voices. Proceedings of the IEEE, 73(11):1651–1664.

Eriksson, A. (2004) SweDia 2000: A Swedish dialect database. In Babylonian Confusion Resolved. Proc. Nordic Symposium on the Comparison of Spoken Languages, ed. by P. J. Henrichsen, Copenhagen Working Papers in LSP 1 – 2004, 33–48.

Fauve, B., Matrouf, D. et al. (2007) State-of-the-Art Performance in Text-Independent Speaker Verification through Open-Source Software. IEEE Transactions on Audio, Speech and Language Processing 15, Issue 7: 1960–1968.

Guillaume, G. (2004) SPro: speech signal processing toolkit, Software available at http://gforge.inria.fr/projects/spro.

Kahn, J. (2008). Caractéristique propres au locuteur : Traitement automatique et distance perceptive. Université-Stendhal-Grenoble 3 : Unpublished Masters Thesis.

Le, V-B., Mella, O. & Fohr, D. (2007) "Speaker Diarization using Normalized Cross Likelihood Ratio". Proceedings of Interspeech 2007, Antwerp, Belgium, 2007.

Martin, A. F. and Przybocki, M. A. (1999) The NIST 1999 Speaker Recognition Evaluation-An Overview. Digital Signal Processing 10: 1–18.

Melin H., Koolwaaij J.W., Lindberg J., Bimbot F. (1998). A Comparative Evaluation of Variance Flooring Techniques in HMM-based Speaker Verification. Proc. of ICSLP ’98:1903–996.

Melin, H. (2006) Automatic speaker verification on site and by telephone: methods, applications and assessment, PhD thesis, KTH, Stockholm, December 2006.

Murry, T. and S. Singh (1980). "Multidimensional analysis of male and female voices." Journal of the Acoustical Society of America 68: 1294–1300.

Reynolds, D., (1995) Speaker identification and verification using Gaussian mixture speaker models, Speech Communication, vol. 17, issue 1–2, pp. 91–108, 1995.

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., (2000) Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, 2000.

Rietveld, A. C. M. and A. P. A. Broeders (1991). Testing the fairness of voice parades: the similarity criterion. Proceedings of the 12th International Congress of Phonetic Sciences. Aix-en-Provence, Université de Provence, Service des Publications. 5: 46–49.

Zetterholm, E., D. Elenius and M. Blomberg (2004) A comparison between human perception and a speaker verification system score of a voice imitation. Proceedings SST2004, Sydney, Australia, Dec 8–10 2004.

Cross-modal magnitude matching as a precursor of multi-modal speech perception

Ellen Marklund, Anna Ericsson and Francisco Lacerda
Department of Linguistics, Stockholm University

Abstract

Four- to ten-month-old infants (n=58) were examined on their ability to match magnitude across modalities. Their looking behaviour was recorded as they were presented with an intensity modulated auditory stimulus and three possible visual matches. The mean looking times towards a visual target (size envelope matching intensity envelope of the auditory stimulus) and a non-target were calculated. Five-month-olds and seven- to ten-month-olds show a significant preference looking towards the target, as do an adult control group. Four- and six-month-olds do not.

Background

The current paper will briefly cover multimodal perception of speech as well as the power law of phsycophysics. Cross-modal magnitude matching in infancy will be proposed as a result of the latter and a precursor of the former. Lastly, the experiment of the current study will be presented.

Multimodal speech perception

Speech perception is a multimodal phenomenon. It is not the acoustic signal alone that provides the listener with information; if available, the articulatory movements of the speaker's face influence the perception of the speech signal. In noisy environments, listeners are better able to understand what is being said if able to see the speaker (Erber, 1975; Ewertsen and Nielsen, 1971; Sumby and Pollack, 1954) and when presented with conflicting auditory and visual cues, listeners' perception of speech sounds is heavily influenced by what is seen (McGurk and MacDonald, 1976; Traunmüller and Öhrström, 2007). Similarly, perception of the acoustic speech signal can be enhanced or influenced by information presented in other modalities, such as touch (Bernstein and Benoît, 1996).

Multimodal perception of speech is present already in infancy. If presented with conflicting visual and auditory cues, 5-month-olds' perception of syllables is influenced by the visual component much like in adults (Rosenblum, Schmuckler and Johnson, 1997). In addition, several studies have shown that infants are able to identify the articulatory movements related to syllables or speech passages (Barker and Tomblin, 2004; Kuhl and Meltzoff, 1982; Kuhl and Meltzoff 1984; Kuhl and Meltzoff, 1988; MacKain, Studdert-Kennedy, Spieker and Stern, 1983).

In a classic study, Kuhl and Melzoff (1982) demonstrated 4- to 4.5-month-old infants' ability to identify articulatory movements of vowels. The infants heard either /a/ or /i/ repeatedly, while visually presented with recordings of two faces, each articulating one of the syllables. The infants' looking time towards the correct face was significantly above chance, showing that infants at this age are able to connect the articulatory movements with the acoustic signal.

Attempting to replicate and expand Kuhl and Meltzoff's studies with 6- to 8-month old infants, Lacerda and colleagues visually presented four faces articulating syllables, while one syllable was presented auditorily, thereby reducing chance level to 25% instead of 50%. Syllables used in the study were /a/, /y/, /ba/ and /by/, and the gaze-measure used was gain (how much longer subjects looked to a certain area during the test compared to during baseline). The infants, however, did not alter their looking behaviour significantly towards the visually matching target syllable. Instead they looked more to the articulation of /ba/ regardless of what was presented auditorily (Klintfors, 2008; Lacerda, Klintfors, Gustavsson, Marklund and Sundberg, 2005), a behaviour hypothesized to be related to a general cross-modal matching ability compatible with the phsycophysical power law.

The power law of psychophysics

Psychophysics pertains to the relationship between magnitude of physical stimuli and the perceived magnitude. The psychophysical power law as proposed by Stevens states that “equal stimulus ratios produce equal subjective ratios” (1957). This means that if subjects are asked to adjust the level of a stimulus (e.g. the intensity of a tone or the brightness of a light) to half of the original, they will reduce the stimulus magnitude to a certain percentage of the original, regardless of the level on which they start. The relationship between subjective and stimulus magnitude can thus be described as a power function[1] where M_{subj} is the subjective magnitude, M_{stim} is the stimulus magnitude and k is a constant dependent on which units are used.

$$M_{subj} = kM_{stim}{}^{n}$$

Different sensory impressions have their own characteristic exponent *n*, derived from experimental data (Stevens, 1966). For instance, the exponent of loudness is 0.3 while for heaviness it is 1.45, resulting in different functions which describe how the perceived magnitude corresponds to the stimulus intensisty and stimulus weight respectively (Stevens, 1957). Since all sensory representations are expected to follow the power law, there is a potential direct proportionality between sensory dimensions. Subjects are indeed able to match sensations in one modality to those in another (Stevens, 1966); if, for instance, subjects are asked to adjust the loudness of a tone to the intensity of vibrations applied to fingertips or to the brightness of a light, they will do so systematically according to the relation between the power functions for each of the two modalities (Stevens, 1962).

Cross-modal magnitude matching and the current study

As reported by Stevens (1962), humans have the ability (and possibly a predisposition) to associate degrees of magnitude across modalities. This could explain the results of Lacerda et al. (2005). Infants looked more towards faces with more noticable articulation, suggesting they did not necessarily connect the syllable they heard with any particular articulation, but instead looked towards the most visually prominent event when presented with any sound (contrasted to the silent baseline), resulting in greater gain for the /ba/-articulation than for the others.

In this light, the general process of cross-modal magnitude matching can be seen as a possible precursor of multimodal perception of speech, serving as a fundament of speech development and first language acquisition. Indeed, while there is no reliable evidence for audio-visual speech perception before the age of 3 months (Burnham, 1998), infants’ ability to match intensities across modalities has been demonstrated already at three weeks of age (Lewkowicz and Turkewitz, 1980), suggesting that the cross-modal magnitude matching ability may precede the onset of multimodal perception of speech.

To assess the potential role of general cross-modal magnitude matching in the devopment of multimodal speech perception, it is important to study infants’ spontaneous magnitude matching across modalities in non-speech contexts. The present study thus examines 4- to 10-month-olds ability to match intensity modulated white noise to size modulated images of suns.

Method

Participants were 58 infants between 4 and 10 months of age (mean age = 7.3 months), randomly selected from the Swedish national address register, based on date of birth and geographical criteria. A control group of 12 adults also participated in the study.

The experiment consisted of several short film sequences presented in random order, and had a total duration of 140 seconds. The current paper is based on the analysis of one of the two types of film sequences presented in the experiment. The relevant sequence type had a duration of 20 seconds and showed three suns lined up next to each other (see figure 1). The sizes of the suns were modulated while intensity modulated white noise was presented and the size envelope of one of the suns was congruent with the intensity envelope of the noise (the target). There were three occurrences of the relevant film sequence, balanced for target-sun position.

For each version of the film sequence three areas of interest were defined (see figure 2) and the total looking time for each subject within each of those areas were measured. Finally, each subject’s average looking times to non-target suns and target suns respectively (regardless of screen position) were calculated.

[1] This description is valid the middle ranges; for magnitude values near the minimum and maximum thesholds, there are known departures from this power law.

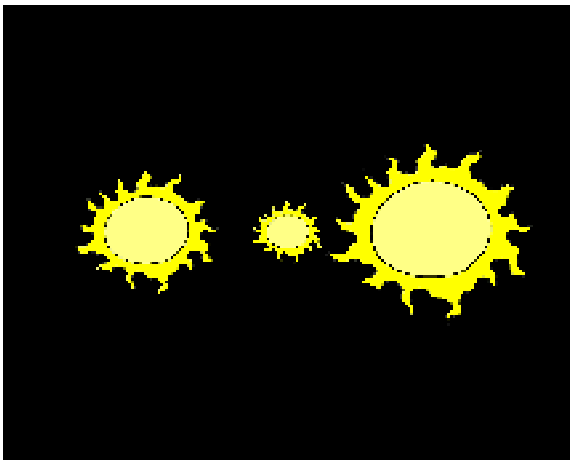


Figure 1. Snapshot of the film sequence. The suns had different size envelopes, one of which matched the noise envelope.

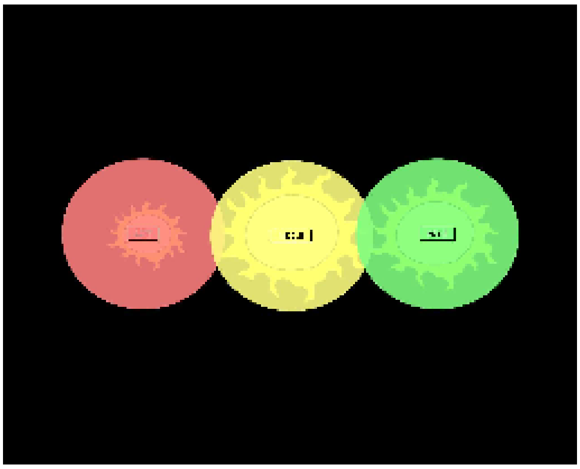


Figure 2. The areas of interest, each covering one sun.

Data recording and pre analysis processing were performed using Tobii T120 and Tobii XL hardware, and Tobii Studio 2.0 software. Data analysis was performed in SPSS 17.0.

Results

The participans were divided into different age groups as shown in table 1.

Table 1. Age groups of the participants.

Age (months)	Number of participants
4	4
5	12
6	5
7	8
8	9
9	9
10	11
Adult (control)	12

The looking behaviour of the infants is shown in figure 3. A significant difference in mean looking time towards target versus non-target was found using a repeated measures ANOVA (F(1,51)=20.185, p<0.0005). There was no interaction between looking behaviour and age group. When analysed separately, the looking preferences of 4-month-olds and 6-month-olds were not significant.

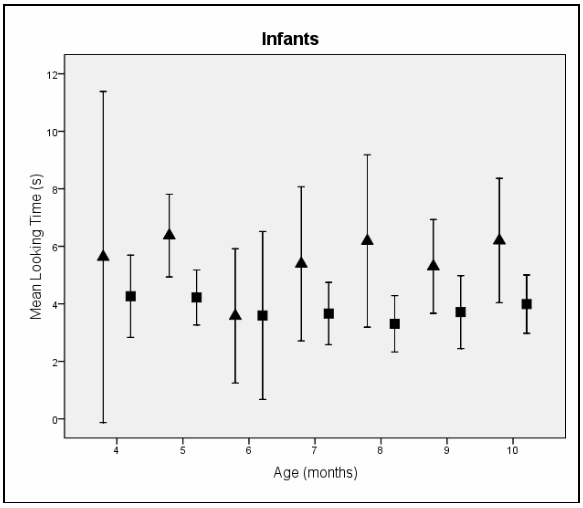


Figure 3. The average looking time (y-axis) towards target (triangles) and non-target (squares) respectively for all age groups (x-axis), with a confidence interval of 95%.

Using a paired samples 2-tailed t-test it was demonstrated that the adults looked significantly longer to the target sun (t(11)=2.487, p<0.03). For target suns, the mean looking time was 8.3 seconds and for non-target suns it was 6.2 seconds (see figure 4).

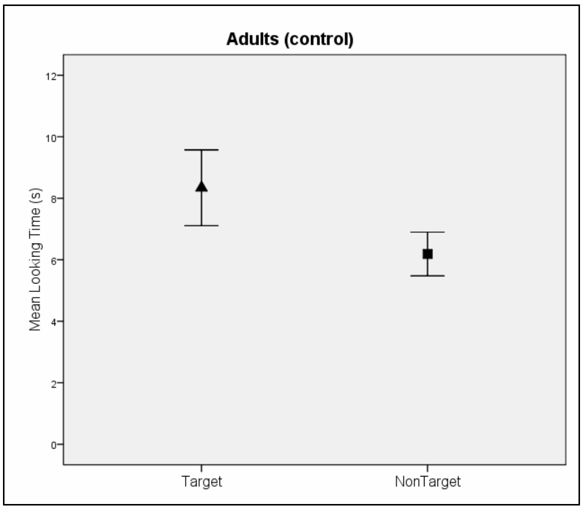


Figure 4. The average looking time (y-axis) towards target and non-target respectively for adults, with a confidence interval of 95%.

Discussion

Adults looks longer to the target sun than they do to the non-target sun and can so be assumed to detect the cross-modal magnitude match.

Infants between the ages of 7 and 10 months show a similar behaviour, as do infants at 5 months of age. The results from both 4- and 6-month-olds are non-significant and might be explained by the low number of subjects in those groups (n=4 and n=5 respectively). Additional data from 4- and 6-month-olds will be collected and added to the present data.

If adding data from more 4- and 6-month-old subjects confirms the present pattern, it may be assumed that the behaviour does not undergo any major changes during the age span covered in the experiment. However, pooling across all age groups suggests a stable looking preference towards the target sun. Expanding the age span covered would also be of interest to determine the youngest age at which cross-modal magnitude matching can be demonstrated.

If, even with added data the target-preference pattern does not emerge, different explanations as to why would be in order for 4- and 6-month-olds. Since 4 months is the youngest age covered in the current study, one possible explanation could be that they are too young for the cross-modal magnitude matching ability to be fully developed. This would, however, contradict the results of Lewkowicz and Turkewitz (1980), where even younger infants were reported to have this ability. One possible (if rather far-fetched) idea as to why 6-month-olds would behave differently than the other age groups is that at around 6 months of age, infants' perception of the world is largely restructured (in terms of caterorization and/or generalization). It is however far more likely that 6-month-olds will display the same pattern as the other age groups.

Once it has been thoroughly established that both multimodal perception of speech and the cross-modal magnitude matching ability is present in infancy, an interesting next step would be to present infants with conflicting cues. Infants' behaviour when presented with e.g. vowels whose intensity has been altered so that a vowel whose intensty is relatively low when occurring naturally is increased (and vice versa), as well as their articulations to match them to, would give answer to which of the two phenomena is more heavily relied upon. If infants in this situation rely more on general magnitude matching, they would prefer the most visually prominent articulation when presented with the loudest syllable. However, if they are more sensitive to the correlation between articulation and vowel quality, they would prefer the face with the correct articulation.

In conclusion, some evidence for cross-modal magnitude matching is found in 5- and 7- to 10-month-old infants and adults. Additional subjects in the groups of 4- and 6-month-olds will probably give more conclusive results for the entire age span investigated. Further experiments with conflicting cues (articulation matching vs. magnitude matching) are suggested.

Acknowledgments

The current study was funded by the Bank of Sweden Tercentenary Foundation as a part of the MILLE-project (K2003-0867).

The authors would like to thank Eeva Klintfors for help with experiment design and data collection and Simon Carlgren, Mathilda Eriksson and Tove Jörgensen for help with data collection. Thanks also to Kelly Smith and Fredrik Myr for proof-reading the paper.

References

Barker BA and Tomblin JB (2004). Bimodal speech perception in infant hearing aid and cochlear implant users. *Archives of Otolaryngology – Head & Neck Surgery,* 130: 582-586.

Bernstein LE and Benoît C (1996). For speech perception by humans or machines, three senses are better than one. In: *Proceedings of ICSLP-96.*

Burnham D (1998). Language specificity in the development of auditory-visual speech perception. In: R Campbell, B Dodd and D Burnham, eds, *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech.* UK: Psychology Press/Erlbaum (U.K.) Taylor & Francis, 27-60.

Erber NP (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40: 481-492.

Ewertsen HW and Nielsen HB (1971). A comparative analysis of the audiovisual, auditive and visual perception of speech. *Acta Oto-laryngologica*, 72: 201-205.

Klinfors E (2008). Emergence of words: multisensory precursors of sound-meaning associations in infancy. Doctoral thesis in Phonetics, Stockholm University, Stockholm.

Kuhl PK and Meltzoff AN (1982). The bimodal perception of speech in infancy. *Science*, 218: 1138-1140.

Kuhl PK and Melzoff AN (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, 7: 631-381.

Kuhl PK and Melzoff AN (1988). Speech as an intermodal object of perception. In: A Yonas, ed, *Perceptual development in infancy. The Minnesota symposia on child psychology: Vol. 20.* NJ: Lawrence Erlbaum Associates, Inc., 235-266.

Lacerda F, Klintfors E, Gustavsson L, Marklund E and Sundberg U (2005). Emerging linguistic functions in early infancy. In: L Berthouze, F Kaplan, H Kozima, H Yano, J Konczak, G Metta, J Nadel, G Sandini, G Stojanov and C Balkenius, eds, *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems.* Sweden: LUCS, 55-70.

Lehiste I and Peterson GE (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, 31: 428-435.

Lewkowicz DJ and Turkewitz G (1980). Cross-modal equivalence in early infancy: auditory-visual intensity matching. *Developmental Psychology*, 16: 597-607.

MacKain K, Studdert-Kennedy M, Spieker S and Stern D (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, 219: 1347-1349.

McGurk H and MacDonald J (1976). Hearing lips and seeing voices. *Nature*, 264: 229-239.

Rosenblum LD, Schmuckler MA and Johnson JA (1997). The McGurk effect in infants. *Perception & Psychophysics*, 59: 347-357.

Stevens SS (1957). On the psychophysical law. *Psychological Review*, 64: 153-181.

Stevens SS (1962). The surprising simplicity of sensory metrics. *American Psychologist*, 17: 29-39.

Stevens SS (1966). A metric for the social consensus. *Science*, 151: 530-541.

Sumby WH and Pollack I (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26: 212-215.

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

Phonetic and phonological aspects in relation to theory of mind and possible consequences for speech training in autism

Pia Nordgren
Department of Philosophy, Linguistics & Theory of Science, University of Gothenburg

Abstract

Recent neuroscientific research concerning dysfunctional mirror neurons in autism spectrum disorders may contribute to the way we look at autism from a linguistic point of view. Can areas that overlap neurologically, according to neuroscientific research, be proven to be related in a cognitive linguistic way? This paper will discuss theoretical aspects concerning phonology – autism – mirror neurons.

Introduction

Kuhl et al (2005) write: *" Infant learning from exposure to language may be dependent on both an initial ability to discriminate phonetic units and an early interest in listening to speech".*

In this sense language and social abilities can be said to be related. The early stages of language development include the ability to percept segments in language, that takes place during early stages of perceptual development. There is an ongoing discussion about early precursors for language development. Phoneme recognition could be a so called precursor because it relates both to social development and language development. According to Kuhl (2005) there is a relation between early perceptual abilities and later language acquisition. This can be studied in autism, which is a disorder with a variety of symptoms related to both language and social development. Autism spectrum disorders (ASD) is characterized by difficulties with social interaction, absence of theory of mind, pragmatic and other language difficulties such as phonological impairment. The pragmatic difficulties may be secondary to other difficulties with theory of mind and language. Current research on the mirror neuron system link imitation to motor actions (Lacoboni, Dapretto 2006). Mirror neurons are related to action understanding and theory of mind. The Mirror Neuron System is also thought to be involved in hearing and is therefore directly connected to perception and production of speech. This conclusion may have learning aspects in that when we hear other persons speak or say different sounds we are stimulated to imitate these sounds in our production of speech. If it is the case that motor areas for speech – Brocas area – is activated during listening to sounds (motor theory for speech perception), mirror neurons could be influenced because of the cortical overlap with Brocas area. This could theoretically affect theory of mind, which may be be trained because of the plasticity of the human brain. Pickering/ Garrod (2007) claim that motorically related brain areas may be activated before perceptual prediction and perceptual actions. (This is the case in monkeys.) They claim that articulation is involved in language comprehension. Listening to sounds and language comprehension thus activate the speech production system.

In summary neuroscientific research has a large proportion of studies that conclude the existence of an overlap between Brocas area and the mirror neuron regions. This leads to questions whether there exists a linkage between mirror neurons – theory of mind – and language, for example phonological processing. Both lexical, grammatical and phonological information are processed within Brocas area according to several authors. This is of relevance for autism spectrum disorders, which is a disorder thought to have dysfunctional mirror neurons.

Relationship between phonology and autism

In an article by Wolk/Edwards (1993) the relationship between phonological difficulties and language delay is discussed. Four possible alternatives are given:

- Phonological limitations may restrict language development

- Limitations concerning general language development may restrict phonological development

- Phonological and other language difficulties may be independent of one another

- Phonological and language difficulties may share a common basis

If we analyse these statements further we may see that limitations in auditive perception, auditive processing and speech production may restrict syllable level, word level, phonological phrases or phonological utterances. Limitations on language development on the other hand, may restrict morphosyntactic structure, the phonological utterance and have an impact on lexical ability and word acquisition. Phonological and language difficulties may also share a common basis, together with other domains. Wolk/Edwards (1993) discuss that the most plausible explanation is an underlying neurolinguistic substrate for both phonology and language skills. This leads us back to the question whether theory of mind and language skills are linked in some way. We know that mirror neuron regions and Brocas area are linked anatomically in the human brain, but can we also prove that theory of mind and phonological skills are linked in a cognitive linguistic way? Researchers have in some previous research tried to examine the phonology of children with autism, who we know exhibit difficulties with theory of mind. In the last decades there are only a few articles concerning phonology and autism. Goldstein (2002) finds only one article (Koegel, ODell, Dunlap 1988) during a period of twenty years that focuses on speech production in autistic children. Koegel, ODell, Dunlap (1988) investigated the difference between reinforcing motor speech versus speech attempts. Reinforcing speech attempts (with motivation as an important aspect) was more effective than reinforcing motor speech.

One question that has been asked is: Do they develop phonology in the same way as other children? Wolk/Giesen, (2000) and Wolk/Edwards, (1993) found evidence for delayed (deviant) phonological development as well as unusual sounds in autistic children. Bishop (2004) found no indication of disproportionate difficulties in autistic people of normal ability. Tager-Flusberg (1990) found that autistic children seem to follow the same general path as Downs syndrome children in their study and normal children reported in the literature. Bartolucci et al (1976) discussed the matter of auditory input. Autistic subjects are suggested to be deficient in their ability to extract the components that structures auditory input. The study found that autistic subjects differ significantly from the mentally retarded in the phonemic substitutions which they make. Autistic subjects were also characterized by a high correlation between high frequency of phonological errors and low level of overall language development. The study also concluded that systematic investigations of the speech sound development of autistic children are absent in the literature. Bartolucci made an assumption that autistic children may generate phonemes by utilizing different strategies, which was weekly supported in the study. In that case the same end result at the phonological level could be reached by different processes. Rapin and Dunn (2003) found in their review article two subtypes in autistic children: one subtype concerning phonology and syntax and one subtype related to semantic difficulties and pragmatics. The authors of this article claim that there is a support that young autistic children are language disordered as well as autistic. Bortolini/Leonard, (2000) address the question of whether phonology is an area of special difficulties for these children or if phonology is impaired as a natural consequence of a more general limitation in language learning? Another aim was to investigate the centrality of phonology in the disorder in relation to morphosyntax. The authors of this article also discuss the impact of prosody on grammatical morphology in these children. This study revealed that the childrens phonological and grammatical morpheme limitations were separable to some degree.

Discussion and further work

We have now seen that there are very few articles about phonology and autism. Three articles (Bishop, 2004; Wolk/Edwards, 1993; Wolk/Giesen, 2000) used phonetic inventory/ process analysis as a linguistic method.

Generative phonology is not generally referred to, but may be the ground for much of the research implicitly. Distinctive features are referred to by Bartolucci et al (1976) and Koegel, ODell and Dunlap (1988), who use the distinctive feature analysis in their research. Wolk/Giesen use natural segment classes. Concerning phonological development in autistic children, some researchers have found delayed pattern in autistic children similar to mentally retarded children (Bartolucci/Pierce, 1977; Tager-Flusberg et al 1990). Wolk/Giesen, (2000) and Wolk/Edwards (1993) found delayed/ deviant development in autistic children concerning phonology. Only one article investigates perceptual abilites in autistic children – the article by Bartolucci/Pierce (1977). Bortolini/Leonard (2000) examined the relationship between structural constraints and use of grammatical morphology in SLI (=specific language impairment). No studies of this kind exist concerning children with autistic disorder. Mostly segmental levels are investigated in the literature. Tager-Flusberg, (1990) however, investigated morphological structures in a speech corpus and Bortolini/Leonard examined morphology in relation to constraints. The two subtypes found by Rapin & Dunn (2003), need to be examined more. If it is true that there are different subtypes: one semantic – pragmatic and one phonological – syntactic this must be taken into consideration when it comes to treatment of autistic children. There are several possible questions to ask concerning this subject. How are different language modules related? How is phonology related to word acquisition? How is language and joint attention related? Can imitative ability as a motor event improve by listening to sounds?

From my literature study I conclude the following:

Listening to sounds may influence cortical brain areas and affect linguistic prediction – phonological awareness and speech production in children with for example SLI/autism. If motor areas for speech – the so called Brocas area is activated according to motor theory for speech perception, mirror neurons could be influenced because of the overlap between Brocas area and theory of mind. This could theoretically affect theory of mind in autism because of the brain´s plasticity. The knowledge of mirror neurons can be summarized as "knowing how" to perform an action. This would be of importance for children with autism, who lack functional mirror neurons.

For my future empirical study I will hypothesize that children with autism will improve their speech and language by perceptual speech therapy. I will assume that cortical brain areas will be influenced by listening to sounds and that motor areas for speech may also be affected according to the motor theory for speech perception. This will also be in accordance with development of theory of mind because of the overlap of brain regions (Brocas area and mirror neuron regions).

My general hypothesis is:

1. Children with autism will improve their phonological ability and maybe their general speaking ability in that prevalence of certain target phonemes composed of distinctive features will improve by training with auditive stimuli.

Specific hypotheses:

2. Phonological (or semantic) errors will be reduced by perceptual speech therapy.
3. Secondary effects will be seen on words in different categories, that will increase, such as prevalence of nouns, verbs etc.
4. Syllable constructions will be affected.
5. Mean length of utterance – MLU will increase after therapy.
6. Complexity of syntactic structure will be affected.

Subjects: Subjects are autistic children, that are diagnosed with autism. The plan is to involve high-functioning autistic children with an IQ above 70, in the study. Preferably the group of subjects is as homogenous as possible. The children must have normal vision and hearing and preferably be about 6 years of age at therapy start. If possible only boys will be enrolled in the study to rule out gender as a variable. The children must have difficulties with language form and communication.

The plan is to try these hypotheses with a single subject design across behaviours. In a single-subject design the same individual´s performance is measured repeatedly and you may according to Hegde (1994) get a better result than with pre/posttests. With this design you do not present results in form of averages and subjects are not selected randomly. Focus will be on distinctive features in order to get a better possibility to analyze patterns in the children´s speech than with phonetic/phonemic inventories. This study may involve types of distinctive features; place features, laryngeal features and manner features. When the first feature is treated, the two remaining features still remain in baseline and so on.

Measures will be taken on prevalence of target phonemes. A qualitative analysis of semantic/phonological errors, word production per se, prevalence of words in different categories, syllable constructions will be made. Complexity of syntactic structure will then be performed together with MLU. Correlations between non-linguistic mental age and number of phonological errors in perception and production tasks will also be looked upon together with correlations between perception and production, IQ-level and language, language ability related to level of autism and age in relation to other variables such as language ability and treatment results. Further questions are: Is there a correlation between syntactic level and number of phonemes? How is lexical ability related to phoneme complexity? How is syntactic level and MLU related? This also needs to be found out in the study.

References

Bartolucci G et al (1976). Phonological Investigation of Verbal Autistic and Mentally Retarded Subjects. *Journal of Autism and Childhood Schizophrenia,* 6:4, 303-316.

Bartolucci G, Pierce S J (1977). A Comparison of Phonological Development in Autistic, Normal and Mentally Retatarded Subjects. *British Journal of Disorders of Communication,*12:2, 137 – 147.

Bishop D, Maybery M et al (2004). Are Phonological Processing Deficits Part of the Broad Autism Phenotype? *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics),* 128 B: 54-60.

Bortolini U, Leonard L (2000). Phonology and children with specific language impairment: status of structural constraints in two languages. *Journal of Communication Disorders,* 33:131-150.

Goldstein H (2002). Communication Intervention for Children with Autism: A Review of Treatment Efficacy, *Journal of Autism and Developmental Disorders*: 373–396.

Hegde M N (1994). Clinical Research in Communicative Disorders – Principles and Strategies, Pro-Ed, Austin, Texas.

Koegel R, ODell M et al (1988). Producing speech use in nonverbal autistic children by reinforcing attempts, *Journal of Autism and Developmental Disorders* 18 (4): 525-538.

Kuhl P K, Conboy B T et al (2005). Early Speech Perception and Later Language Development: Implications for the “Critical Period”. *Language Learning and Development* 1 (3): 237 – 264.

Lacoboni M and Dapretto M (2006). The Mirror Neuron System and the Consequences of its dysfunction,NatureReviewsNeuroscience,7: 942–951.

Pickering M and Garrod S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11:3, 105 – 110.

Rapin I, Dunn M (2003). Update on the language disorders of individuals on the autistic spectrum, *Brain & Development* 25 (3): 166-172.

Tager-Flusberg, H et al (1990). A Longitudinal Study of Language Acquisition in Autistic and Down Syndrome Children, *Journal of Autism and Developmental Disorders,* 20 (1):1-21.

Wolk L, Edwards M L (1993). The Emerging Phonological System of an Autistic Child, *Journal of Communication Disorders* 26 (3):161 – 77.

Wolk, L, Giesen, J (2000). A Phonological Investigation of Four Siblings with Childhood Autism, *Journal of Communication Disorders* 33 (5): 371 – 89.

The effect of speaking rate on the perception of vowel-initial and vowel-final [h]

Jörgen L. Pind

Faculty of Psychology, University of Iceland

Abstract

Numerous studies have shown that the perception of speech segments is often rate-dependent. Thus, with faster speaking rate the boundary between, say, [b] and [p], measured in terms of voice-onset time (VOT), moves to shorter values of VOT. The present paper reports an experiment where the perception of vowel-initial [h] and vowel-final [h], i.e. preaspiration, is compared in words at three different speaking rates. With slower speaking rate the boundaries, as expected, move to longer values of aspiration and preaspiration respectively, though, interestingly, less so in the case of aspiration than in the case of preaspiration.

Introduction

It is common knowledge that speech is a highly variable signal and invariant cues for individual speech segments are not easily established, (though see Pind, 1995b; Stevens and Blumstein, 1981). The sources of variability in the speech signal are many. Some of these sources are extra-linguistic, do not directly concern the linguistic message being conveyed. One such factor is speaking rate which has a significant effect on the duration of individual speech segments. Since some speech cues are primarily temporal in nature, defined by duration or other such temporal measures, speaking rate can obviously affect such cues. To take just one example: Icelandic has long and short vowels and consonants that are for the most part defined by their durations (Einarsson, 1927; Pind, 1986, 1999). Since speech rate affects the segmental durations it can easily happen that a quantitatively long vowel, spoken at a fast utterance rate, will be shorter than a quantitatively short vowel, spoken at a slow rate.

How does the listener deal with such temporal variability? In the main, two different theoretical viewpoints have been advanced, both boasting of long traditions in perceptual psychology. One of them posits a process of *normalization* where the perceptual system is thought to 'take account of' any external factors (for an interesting discussion of this idea in perceptual psychology, see Epstein, 1973). The other viewpoint asserts that invariants can in fact be found in the speech signal, even for speech segments which are highly variable. A case has thus been been made that the ratio of vowel to consonant duration can function as a *higher-order* invariant for quantity in Icelandic (Pind, 1995b), which fits nicely with the perceptual theories of James J. Gibson, e.g. Gibson (1959).

One particular temporal speech cue which has been the focus of great interest is that of VOT or voice-onset-time, the time from the release of a stop consonant to the onset of voicing in the following vowel (Lisker and Abramson, 1964). This speech cue differentiates voiced or unaspirated stop consonants such as [bdg] from voiceless aspirated stops [ptk]. The former series has short VOTs (perhaps 0–20 ms) , the latter longer VOTs (perhaps 50–70 ms). The precise values of VOT are language dependent (Cho and Ladefoged, 1999).

Previous research has shown that VOT is sensitive to speaking rate, with VOTs lengthening at slower speaking rate. This holds especially for the stop category with the longer VOTs, i.e. the voiceless/aspirated series of stops. Perceptual experiments have shown that listener's phoneme boundaries are affected by the duration of surrounding speech segments. Thus, the longer the vowel following a syllable-initial stop, the longer the VOT needed to cue voicelessness/aspiration (Miller et al., 1986; Summerfield, 1981).

Icelandic phonetics has a rather uncommon feature (Ladefoged and Maddieson, 1996) termed *preaspiration*. This is an [h]-like segment at the end of a vowel before stop closure. Consider thus a word-pair like *akur* [aːkʏr] 'field', vs. *akkur* [ahkʏr] 'advantage'. In the first word a long vowel is followed by the closure for the stop [k], in the latter word a short vowel is followed by preaspiration [h] before the closure of the stop. In perceptual research, preaspiration is readily cued by voice-offset time (VOffT), a speech cue which is in most respects the mirror image of VOT. (Usually VOT is defined to include the initial burst for the stop consonant as well. There is of course no burst following the VOffT, the aspiration is followed by the closure of the stop.)

Previous experiments (Pind, 1995a, 1998) have shown that the perception of preaspiration in Icelandic is sensitive to the duration of the previous vowel. The longer the vowel, the longer the aspiration needs to be for the listener to perceive a pre-aspirated segment.

The experiment reported in this paper compares the influence of speaking rate on the perception of preaspiration and initial [h]. This is done using the word-pair [aːka]–[ahka] on the one hand with the word-pair [aːka]–[haːka] on the other hand. It is hypothesized that in both cases, the slower the speaking rate, the longer the aspiration needs to be to cue the perception, either of vowel-initial aspiration [h-] or of vowel-final aspiration [-h], i.e. preaspiration.

Experiment 1

Method

Participants

Ten undergraduate students of psychology at the University of Iceland participated in the experiment. They were all native speakers of Icelandic and reported normal hearing.

Stimuli

A total of six synthetic speech continua were used in the experiment, made with the Sensyn speech synthesizer, a version of the Klatt-synthesizer (Klatt, 1980; Klatt and Klatt, 1990). Three of the continua were of the type [aːka]–[haːka], three of the type [aːka]–[ahka]. The three continua were distinguished by the length of first vowel and following closure in the first syllable. In the first continuum the vowel was 150 ms long and the closure 75 ms, in the second the vowel was 200 ms long and the closure 100 ms, and in the third and final continuum the vowel was 250 ms long and the closure 125 ms long. In all cases the ratio of vowel duration to the duration of the following closure is thus constant and typical for a phonemically long vowel, followed by a phonemically short stop. The VC durations of the first syllable thus range from 225 ms, through 300 ms to 375 ms. This lengthening of the initial syllable is perceived as a slowing of speaking rate.

In all cases the second syllable started after the closure with a 25 ms long burst and aspiration, followed by a 75 ms long voiced vowel, the second [a] vowel in the word.

The steady state vowel formants for both [a] tokens were set at 770 Hz for F1, 1280 Hz for F2 and 2425 Hz for F3. For the transitions into the [k]-closure the formants moved linearly to respectively 200 Hz, 1600 Hz and 2000 Hz over 45 ms. The transitions into the second vowel were a mirror image of the transitions into the closure.

Within each stimulus continuum aspiration (or preaspiration) was varied in 8 ms steps from 0 to 80 ms. Each continuum thus contained 11 stimuli, bringing the total number of stimuli to 66 for the whole experiment. Aspiration was cued by disabling voicing (synthesis parameter **AV** set to 0 dB), turning on aspiration (**AH** = 40 dB) and frication (**AF** = 55 dB, **A2F** = 50 dB, and **A3F** = 40 dB), and increasing the bandwidth of the first formant (**B1**) from 60 to 200 Hz. The fundamental frequency during the voiced portion of the stimuli was fixed at 125 Hz and the synthesizer was set to use an update interval of 4 ms. The sampling frequency was 11.025 Hz. Figure 1 shows spectrograms of three stimuli from the experiment.

The stimuli were recorded onto two tapes, by playing them through a Turtle Beach Multisound card. One tape contained the stimuli for the [aːka]–[haːka] continua, the other the tapes for the [aːka]–[ahka] continua. Each tape started with a practice block consisting of all 33 stimuli played in randomized order. This was followed by five blocks, each of which contained two tokens of each stimulus. These were presented randomly, a total of 66 stimuli in each block. The interstimulus interval was 2.5 seconds.

Procedure

Five participants listened to the [aːka]–[haːka] tape followed by the [aːka]–[ahka] tape, for the other five participants the order was reversed. The testing took place in a quiet room. The participants listened to the stimuli, which were played at a comfortable listening level, over Sennheiser HD-530-II circumaural headphones.

Participants were provided with response sheets which contained two fields for each stimulus presented. In the [aːka]–[haːka] test the fields were marked with the words *aka* on the one hand and *haka* on the other. In the [aːka]–[ahka] test the words were *aka* on the one hand and *akka* on the other using the normal Icelandic orthography. Participants were instructed to mark the appropriate box for each stimulus presented and to guess if they were not sure which response was appropriate. Participants did not report any difficulty in carrying out this task. With the first 33 trials

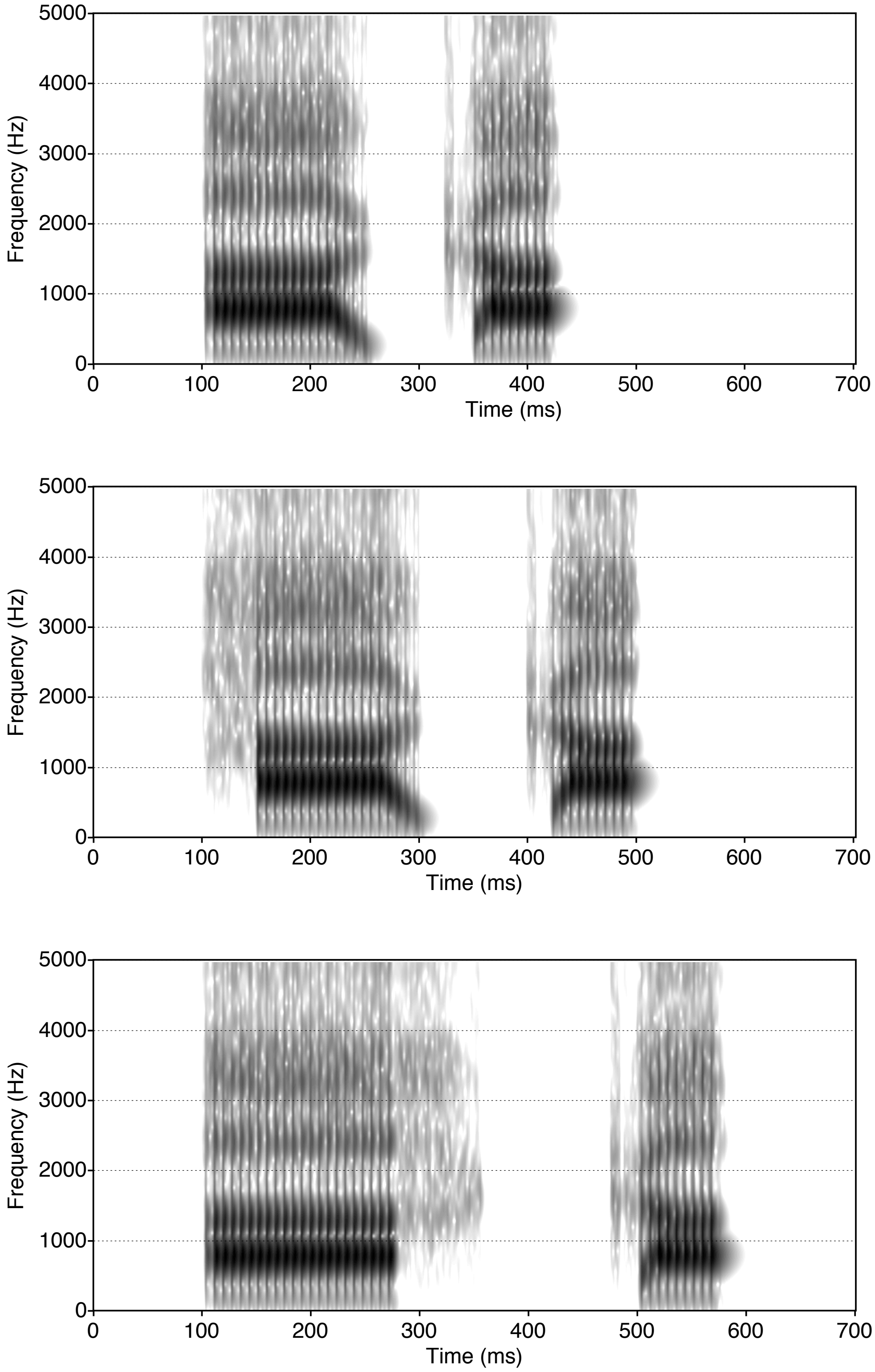


Figure 1: Spectrograms of three stimuli (out of a total of 66) used in the experiment. The topmost spectrogram shows a stimulus with an initial 150 ms long vowel followed by a 75 ms long closure. The stimulus contains neither aspiration nor preaspiration. In the middle spectrogram the first vowel is 200 ms long (including 48 ms of VOT (aspiration) at the beginning of the vowel) followed by a 100 ms long closure. The final spectrogram shows a stimulus with a 250 ms long vowel (including 80 ms of VOffT (preaspiration) at the end of the vowel) followed by a 125 ms long closure.

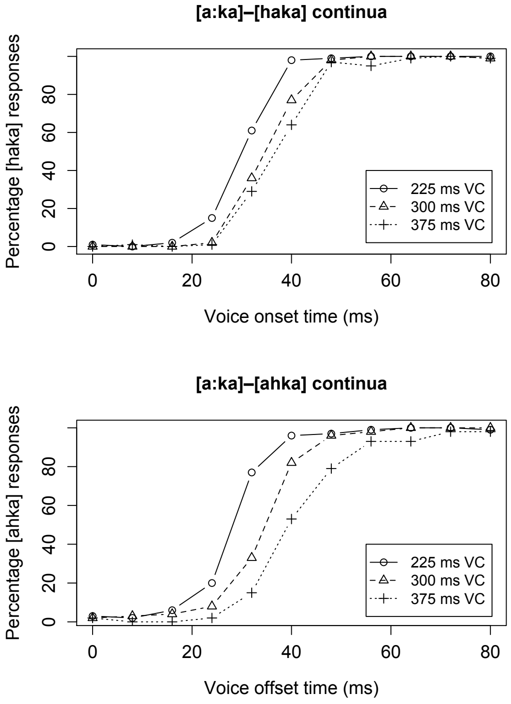


Figure 2: Pooled identification curves for all 10 participants in the present experiment. The upper graph shows the results for the stimulus continua with vowel-initial aspiration. The lower graph shows the results for the three preaspiration continua.

of each run used for familiarization, 10 responses were tabulated for each participant per stimulus.

Results and discussion

Pooled identification curves for all 10 participants are shown in figure 2. The figure clearly show that the percentage of aspiration responses ([haka] or [ahka]) increased as the duration of vowel-initial aspiration, voice-onset time, or vowel-final aspiration, voice-offset time, increases. Additionally the figures clearly show the effect of VC duration on the percentage of aspiration responses. As VC duration increases the phoneme boundaries for the perception of aspiration or preaspiration move to longer values of VOT and VOffT respectively.

Phoneme boundaries were calculated for individual participants using the method of probits (Finney, 1947). The calculation were done in R with the help of the MASS library (Venables and Ripley, 1999). The average phoneme boundaries are shown in Table 1.

Table 1: Average phoneme boundaries (10 participants) for the six stimulus continua of the present experiment. The values denote milliseconds of aspiration.

VC duration (ms)	[aːka]–[ahka] continua	[aːka]–[haːka] continua
225	28.30	29.69
300	33.85	35.10
375	41.58	37.48

A two-factor repeated measures ANOVA (stimulus series × VC duration) shows that the effect of stimulus series is not significant, $F(1,9)$ = 0.028, whereas the effect of VC duration is significant, $F(2,18)$ = 92.25, $p < 0.001$. The interaction of series and VC duration is significant, $F(2,18)$ = 7.83, $p < 0.01$.

Repeated paired t-tests show that average phoneme boundary in the words with the 300 ms long VC is significantly longer (34.46 ms) than in the words with the 225 ms long VC (29 ms), $t(38)$ = 3.66, $p < 0.001$. Again the average phoneme boundaries in the words with the 375 ms long VC is significantly longer (39.53 ms) than in the words with the 300 ms long VC, $t(38)$ = 2.674, $p < 0.05$.

The interaction of stimulus series and VC duration can be inferred from Figure 2. Here it can be seen that the phoneme boundaries move to steadily longer values of voice-offset time in the [aːka]–[ahka]] series, in approximately equal steps, confirm also Table 1. In the [aːka]–[haːka] series there is a clear movement of the phoneme boundaries as VC duration increases from 225 to 300 ms (change of 5.41 ms), with a much smaller increase from 300 to 375 ms (change of 2.38 ms). In the [aːka]–[ahka] series the corresponding changes in the phoneme boundaries are 5.55 ms and 7.73 ms. So in the case of preaspiration the movement of the phoneme boundaries increases as the vowel is lengthened, in the case of vowel-initial aspiration it is the other way around, the movement of the phoneme boundaries decreases with increasing vowel length.

How can this be explained? From earlier experiments it is clear that vowel quantity has a decisive influence on the perception of preaspiration and this is so regardless of whether quantity is cued by vowel duration (Pind, 1996a) or the spectrum of the vowel (Pind, 1998). (The vowel spectrum plays an important role in the perception of quantity of the three central vowels of Icelandic, Pind (1996b).) Since care was taken in this experiment to keep the quantity constant over the different stimulus continua by keeping the vowel to closure duration ratio fixed the effects of quantity would seem to be ruled out, though admittedly a direct test of perceived vowel duration was not undertaken. This needs to be addressed in further experiments. If there is a tendency for the stimuli with the longest VC duration to be perceived as perhaps having a more robust – or more prototypical – long vowel than in the stimuli with the shortest VC then this would undoubtedly influence the perception of preaspiration more than the perception of initial [h]. If this is not the case, then an explanation needs to be sought elsewhere. Perhaps the weak [h] sound following the loud vowel is more susceptible to masking than the word-initial [h] (cf. Bladon, 1986, for a similar hypothesis). Further experiments are needed to explore this issue in greater detail.

References

Bladon A (1986). Phonetics for hearers. In G McGregor, ed., *Language for hearers*, 1–24. Oxford: Pergamon Press.

Cho T and Ladefoged P (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, 27:207–229.

Einarsson S (1927). *Beiträge zur Phonetik der isländischen Sprache*. Oslo: A. W. Brøgger.

Epstein W (1973). The process of 'taking-into-account' in visual perception. *Perception*, 2:267–285.

Finney D J (1947). *Probit analysis*. Cambridge: Cambridge University Press.

Gibson J J (1959). Perception as a function of stimulation. In S Koch, ed., *Psychology: A study of a science. Volume I. Sensory, perceptual and physiological formulations*, 456–501. New York: McGraw-Hill Book Company.

Klatt D H (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971–995.

Klatt D H and Klatt L C (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857.

Ladefoged P and Maddieson I (1996). *The sounds of the world's languages*. Oxford: Blackwell.

Lisker L and Abramson A S (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20:384–422.

Miller J L, Green K P and Reeves A (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, 43:106–115.

Pind J (1986). The perception of quantity in Icelandic. *Phonetica*, 43:116–139.

Pind J (1995a). Constancy and normalization in the perception of voice offset time as a cue for preaspiration. *Acta Psychologica*, 89:53–81.

Pind J (1995b). Speaking rate, VOT and quantity: The search for higher-order invariants for two Icelandic speech cues. *Perception & Psychophysics*, 57:291–304.

Pind J (1996a). Rate-dependent perception of aspiration and pre-aspiration in Icelandic. *Quarterly Journal of Experimental Psychology*, 49A:745–764.

Pind J (1996b). Spectral factors in the perception of vowel quantity in Icelandic. *Scandinavian Journal of Psychology*, 37:121–131.

Pind J (1998). Auditory and linguistic factors in the perception of voice offset time as a cue for preaspiration. *Journal of the Acoustical Society of America*, 103:2117–2127.

Pind J (1999). Speech segment durations and quantity in Icelandic. *Journal of the Acoustical Society of America*, 106:1045–1053.

Stevens K N and Blumstein S E (1981). The search for invariant acoustic correlates of acoustic features. In P D Eimas and J L Miller, eds., *Perspectives on the study of speech*, 1–38. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Summerfield Q (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7:1074–1095.

Venables W N and Ripley B D (1999). *Modern applied statistics with S-PLUS*. Springer, third edn.

Synthesising intonational varieties of Swedish

Susanne Schötz[1], Jonas Beskow[2], Gösta Bruce[1], Joakim Gustafson[2], Björn Granström[2], My Segerup[1]

[1] Dept. of Linguistics & Phonetics, Centre for Languages & Literature, Lund University
[2] Dept. of Speech, Music & Hearing, School of Computer Science & Communication, KTH

Abstract

Within the research project SIMULEKT (Simulating Intonational Varieties of Swedish), our recent work includes two approaches to simulating intonation in regional varieties of Swedish. The first involves a method for modeling intonation using the SWING *(*SW*edish* IN*tonation* G*enerator) tool, where annotated speech samples are resynthesised with rule-based intonation and audio-visually analysed with regards to the major intonational varieties of Swedish. The second approach concerns a method for simulating dialects with HMM synthesis, where speech is generated from emphasis-tagged text. We consider both approaches important in our aim to test and further develop the Swedish prosody model, as well as to convincingly simulate Swedish regional varieties using speech synthesis.*

Background

Our object of study in the research project SIMULEKT (Simulating Intonational Varieties of Swedish) (Bruce et al., 2007) is the prosodic variation characteristic of different regions of the Swedish-speaking area. The SIMULEKT project, supported by the Swedish Research Council 2007-2010, is a collaboration between Linguistics, Lund University, and Speech, Music, Hearing, KTH, Stockholm. The primary goal of the project is to gain more precise knowledge about the major intonational varieties of Swedish. A concomitant goal is to develop the Swedish prosody model theoretically and experimentally. In addition to the regular description of intonational patterns from studying F_0contours, the use of speech synthesis in different forms is a major feature of our research project. Our starting-point is that prosody and specifically intonation is a fundamental constituent of the different, native accents characterising the distinct regional varieties of Swedish. We believe that studying the pitch patterns of different varieties of a language like Swedish contrastively will sharpen our analysis and description of their intonation. The main regional varieties or dialect groups of Swedish are South, Göta, Svea, Dala, Gotland, North, and Finland Swedish. They are also considered to be the major intonational varieties of Swedish. Figure 1 shows a map of these regions, corresponding to our present dialect classification scheme.

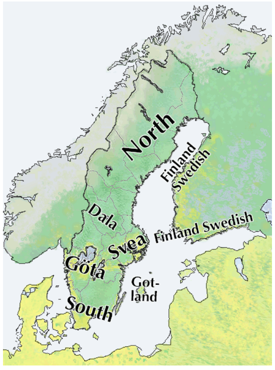


Figure 1: Approximate geographical distribution of the seven main regional varieties of Swedish.

The Swedish prosody model

The main parameters of the Swedish prosody model (Bruce and Gårding, 1978; Bruce and Granström, 1993; Bruce, 2007) are for word prosody 1) word accent timing, i.e. timing characteristics of pitch gestures of word accents (accent1/accent 2) relative to a stressed syllable, and 2) pitch patterns of compounds, and for utterance prosody 3) intonational prominence levels (focal/non-focal accentuation), and 4) patterns of concatenation between pitch gestures of prominent words. Thus the specific timing of pitch patterns of word accents in relation to stressed syllables (both primary and secondary), i.e. in simplex as well as in compound words, is considered to be a potential distinctive feature for the classification of different intonational varieties. Major emphasis in the SIMULEKT project has been given to postlexical prosody, specifically phrase and utterance intonation, assuming intonational prominence levels and tonal concatenation patterns to be particularly revealing for the impression and identification of different native accents of Swedish. There are regional varieties of Swedish, which can be shown to be exploiting two levels of intonational prominence (focal vs. non-focal accentuation), specifically the Central Swedish regional varieties Svea and Göta. In these varieties there will be a fairly regular alternation between focal and non-focal accentuation for successive accented words of a phrase. Other varieties have more equal weighting between such words of a phrase and regularly exploit only one level of intonational prominence as a characteristic feature, namely the geographically more peripheral varieties South, Gotland, North, and also Finland Swedish as well as Dala. We have also been emphasising patterns of tonal concatenation, both as coherence patterns between prominent words – high/low plateau, downslope/upslope – and as boundary signalling. These patterns would also appear to contribute to our differentiating between distinct varieties of Swedish.

Methodology

Speech databases

Our main sources for analysis here are the three Swedish speech databases SweDia 2000, SpeechDat and NST (Nordisk Språkteknologi 'Nordic Language Technology').

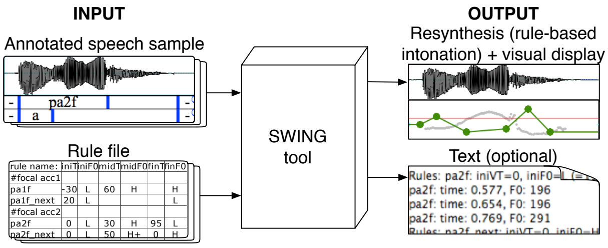
The SweDia 2000 database comprises a word list, an elicited prosody material, and spontaneous monologues from 12 speakers (younger and elderly men and women) each from more than 100 different places in Sweden and Swedish-speaking parts of Finland, selected for dialectal speech. From this database a minor elicited prosody material and primarily the extensive spontaneous speech material are relevant for our project work.

SpeechDat (Elenius, 1999) contains speech recorded over the telephone from 5000 speakers, registered by age, gender, current location and self-labeled dialect type, according to Elert's suggested Swedish dialect groups (Elert, 1994) that is a more fine-grained classification with 18 regions in Sweden. This database contained two particularly interesting read sentences to our project: *Mobiltelefonen är nittiotalets stora fluga, både bland företagare och privatpersoner.* 'The mobile phone is the big hit of the nineties, both among business people and private persons' and *Flyget, tåget och bilbranschen tävlar om lönsamhet och folkets gunst* 'Airlines, train companies and the automobile industry are competing for profitability and people's appreciation'.

The main data used in our HMM approach are from the Norwegian Språkbanken. This large speech synthesis database from a professional speaker of standard Swedish was recorded as part of the NST synthesis development. About 5000 read sentences are included in the corpus, adding up to about 11 hours of speech.

Analysing intonation with SWING

An important part of our work concerns analysis and modeling of Swedish intonation by resynthesis. The SWING (SWedish INtonation Generator) tool was developed for this task. It comprises several parts joined by the speech analysis software Praat (Boersma and Weenink, 2010), which also serves as graphical interface. Using an input annotated speech sample and an input rule file, SWING generates and plays PSOLA resynthesis – with rule-based and speaker-normalised intonation – of the input speech sample. Additional features include visual display of the output on the screen, and options for printing various kinds of information to the Praat console (Info window), e.g. rule names and values, the time and F_0 of generated pitch points etc. Figure 2 shows a schematic overview of the tool.



rule name:	iniT	iniF0	midT	midF0	finT	finF0
#focal acc1						
pa1f	-30	L	60	H		H
pa1f_next	20	L				L
#focal acc2						
pa2f	0	L	30	H	95	L
pa2f_next	0	L	50	H+	0	H

```
Rules: pa2f: iniVT=0, iniF0=L (=
pa2f: time: 0.577, F0: 196
pa2f: time: 0.654, F0: 196
pa2f: time: 0.769, F0: 291
Rules: pa2f_next: iniVT=0, iniF0=H
```

Figure 2: Schematic overview of the SWING *tool.*

Input speech material

The input speech sample to be used with the tool is manually annotated. Stressed syllables are labelled prosodically and the corresponding vowels are transcribed orthographically. Figure 3 displays an example utterance with prosodic annotation: *tävlar om lönsamhet och folkets gunst* 'are competing for profitability and people's appreciation', while Table 1 shows the prosodic labels that are handled by the current version of the tool.

Rules

The Swedish prosody model is implemented as a set of rule files – one for each regional variety in the model – with timing and F_0 values for critical points in the rules. These files are text

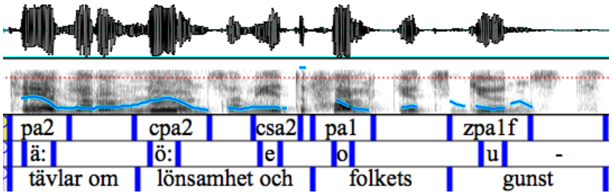


Figure 3: Example of an annotated input speech sample.

Table 1: Prosodic labels used for annotation of speech samples to be analysed by SWING.

Label	*Description*
pa1	primary stressed (pre-focal) accent 1
pa1f	focal focal accent 1
pa1po	primary stressed (post-focal) accent 1
zpa1f	finally stressed focal accent 1
pa2	primary stressed (non-focal) accent 2
pa2f	focal focal accent 2
cpa	primary stressed (ps) compound accent
csa	secondary stressed (ss) compound accent
cpaf	focal ps compound accent
csaf	focal ss compound accent
cpapo	post-focal ps compound accent
csapo	post-focal ss compound accent

files with a number of columns; the first contains the rule names, and the following comprise three pairs of values, corresponding to the timing and F_0 of the critical pitch points of the rules. The three points are called *i* (initial), *m* (medial), and *f* (final). Each point contains values for timing (*T*) and F_0 (*F0*). Timing is expressed as a percentage into the stressed syllable, starting from the onset of the stressed vowel, which is the default. Three values are used for F_0: L (low), H (high) and H+ (extra high, used e.g. in focal accents). The pitch points are optional; they can be left out if they are not needed by a rule. New rules can easily be added and existing ones adjusted by editing the rule file. Table 2 shows an example of the rules for Gothenburg (Göta) Swedish, adjusted in accordance with the Swedish prosody model and Segerup (2005). Several rules change values when applied to phrase-final position (see *(pf)* in Table 2), and some rules contain a second part, which is used for the pitch contour of the following (unstressed) interval (segment) in the annotated input speech sample. This extra part has '*_n*' (n stands for next segment, the domain of the extra gesture) attached to its rule name. Examples of such rules are *pa1f*, *zpa1f*, *pa2f* and *csaf* in Table 2.

Table 2: Example rule file for Gothenburg (Göta) Swedish with timing (T) and F_0 (F0) values for initial (i), mid (m) and final (f) points. Values in italics within parentheses are phrase-final (pf) rules. For rule name abbreviations, see Table 2.

Rule name	iT/*(pf)*	iF0/*(pf)*	mT/*(pf)*	mF0/*(pf)*	fT/*(pf)*	fF0/*(pf)*
global		L				L
pa1	-40	L		H	70	L
pa1f	-40	L		H	70	L
pa1f_n	20	L		/*(60)*	/*(H+)*	
pa1po				H+	70	L
zpa1f	-50	L	-20	H	30	L
zpa1f_n	-20	H+				
pa2	-50/*(0)*	L/*(H+)*	/*(100)*	H+/*(L)*	/*(180)*	L/*(H)*
pa2f	-40	L	20/*(0)*	H+/*(H)*	/*(70)*	L
pa2f_n	40	L			70/*(90)*	H+/*(H)*
cpa	-20	L	20	H		L
csa						L
cpaf	-20	L	20	H+		L
csaf						L
csaf_n	/*(20)*	/*(L)*			/*(80)*	/*(H+)*
cpapo			20	H+		L
csapo						L

Procedure

Analysis with SWING is fairly straightforward. The user selects one input speech sample and one rule file to be used with the tool, and which (if any) text (rules, pitch points, debugging information) to print to the Praat console. A Praat script generates resynthesis of the input speech sample with a rule-based output pitch contour based on 1) the pitch range of the input speech sample, used for speaker normalisation, 2) the annotation, used to find the time and pitch gestures to be generated, and 3) the rule file, containing the values of the critical pitch points. The Praat graphical user interface provides immediate audio-visual feedback of how well the rules work, and also allows for easy additional manipulation of pitch points with the Praat built-in *Manipulation* feature.

Testing the Swedish prosody model

SWING has been used in our work with testing and developing the Swedish prosody model for simplex and compound words as well as phrasing. Testing is done by selecting an input speech sample and a rule file of the same intonational variety. If the model works adequately, there should be a close match between the F_0 contour of the original version and the rule-based one generated by the tool. Figure 4 shows example SWING output of three phrases for the two intonational varieties Göta and South Swedish. As can bee seen there is a close match between the original pitch of the input speech samples and the simulated pitch contour in all phrases.

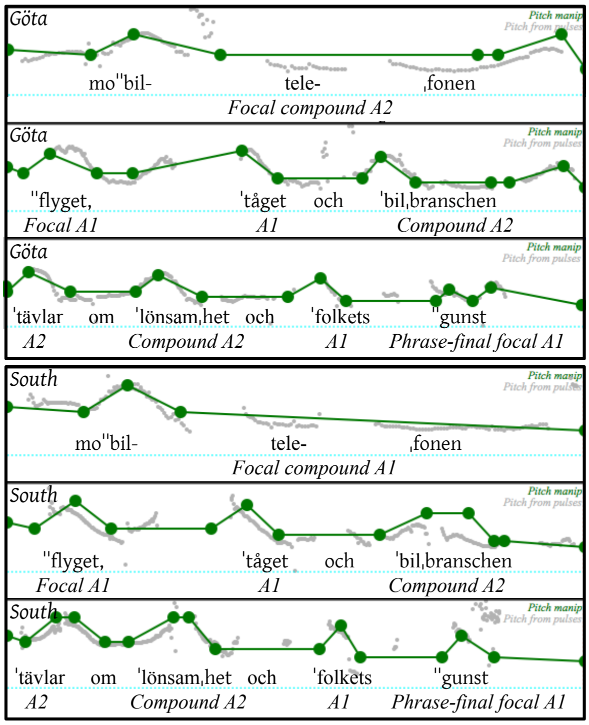


Figure 4: Simulation examples in SWING. *Three phrases in Göta and South Swedish (simulation: circles connected by solid line; original pitch: light-grey line; A1: accent 1; A2: accent 2).*

HMM synthesis

During the last decade, most speech synthesisers have been based on prerecorded pieces of speech resulting in improved quality, but with lack of control in modifying prosodic patterns (Taylor, 2009). The research focus has been directed towards how to optimally search and combine speech units of different lengths. A synthesis approach that has gained interest in recent years is HMM based synthesis (Tokuda et al., 2000). In this solution the generation of speech is based on a parametric representation, while the grapheme-to-phoneme conversion still relies on a large pronunciation dictionary. This approach has been successfully applied to a large number of languages, including Swedish (Lundgren, 2005).

HMM synthesis is an entirely data-driven approach to speech synthesis. As such it gains all its knowledge about segmental, intonational and durational variation in speech from training on an annotated speech corpus. Given that the appropriate features are annotated and made available to the training process, it is possible to synthesise speech with high quality at both segmental and prosodic levels. Another important feature of HMM synthesis that makes it an interesting choice in studying dialectal variation, is that it is possible to adapt a voice trained on a large data set (2-10 hours of speech) to a new speaker with only 15-30 minutes of transcribed speech (Watts et al., 2008). In this study we used 20-30 minutes of dialectal speech for experiments on speaker adaption of the initially trained HMM synthesis voice. The data we used in this study are from the Norwegian Språkbanken (see Section Speech Databases).

Data description

The manuscripts for the recordings were based on the NST corpus, and the selection was done to make them phonetically balanced and to ensure diphone coverage. Though not prosodically balanced, the manuscripts still contain different types of sentences that ensure prosodic variation, e.g. statements, wh-questions, yes/no questions and enumerations. The 11 hour speech database was aligned on the phonetic and word levels using our Nalign software (Sjölander and Heldner, 2004) with the NST dictionary as pronunciation dictionary. This comprises more than 900.000 phonetically transcribed items with syllable boundaries marked. In addition, the text was tagged for part-of-speech using a TNT tagger trained on the SUC corpus (Megyesi, 2002). From the NST database for training of speech recognition we selected a small number of unprofessional speakers from the following Swedish dialectal areas: North, Dala, Göta, Gotland and South (see Figure 1). The data samples were considerably smaller than the speech synthesis database; they ranged from 22 to 60 minutes, compared to the 11 hours by the professional speaker.

HMM contextual features

The typical HMM synthesis model can be decomposed into a number of distinct layers. At the acoustic level, a parametric source-filter model (MLSA-vocoder) is responsible for signal generation. Context dependent HMMs, containing probability distributions for the parameters and their 1st and 2nd order derivatives, are used for generation of control parameter trajectories. In order to select context dependent HMMs, a decision tree that uses input from a large feature set to cluster the HMM models was applied.

In this study, we used the standard model for acoustic and HMM level processing, and we focussed on adapting the feature set for the decision tree for the task of modeling dialectal variation. The feature set typically used in HMM synthesis includes features on segment, syllable, word, phrase and utterance level. Segment level features include immediate context and position in syllable; syllable features include stress and

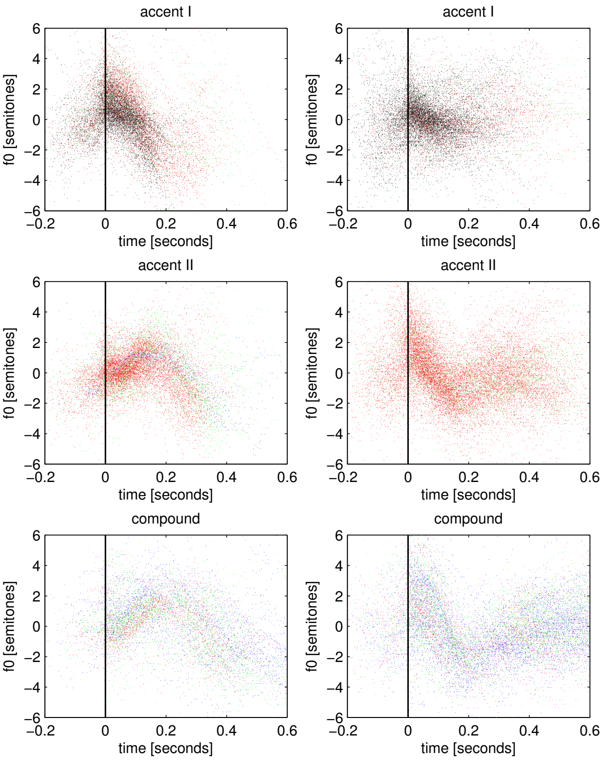


Figure 5: Pitch clouds for South *(left) and* Svea *(right) dialects.*

position in word and phrase; word features include emphasis, part-of-speech tag (content or function word), number of syllables, position in phrase etc., phrase features include phrase length in terms of syllables and words; utterance level includes length in syllables, words and phrases. For our present experiments, we have also added a speaker level to the feature set, since we train a voice on multiple speakers. The only feature in this category at present is dialect group, which is one of North, Dala, Svea, Göta, Gotland and South. In addition to this, we have chosen to add to the word level a morphological feature stating whether or not the word is a compound, since compound stress pattern is often a significant dialectal feature in Swedish (Bruce et al., 2007). At the syllable level we have added explicit information about lexical accent type (accent 1, accent 2 or compound accent).

Data-driven exploration of intonation

The content-rich feature files generated from large annotated speech corpora, used in HMM synthesis training, also allow for statisical and explorative investigation of prosodic characteristics of different speakers and dialects. Figure 5 visualises F_0-patterns for two speakers with different dialects as *pitch clouds*.

We selected approximately 1000 content words, ranging from 1 to 5 syllables, with primary stress on the first syllable, from a large set of read utterances. F_0-curves were extracted, mean-normalised and temporally aligned according to vowel onset in the stressed syllable (marked with a vertical line in the figure). For each dialect, separate clouds were generated for three accent types: accent 1, accent 2 and compounds. The figure clearly shows the dialect difference in accent 2 and compounds, with two peaks in the *Svea* case and a single peak for *South*. For *South* it is clear that the temporal alignment of the peak is later in accent 2 than in accent 1. An additional dimension in the figure is syllable length, which is represented by color. Monosyllabic words are black, 2-5 syllable words are red, green, blue and magenta respectively. Not unexpectedly, there is an overrepresentation of monosyllabic accent 1 words, since we selected only those with stress on the first syllable. Accent 2 words are primarily disyllabic, while a majority of the longer words are compounds. This type of analysis gives insight into features that influence prosodic realisations, which is valuable both in HMM synthesis and for fine-tuning the SWING rules.

Synthesis of our approaches

The SWING tool requires information about phoneme alignment, pitch range, syllable stress and accents. These features are all automatically generated in the HMM synthesis process, which makes it possible to use SWING rules to generate pitch contours automatically from an emphasis-tagged text, which in turn can be used to replace or supplement the HMM-generated pitch curves prior to sound synthesis.

Current work in our project concerns using the rules obtained with SWING to generate intonation for the seven main regional varieties of Swedish together with the HMM synthesiser. We have integrated SWING into the HMM synthesis framework, so that it may be driven by the same input features as the HMM synthesis, and the durations generated by the HMMs trained on a speaker of the target dialect. This makes it possible to replace or adapt the HMM-generated F_0 track by the one generated by SWING before rendering the waveform. As an example of how the new hybrid SWING/HMM synthesiser works, Figure 6 shows the F_0 tracks generated by the two systems for South Swedish SWING.

The new hybrid HMM/SWING synthesiser will allow more careful investigation of the SWING rules, since large sets of perceptual stimuli can be automatically generated under controlled conditions.

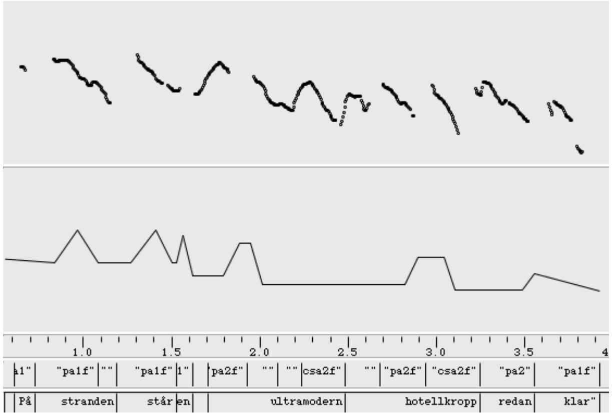


Figure 6: Log F_0 curves generated by HMM (top) and SWING *(bottom) for the* South *dialect.*

Discussion and future work

We are planning to run some perceptual testing to see whether listeners will be able to reliably identify a speaker as belonging to the different regional varieties of Swedish depending on the specific pitch shapes of target utterances. We will also perform listening tests comparing different methods of merging the data-driven and rule-generated F_0 tracks in the hybrid HMM/SWING synthesiser.

Acknowledgement

This work is supported by a grant from the Swedish Research Council.

References

Boersma P and Weenink D (2010). Praat: doing phonetics by computer (version 4.6.17) [computer program]. Webpage http://www.praat.org/, visited 14-March-10.

Bruce G (2007). Components of a prosodic typology of Swedish intonation. In T Riad and C Gussenhoven, eds., *Tones and Tunes*, vol. 1, 113–146. Berlin: Mouton de Gruyter.

Bruce G and Gårding E (1978). A prosodic typology for Swedish dialects. In E Gårding, G Bruce and R Bannert, eds., *Nordic Prosody*, 219–228. Lund: Department of Linguistics.

Bruce G and Granström B (1993). Prosodic modelling in Swedish speech synthesis. a prosodic typology for Swedish dialects. *Speech Communication*, 63–73.

Bruce G, Granström B and Schötz S (2007). Simulating intonational varieties of Swedish. In *Proc. of ICPhS XVI, Saarbrücken, Germany*.

Elenius K (1999). Two Swedish SpeechDat databases - some experiences and results. In *Proc. of Eurospeech 99*, vol. 4, 2243–2246.

Elert C C (1994). Indelning och gränser inom området för den nu talade svenskan - en aktuell dialektografi. In L Edlund, ed., *Kulturgränser - myt eller verklighet.*, vol. 1, 215–228. Umeå, Sweden: Diabas.

Lundgren A (2005). *HMM-baserad talsyntes*. Master's thesis, KTH, TMH, CTT.

Megyesi B (2002). *Data-Driven Syntactic Analysis - Methods and Applications for Swedish*. Ph.D. thesis, KTH, Department of Speech, Music and Hearing, KTH, Stockholm.

Segerup M (2005). The interaction of word accent and quantity in Gothenburg Swedish. In *Proc. of the XVIIIth Swedish Phonetics Conference, Fonetik 2005*. Department of Linguistics, Göteborg University.

Sjölander K and Heldner M (2004). Word level precision of the NALIGN automatic segmentation algorithm. In *Proc. of The XVIIth Swedish Phonetics Conference, Fonetik 2004*, 116–119. Stockholm University.

Taylor P (2009). *Text-To-Speech Synthesis*. Cambridge University Press.

Tokuda K, Yoshimura T, Masuko T, Kobayashi T and Kitamura T (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *Proc. of CASSP 2000*, 1315–1318.

Watts O, Yamagishi J, Berkling K and King S (2008). Hmm-based synthesis of child speech. In *Proc. of The 1st Workshop on Child, Computer and Interaction*.

English *of* in L1 and L2 speakers' read and spontaneous speech

Helena Spilková and Wim A. van Dommelen
Department of Language and Communication Studies, NTNU

Abstract

This paper compares realizations of the English function word of *in read and spontaneous speech produced by Czech and Norwegian speakers with native productions. Acoustic analysis involved word and segment durations, voicing, formant and band energy measurements. Non-natives appeared to produce longer durations than natives but read productions were longer than spontaneous tokens for natives and non-natives alike. Relative segment durations, however, varied between speaker groups. F1-F0 values in* of *did not differ between groups. Higher F3-F2 values indicated less fronted vowel quality for the non-natives than for the natives. In this case there was no effect of speaking style. Amount of fricative voicing and friction intensity varied systematically with absence or presence of voicing in the following segment.*

Introduction

It is a well-known fact that spoken language is characterized by reduction processes making actual word tokens differ from their canonical forms. Reduction phenomena have been studied in e.g. Dutch (van Son and Pols 1999; Keune et al. 2005), English (Bell et al. 2003; Baker and Bradlow 2009), Greek (Dauer 1980), Japanese (Nakamura et al. 2008), Russian (Bolotova 2003) and Spanish (Harmegnies and Poch-Olivé 1992). One of the most important factors in reduction of word forms is speaking style. Usually, investigations compare more formal presentations (often read speech) with less formal speech (often called spontaneous speech). Although *spontaneous speech* might seem to be appropriate as a cover term, it should be kept in mind that in a number of studies speaking strategies have been shown to be individually different. For example, different approaches to speaking styles were found for the two Dutch speakers in Laan (1997), five speakers of Greek in Dauer (1980), five English subjects in Lavoie (2002), and seven Dutch, six Finnish and five Russian speakers in de Silva et al. (2003).

For second language (L2) speakers it may be a challenge to produce appropriately reduced forms. During the initial period of second language instruction normally attention will be paid to the L2 sound system, in particular to the sounds that are different from the learner's native language (L1). Naturally, much emphasis will be put on the acquisition of specific sounds as they appear in canonical word forms. In a more advanced stage, efforts may be devoted to mastering language-specific reduction rules. Depending on factors like frequency and predictability of a word, its position in an utterance, speaking style, etc., native speakers' realizations may show varying degrees of reduction (cf. Lavoie 2002). Previous investigations have produced evidence of L2 learners showing phrase-level effects that differ from what is found for L1 speakers. Native speakers of Spanish in Lowenstein Mairs (1989) produced incorrect patterns of stress assignment in English. Further, Wenk (1985) found improper reduction of vowel quality in francophone English. In a study by Bond and Fokes (1985), native speakers of Thai, Malaysian and Japanese were shown to have insufficient awareness of typical English patterns of word compression due to addition of syllable suffixes. In Gut (2007) English learners of German demonstrated insufficient degrees of vowel reduction. The same was true for German learners of English.

The goal of the present study was to investigate reduction phenomena in the English function word *of* occurring in (quasi-) spontaneous vs. read speech produced by native as well as non-native speakers. The spontaneous material was collected from different types of task-elicited dialogues recorded in studios (see Method section). Non-native speakers came from two typologically different languages, Norwegian and Czech (cf. Swan and Smith 2001). It was hypothesized that in general non-native speakers of English would show less reduced productions than natives. Further, native speakers were expected to exhibit stronger reduction effects in spontaneous vs. read speech than L2 users. Finally, it was postulated that the larger typological distance between English and Czech would cause less native-like productions for Czech than for Norwegian speakers.

Method

Speech material

The material used in this study was obtained from several sources. The read material is represented by recordings of non-professional speakers reading transcripts of BBC news texts. Part of the BBC news recordings was recorded in Trondheim and part was provided by the Institute of Phonetics, Charles University in Prague. The spontaneous material consists of spontaneous dialogues in English, elicited using a picture replication task (part of the Kachna corpus; Spilková et al. 2010) and dialogues elicited using a Map Task (White et al. 2010). All the recordings were made in studio environments with a sampling rate of 32 kHz or higher and 16-bit quantization, using a separate channel for each speaker (in dialogues).

The lexical item chosen for analysis is the English preposition *of*. For both types of material, we aimed to select realizations of this word fluently and naturally integrated in surrounding speech, therefore we excluded all cases where a pause, hesitation or another type of disfluency was present in close proximity of the observed word. Attention was also paid to the context and syntactic status of the observed word, where we avoided, e.g., clause-final use of prepositions (so-called stranded preposition) and strongly lexicalized phrases where a disproportional reduction could be expected. Five tokens per speaker and speaking style were selected (incidentally less than five for a few speakers with a limited number of suitable items).

Speakers

The groups of subjects consisted of ten Norwegian speakers (4 females, 6 males), ten Czech speakers (5 females, 5 males) and two native British English speakers (females) that were recorded in both speaking styles (reading and replication task dialogue). In addition, BBC news recordings of three British speakers (males) and recordings of Map Task dialogue of six British speakers (2 females, 4 males) were used. The age of the speakers ranged from 19 to 45 years, and most of the speakers were university students. The speaker pairs in the dialogues were in most cases formed by either classmates or colleagues.

In Norway, the well-established system of English instruction and high exposure to English language (e.g. most movies in English are not dubbed) result in an overall high competence in English in young population. The speakers were therefore selected from university students which guaranteed sufficient proficiency. In Czech Republic, however, such a proficiency standard cannot be generally expected and we had to select speakers from more carefully chosen groups, namely university students of English, and employees in a company using English as the official work language. The dialects of the native English speakers mostly belonged to the Southern English dialect group; one speaker spoke a Northern English dialect.

Acoustic analysis

The selected items were segmented using Praat (Boersma and Weenink 2009). Segment durations were obtained for the vowel, the fricative and the portion with phonetic voicing in the fricative. Formant values in Bark were measured as means of values obtained from the whole duration of the vowel in the observed item. To be able to eliminate relatively frequent errors in automatic formant tracking, an additional semi-automatic method was used to detect any abrupt jumps between nearby formant measurements. The resulting formant values were used to calculate F1-F0 and F3-F2 values in Bark that correspond to vowel height and backness while reducing anatomical variation (corresponding to, e.g., gender; cf. Syrdal and Gopal 1986; Adank et al. 2004). The value of F0 necessary for this transformation was measured in the centre of the vowel interval, avoiding the portions with a creaky voice quality where possible. Furthermore, band energies (low band: 0 - 5000 Hz, high band: over 5000 Hz) were also measured in the fricative. These were used to calculate *high-frequency band – low-frequency band* differences in dB, corresponding to relative friction intensity (that represents the fortis character of the sound).

Results

Segment durations

In this section the results of measurements of segmental durations will be presented. As can be seen from Figure 1, total word durations were longer in read than in spontaneous speech (pooled across the three groups of speakers 123ms vs. 105ms). In addition, word durations were longer for both Czech (133ms) and Norwegian speakers (109ms) compared to natives (91ms). An analysis of variance with speaking style and language as factors revealed that whereas the effects of speaking style as well as language were significant (F(1, 255)= 6.610; p= 0.011 and F(2, 255)= 12.601; p< 0.001), the speaking style x language interaction did not reach significance (F(2, 255)= 1.057; p= 0.349). Bonferroni-adjusted paired comparisons showed that only Czech speakers had significantly longer word durations than natives. Czech word durations were also reliably longer than those produced by Norwegians.

Further, we investigated whether V/C ratios differed across speaking styles and language groups. The data showed that vowel duration in the word *of* in read speech amounted to 50% of the total word duration vs. 55% in spontaneous speech. For the English speakers the corresponding percentages were 47% vs. 55%. While for Czech speakers similar ratios were observed (43% vs. 55%), Norwegian speakers showed an opposite pattern (59% vs. 56%). According to an analysis of variance the effects of the factors speaking style (F(1, 255)= 6.253; p= 0.013), language (F(2, 255)= 6.585; p= 0.002) as well as their interaction (F(2, 255)= 5.396; p= 0.005) were significant.

Closer inspection of the data revealed that speaking style also affected the degree to which the fricative was filled with voicing (69% in read speech vs. 84% in spontaneous speech). For the English and, especially, the Czech speakers this result can be explained by relatively long fricative durations in read speech (read vs. spontaneous: Czech 84ms vs. 56ms; English 58ms vs. 40ms). Norwegian speakers had similar fricative durations for these two conditions (both 48ms). Pooled across the two speaking styles, Czech and English subjects had comparable amounts of voicing (73% and 70%, respectively), while Norwegian speakers made the fricative more voiced (85%). Statistical analysis revealed that the effect of both speaking style and language on the amount of voicing in the fricative was significant (F(1, 242)= 12.435; p= 0.001 and F(2, 242= 5.082; p= 0.007) with a significant speaking style x language interaction (F(2, 242)= 3.875; p= 0.022). Phonetic classification of immediately neighboring segments as voiceless/voiced revealed that the amount of voicing in the fricative correlated with the voicing status of the following segment. This issue will be dealt with in the section Context effects below.

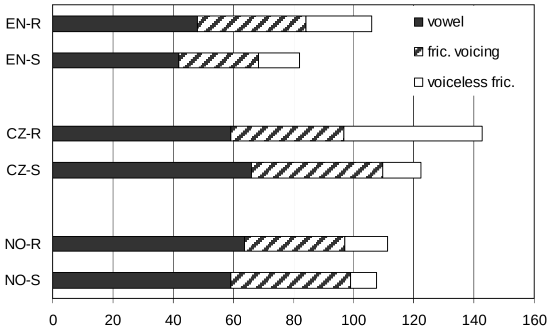


Figure 1. Segment durations in ms in of *spoken by English (EN), Czech (CZ) and Norwegian (NO) speakers. R= read, S= spontaneous; fric. voicing= voiced part, voiceless fric.= voiceless part of fricative.*

Vowel quality

In the following we will investigate whether the quality of the vowel in *of* differed systematically due to speaking style and between natives and non-natives. F1-F0 and F3-F2 values for different conditions are presented in Figures 2 and 3. For F1-F0 an analysis of variance with the factors speaking style and language showed no significant differences at all (for both factors: F< 1). Only Norwegian subjects' F1-F0 values were larger for read than spontaneous speech (3.3 Bark vs. 2.9 Bark; t(92)= 2.168; p= 0.033).

In contrast, for the F3-F2 measure a significant effect of language was found (F(2, 252)= 24.416; p< 0.001). Bonferroni-adjusted paired comparisons showed that across speaking styles F3-F2 values for Czech (4.6 Bark) as well as Norwegian (4.5 Bark) were different from English (3.6 Bark). This indicates more peripheral vowel qualities for the non-native speakers. Both speaking style (read vs. spontaneous: 4.4 Bark vs. 4.3 Bark) and its interaction with the factor language, however, did not reach statistical significance (F< 1 and F(2, 252)= 1.668; p= 0.191, respectively).

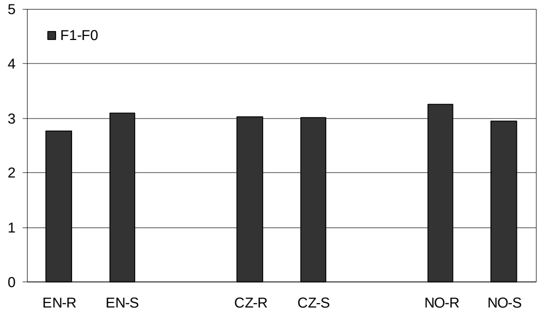


Figure 2. F1-F0 distances in Bark in of. *Legend: see Figure 1.*

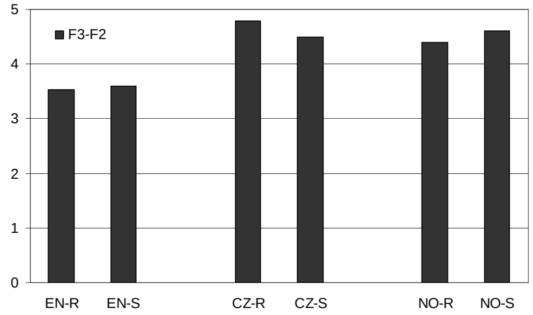


Figure 3. F3-F2 distances in Bark in of. *Legend: see Figure 1.*

Context effects

Fricative voicing

It was speculated that apart from the factors speaking style and speakers' language background, phonetic context may affect realization of the fricative in *of.* To study the effect of context, the immediately neighbouring segments were classified as (phonetically) voiceless or voiced. First we will look into contextual effects on segment durations. As was shown by an analysis of variance with factors left context and right context (in addition to speaking style and language), absence vs. presence of voicing in neither left nor right context did affect total word duration. Also, no significant effect on relative vowel duration was found. In contrast, absence vs. presence of voicing in the segment following *of* had a strong influence on the amount of voicing in the fricative. Across speaking styles and language groups, voicing percentages amounted to 87% before a voiced segment vs. 54% before a voiceless segment (see Figure 4). According to an analysis of variance with factors speaking style, language, left context and right context this effect was highly significant (F(1, 224)= 56.062; p< 0.001). Left context had no significant effect at all (F< 1). The only significant interaction was found between language and speaking style (F(2, 224)= 5.316; p= 0.006). Presumably due to small number of observations in some categories, the main effects of speaking style and language did not reach statistical significance any longer (cf. section Segment durations above).

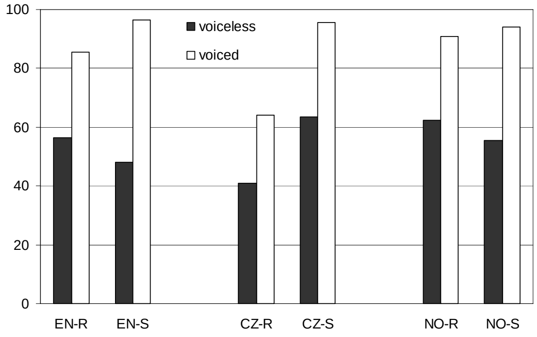


Figure 4. Amount of voicing in the fricative in of *(in %) followed by a voiceless/voiced segment. Legend: see Figure 1.*

Fricative intensity

One of the acoustic correlates of the status of a fricative as fortis vs. lenis is the intensity of the friction. Through a similar analysis as for fricative voicing, in this section we investigate whether the absence vs. presence of voicing in the neighbouring segments also affected friction intensity. The measure used was the difference in spectral energy above-below 5000Hz in the fricative (see Method section, Acoustic analysis).

As can be seen from the results presented in Figure 5, intensity of the friction varies with voicing status of the following segment. When followed by a voiceless segment, the amount of spectral energy above 5000Hz is considerably larger than preceding a voiced segment. An analysis of variance with factors speaking style, language, left context and right context showed that the latter effect was highly significant (F(1, 224)= 52.742; p< 0.001). The same was true for the main effects speaking style and language (F(1, 224)= 7.856; p= 0.006 and F(2, 224)= 3.887; p= 0.022) but not for left context (F(1, 224)=1.04; p= 0.295. The only significant interaction was between speaking style and language (F(2, 224)= 5.975; p= 0.003).

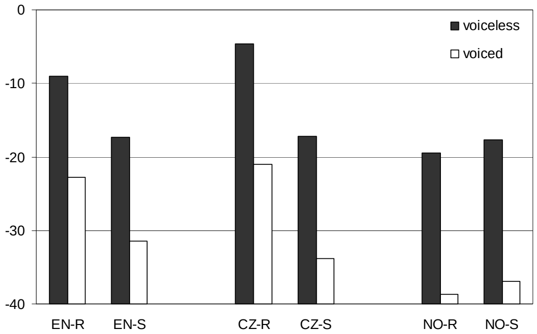


Figure 5. Difference in spectral energy above-below 5kHz in the fricative in of *(in dB) followed by a voiceless/voiced segment. Legend: see Figure 1.*

Discussion

The combined results of the present acoustic measurements on read and spontaneous *of* tokens produced by L2 speakers revealed partly consistent tendencies on the one hand and partly diverging reduction patterns on the other. In the temporal domain there was a general tendency to longer word durations for the non-native speakers than for the natives (though only significant for the Czech). This result is in line with Gut's (2007) observations of longer syllable durations in non-native English spoken by Germans vs. native English and *vice versa* in non-native German produced by English speakers. Also, White and Mattys (2007) found speech rate to be slower for non-natives compared to natives in Spanish sentences read by English speakers as well as English sentences read by Spanish speakers. In the study by Mackay and Flege (2004) early and late Italian–English bilinguals repeated matched English and Italian sentences following an aural model. The late bilinguals produced longer English than Italian sentences. This result was interpreted as due to speakers' need for resources to suppress their Italian subsystem. A similar interpretation might be valid for the present results.

Also in the spectral domain a general L1-L2 effect was observed. As evidenced by the F3-F2 measure, the Czech as well as Norwegian subjects produced both read and spontaneous *of* with a less reduced vowel than the natives. This result is in congruence with Flege, Bohn and Jang (1997) who observed inappropriate spectral contrasts in English vowel pairs produced by German, Mandarin, Spanish and Korean speakers. In Wenk (1985) French L2 users of English showed insufficient reduction of vowels in pre-tonic syllables. Lowenstein Mairs (1989) reports inappropriate stress assignment by Spanish speakers of English, which presumably caused insufficient vowel reduction (neither her nor Wenk's study involved acoustic measurements).

At odds with the general differences between L1 and L2 performance was the absence of significant interactions for language and speaking style. Word durations were shorter in spontaneous than in read speech for natives and non-natives alike. Further, no vowel quality reduction due to speaking style was found for L1 as well as L2 speakers. For the English speakers this might be due to high degree of vowel quality reduction present already in their read tokens.

Also the effects of context on *of* realizations were similar for the native and non-native speakers. For all three groups, absence vs. presence of phonetic voicing in the segment following *of* affected voicing during the fricative and also its intensity. It can be speculated that we are dealing with assimilation processes that are possibly similar for the three languages (cf. Hall 2003 for Czech; Jansen 2007 for English; not many studies seem to exist for Norwegian; cf. Kristoffersen 2000). More research on assimilation across word boundaries is needed to answer this question.

Further, internal syllable structure (i.e., V/C ratio) differed for natives vs. Norwegian but not Czech speakers. Only the Norwegians had a larger percentage of voicing in the fricative which was explained by their relatively short fricative durations.

The lack of consistent results might at least partly be explained by individually different reduction strategies. It might be speculated that in the present study idiosyncratic behaviour to some degree outweighed L2-specific influences.

Acknowledgements

The authors were supported by the EU Marie Curie Research Training Network Sound to Sense. We would like to thank Radek Skarnitzl for providing the read Czech material and to Laurence White for making available the recordings of native English speakers' Map Task.

References

Adank P, Smits R, van Hout R (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America,* 116: 3099-3107.

Baker R E, Bradlow A R (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech,* 52: 391-413.

Bell A, Daniel J, Fosler-Lussier E, Girand C, Gregory M, Gildea D (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America,* 113: 1001-1024.

Boersma P, Weenink D (2009). Praat: Doing phonetics by computer (Version 5.1.06) [Computer program]. Retrieved May 11, 2009, from http://www.praat.org/.

Bolotova O (2003). On some acoustic features of spontaneous speech and reading in Russian (quantitative and qualitative comparison methods). *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 3-9 August 2003 (1): 913-916.

Bond Z S, Fokes J (1985). Non-native patterns of English syllable timing. *Journal of Phonetics*, 13: 407-420.

Dauer R M (1980). The reduction of unstressed high vowel in modern Greek. *Journal of the International Phonetic Association,* 10(1): 17-27.

de Silva V, Iivonen A, Bondarko L V, Pols L C W (2003). Common and language dependent phonetic differences between read and spontaneous speech in Russian, Finnish and Dutch. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 3-9 August 2003 (3): 2977-2980.

Flege J E, Bohn O-S, Jang S (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25: 437-470.

Gut U (2007). Learner corpora in second language prosody research. In: Trouvain J, Gut U, eds, *Non-native prosody: Phonetic description and teaching practice.* Berlin/New York: Mouton de Gruyter, 145-167.

Hall D C (2003). Laryngeal feature specifications in West Slavic languages. *Toronto Working Papers in Linguistics,* 20: 93-114.

Harmegnies B, Poch-Olivé D (1992). A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. *Speech Communication,* 11: 429-437.

Jansen W (2007). Phonological 'voicing', phonetic voicing, and assimilation in English. *Language Sciences,* 29: 270-293.

Keune K, Ernestus M, Van Hout R, Baayen, R H (2005). Variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory*, 1: 183-223.

Kristoffersen G (2000). *The phonology of Norwegian*. Oxford: Oxford University Press.

Laan G P M (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication,* 22: 43-65.

Lavoie L (2002). Some influences on the realization of for and four in American English. *Journal of the International Phonetic Association*, 32(2): 175-202.

Lowenstein Mairs J (1989). Stress assignment in interlanguage phonology: an analysis of the stress system of Spanish speakers learning English. In: Gass S M, Schachter J, eds, *Linguistic perspectives on second language acquisition.* Cambridge: Cambridge university Press, 260-283.

Mackay I R A, Flege J E (2004). Effects of the age of second language learning on the duration of first and second language sentences: The role of suppression. *Applied Psycholinguistics*, 25: 373–396.

Nakamura M, Iwano K, Furui S (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language,* 22: 171-184.

Spilková H, Brenner D, Öttl A, Vondřička P, van Dommelen W A, Ernestus M (2010). The Kachna L1/L2 Picture Replication Corpus. *International Conference on Language Resources and Evaluation*, 2010, Valletta. (to appear)

Syrdal A K, Gopal H S (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, 79: 1086-1100.

Swan M, Smith B, eds, (2001). *Learner English: a teacher's guide to interference and other problems.* (2nd ed.) Cambridge: Cambridge University Press.

van Son R J J H, Pols L C W (1999). Acoustic description of consonant reduction. *Speech Communication,* 28: 125-140.

Wenk B J (1985). Speech rhythms in second language acquisition. *Language and Speech,* 28: 157-175.

White L, Mattys S L (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35: 501–522.

White L, Wiget L, Rauch O, Mattys S L (2010). Segmentation cues in spontaneous and read speech. *Proceedings of the 5th Conference on Speech Prosody*, 2010, Chicago. (to appear)

How widespread is preaspiration in Italy? A preliminary acoustic phonetic overview

Mary Stevens
Institut für Phonetik und Sprachverarbeitung (IPS), Ludwig-Maximilians Universität, München

Abstract

Preaspiration is a comparatively rare phonetic feature, almost entirely confined to languages spoken in far northwestern Europe (Andersen 2002; Helgason 2002). A recent acoustic phonetic investigation into consonant gemination in Sienese Italian (e.g. Stevens & Hajek 2007), however, found that preaspiration occurred in one third of geminate /pp tt kk/ tokens. Up until that point preaspirated stops had not been reported to occur in Sienese or any other variety of Italian including the standard language. With this in mind, the present paper presents the results of an acoustic phonetic investigation into voiceless geminate stops /pp tt kk/ in a controlled corpus of words read by speakers from 15 other Italian cities. Results are analysed according to city as well as factors (e.g. speaker sex, vowel type) known to favour occurrences of preaspiration in other better known preaspirating languages. Preliminary duration values are presented and results are discussed in terms of two specific hypotheses regarding the rise of preaspiration in Sienese Italian.

Introduction & background

Geminate consonants in Italian

Consonant length is contrastive in (standard) Italian e.g. *pala* 'shovel' v. *palla* 'ball'. The duration of the consonant, that of the preceding vowel, and the ratio between them (C/V) all play a role in signalling phonological consonant length (e.g. Pickett et al. 1999).

Geminate voiceless stops /pp tt kk/, in particular, are described as unaspirated in all prosodic positions in standard Italian (e.g. Bertinetto & Loporcaro 2005). Manner of articulation has not been a focus of existing experimental phonetic studies on geminates in standard Italian (e.g. Pickett et al. 1999), however post-aspiration of geminate voiceless stops is reported to occur in Calabria (Sorianello 1996). Degemination is a feature of northern varieties and the geminates of standard Italian as pronounced by these speakers are often less salient (Bertinetto & Loporcaro 2005).

Standard Italian and regional varieties

The linguistic situation in Italy is complex and the standard language, while historically derived from Florentine Italian, is today perhaps best described as an abstraction only used, and increasingly less so, by trained speakers (Bertinetto & Loporcaro 2005). The pronunciation of standard Italian is influenced by distinct regional accents, which can be very broadly categorized into those of the north, centre (Tuscany) and south.

The present study involves a list of standard Italian words read by speakers from 15 cities spread over these three regions. The controlled circumstances mean less regional influence on speakers' pronunciation than would occur in a less formal setting. However, such data allow us to test whether preaspiration is a regional feature of Sienese or whether it occurs more widely and in standard Italian, as spoken across Italy.

Preaspiration

Preaspiration involves a breathy h-like sound before the closure period for voiceless stops i.e. [hp ht hk], which results from early glottal abduction relative to the supralaryngeal closure gesture. The term preaspiration is often taken as a cover term for a range of finegrained realisations including 'archetypal' [h], breathy voice [ɦ] and preglottalization (e.g. Stevens & Hajek 2007). Preaspiration, like glottal /h/ more generally, is hard for listeners to hear, which may explain why it could go unnoticed or at least unreported, despite occurring relatively frequently in Sienese Italian speech.

Preaspiration in Sienese Italian

An earlier investigation found that third of /pp tt kk/ are realised with preaspiration in spontaneous Sienese speech (e.g. Stevens & Hajek 2004, 2007). While the frequency and distribution of preaspiration are described in detail elsewhere, we note here that preaspirated stops occurred for all 6 speakers in the corpus (albeit more frequently for some than others) and was also found to interact with consonant place (most frequent for velar /k:/) and vowel type (most frequent following low central /a/), but not stress.

In terms of why preaspiration would occur in this variety, here we consider two potential hypotheses. (1) Preaspiration could be related to the *Gorgia toscana* (GT), which weakens intervocalic singleton /p t k/ to voiceless fricatives including glottal [h] in Sienese and some other Tuscan varieties (e.g. Villafaña-Dalcher 2008). With an articulatory similarity to preaspiration (also glottal [h]), GT distinguishes these Tuscan varieties from those spoken elsewhere in Italy, including the standard language. This may explain why preaspiration was found in Siena, but not, to date, elsewhere.

Alternatively, (2): in contrast to other experimental phonetic studies on Italian (e.g. Pickett et al. 1999), the investigation into Sienese was based on spontaneous speech. Keeping in mind the partial debuccalization involved (e.g. [k:] > [hk]), preaspiration may be a kind of weaker, hypoarticulated form of plain unaspirated [pp tt kk] that would typically, and perhaps only, occur in less formal, fast and/or careless speech. However, such an interpretation is not supported cross-linguistically: preaspiration occurs in both controlled and spontaneous Swedish speech (Helgason & Ringen 2008).

Under either hypothesis, recall from above that preaspiration is hard for listeners to hear, even when especially listened for. Therefore it may only have come to light in Sienese Italian because of the detail that more recently developed acoustic phonetic methods allow.

Aims

This paper aims to investigate why preaspiration would occur relatively frequently in spontaneous Sienese Italian speech, and yet had not previously been reported to occur in this or any other variety of Italian. We aim to test the two specific hypotheses suggested above, i.e. whether preaspiration in Sienese may be explained by either (1) the presence of GT in that variety, or (2) the use of spontaneous speech data in the earlier investigation into Sienese Italian. This paper tests both hypotheses by investigating the realisation of voiceless geminate /pp tt kk/ in controlled standard Italian - in which GT is not present. On this point, however, recall from that it is difficult to find examples of standard Italian unmarked by regional influences - such as GT weakening of /p t k/ for speakers native to Siena, Florence and some other parts of Tuscany. Therefore, while the data are essentially standard Italian, speakers' regional origin is also taken into account.

Data & methods

The data were drawn from part of the CLIPS corpus of spoken Italian. Only words read in isolation were analysed - at least for the purposes of this preliminary overview, in order to have a balanced set of data across geographical locations. Sociocultural background and age of the speakers was controlled in the original corpus (see Sobrero 2006 for details). For the present study only word-medial voiceless geminate stops /pp tt kk/ in the *Vignetta A* word list were analysed (the specific words analysed here were *bocca* 'mouth', *bottoni* 'buttons', *cappello* 'hat', *macchina* 'car', *occhi* 'eyes', *specchietto* 'little mirror', and *tetto* 'roof', giving 8 /C:/ tokens). Each word was read once by eight speakers for each variety. Some tokens were eliminated due to background noise, leaving 935 tokens.

These individual words were extracted and for each /VC:/ sequence, the duration of the preceding vowel, the closure period, and the release (comprising the burst and any postaspiration) were manually measured in Praat. These labels are shown in Figure 1.

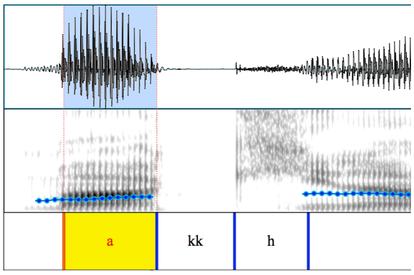


Figure 1. A voiceless velar stop /kk/ in macchina *'car' realised without preaspiration by a male speaker from Florence (LTp1A01F).*

In addition, preaspiration was also labelled separately as hC when it preceded the closure portion, as shown in Figure 2.

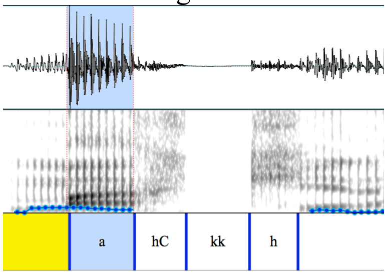


Figure 2. Macchina *produced by a male speaker from Turin (LTp1A04T), with preaspiration (hC) in addition to the three labels seen in Figure 1.*

Preaspiration was measured from the offset of modal voicing in the vowel (consistent with e.g. Helgason & Ringen 2008). This criterion meant that tokens realised with visible breathy voice in the vowel offset were included as preaspirated tokens. For example the /ak:/ sequence in Figure 3 closely resembles that in Figure 2, above, except that here the pitch trace (i.e. voicing) continues throughout the 'preaspiration' portion.

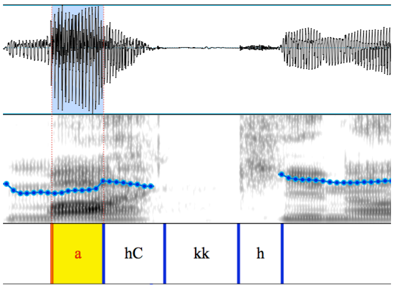


Figure 3. Macchina *produced by a female speaker from Rome (LTp1A04R), with breathy voice (labelled as hC) preceding velar closure.*

Finally, some /pp tt kk/ tokens were realised with preglottalization, i.e. with adducted rather than abducted vocal folds, as in Figure 3, below. These were counted as preaspirated alongside those of the type seen in Figures 2 and 3, primarily because they both differ from expected plain [p: t: k:].

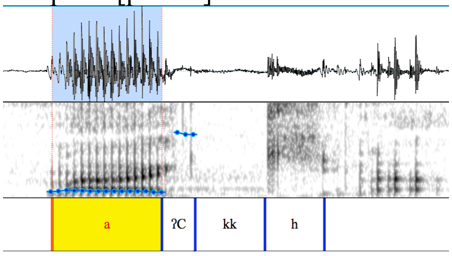


Figure 4. Macchina *produced by a male speaker from Lecce (LTp1A04L), with preglottalization. Creak is also clearly visible on the final vowel.*

Tokens were classified into two groups according to the presence or absence of preaspiration (including preglottalization and breathy voice), and analysed according to the speakers' native city, speaker sex, individual speaker; as well as vowel type, consonant place and lexical stress. Preliminary statistical tests were conducted on the duration values within Excel (single factor Anovas).

Results & discussion

Overall, preaspiration occurred in 269 of the total 935 /pp tt kk/ tokens analysed, i.e. 28.8%. Of these 269 preaspirated tokens, 17 were realised with glottal adduction, further described below. This overall frequency patterns very closely to the Sienese corpus (29.2%; Stevens & Hajek 2007). First, duration values for /VC:/ sequences with and without preaspiration are compared in Figure 5.

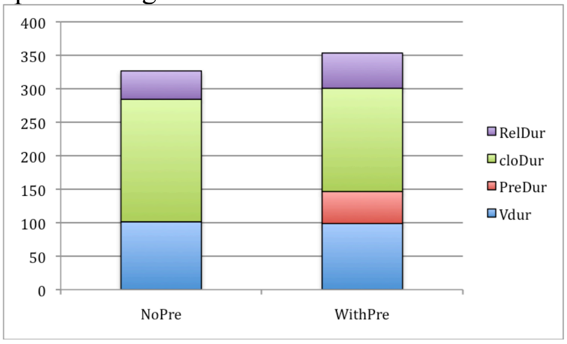


Figure 5. Average duration values (in ms.) for components of /VC:/ sequences with preaspiration (right; 269 tokens) and without (left; 666 tokens).

We can see that the overall duration of /VC:/ sequences is longer when preaspiration intervenes between the vowel and the closure portion. At 26.7ms. this is a statistically significant difference (*p* < 0.005). The duration of the preceding vowel remained relatively stable whether preaspiration occurred or not (*p* = 0.293). The average duration of the preaspiration portion at 47.9ms. is very similar to the value of 44ms. reported by Helgason & Ringen (2008) for preaspiration in isolated words read by Swedish speakers. These authors point to considerable interspeaker variability in terms of preaspiration duration which there is not space to investigate here, although note that preaspiration durations ranged from only 11ms. to 137ms. across the present corpus. The average closure duration was significantly shorter (-28.9ms.; *p* < 0.005) when preaspiration intervened in the /VC:/ sequence, whereas the release was significantly longer (+11ms; $p < 0.005$). Indeed, while /pp tt kk/ are always described as unaspirated in standard Italian, post-aspiration - discussed below - was frequent across the data set and can be seen in all four tokens in Figures 1-4. In terms of the neighbouring segments therefore, only closure duration and not vowel duration appears to interact with preaspiration. However while the shorter closure duration compensates somewhat for the intervening segment, when preaspiration is included as part of the consonant (i.e. preaspiration + closure + release), the preaspirated tokens are 29.4ms. longer, again a significant difference ($p < 0.005$). Given vowel duration remains stable, the overall /VC:/ sequence duration is also significantly (26.7ms.; $p < 0.005$) longer for the preaspirated stops.

Turning to frequency, Figure 6 shows the frequency of preaspiration according to each of the 15 cities included in the CLIPS corpus.

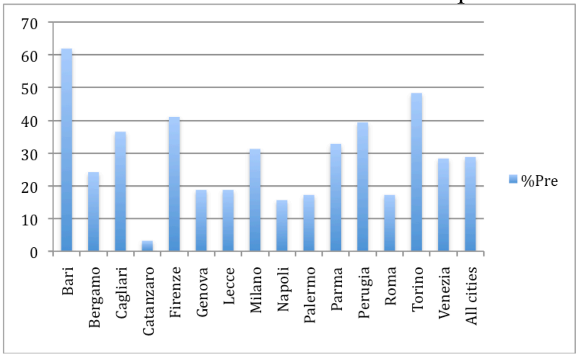


Figure 6. Frequency of preaspiration in the /pp tt kk/ tokens analysed for 15 Italian cities. Actual token numbers ranged between 56 and 64 for each city.

Values are shown as a percentage of the total tokens for each city, which were between 56 and 64 in each case. We can see that in every city in the corpus at least some of the geminate voiceless stops were produced with preaspiration. It is most frequent in Bari, occurring in just over 60% of the tokens analysed, whereas only 2 (3.2%) of the tokens from Catanzaro showed a preaspiration portion. Therefore there is no clear pattern in terms of the geographical distribution of preaspiration e.g. according to northern, central or southern cities: Bari (Puglia) and Catanzaro (Calabria), with the highest and lowest frequencies of preaspiration, respectively, are both in the far south of Italy. Moreover, given preaspiration was first reported in Siena, in central Italy, we might have expected it to be relatively more frequent for the central Italian cities in the present corpus. However again there is no clear pattern: preaspiration occurs in 40% of the tokens drawn from nearby cities of Florence and Perugia, i.e. well above the average of 28.8%, but only half as frequently for speakers from the other central variety spoken in Rome (17.2%). Figure 6 clearly shows that preaspiration is not confined to speakers from Florence, the only region with GT in the sample, so there appears to be no direct link between GT and the rise of preaspiration. Nonetheless, regional allophonic processes like GT may explain the way preaspiration is more frequent for speakers from some regions than others (most notably Catanzaro) in Figure 6. We return to this point below.

Turning to other potential influences on the frequency of preaspiration, it has been reported to be more frequent for female than male speakers (Helgason & Ringen 2008). However in these data there was little difference in the frequency of preaspiration according to speaker sex: preaspiration occurred in 132 of 440 tokens (30%) for females, and it was only slightly less frequent for male speakers at 27.7% of 495 tokens.

Tokens were then further broken down according to individual speaker. Figure 7 shows the percentage of speakers within each variety with at least one preaspirated /VC:/ token amongst the 8 analysed. Here the preglottalized tokens are shown separately, in order to determine whether individual speakers may prefer one gesture over the other.

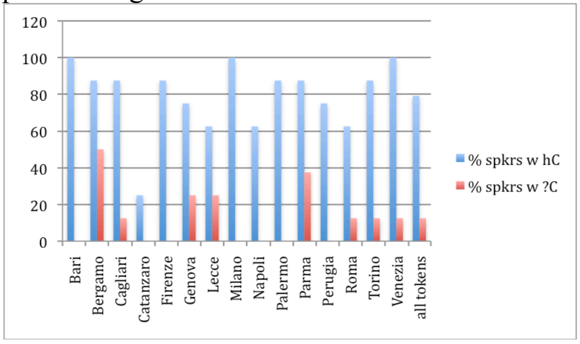


Figure 7. Percentage of speakers (8 for each city) with preaspiration, and preglottalization in /VC:/ sequences. 8 tokens per speaker.

Preaspiration is relatively evenly distributed across speakers in the corpus, inasmuch as for all cities except Catanzaro, half the speakers show at least one preaspirated token. While marginal in comparison to the frequency of preaspiration, it is worth noting the 17 preglottalized tokens were spread across 15 speakers in the corpus, each of whom also realised one or more /pp tt kk/ token with preaspiration. In other words, it does not appear that certain speakers favour either preglottalization or preaspiration.[1] Overall, Figures 6 and 7 show that preaspiration is a relatively widespread feature in Italy, albeit more occurring more frequently for some regions and for some individual speakers than others. In this sense Italian is very similar to Swedish, for which preaspiration is also a non-obligatory but relatively frequent feature (Helgason & Ringen 2008).

The impact of phonemic context on the frequency of preaspiration in the data was then examined, beginning with consonant place of articulation as shown in Figure 8.

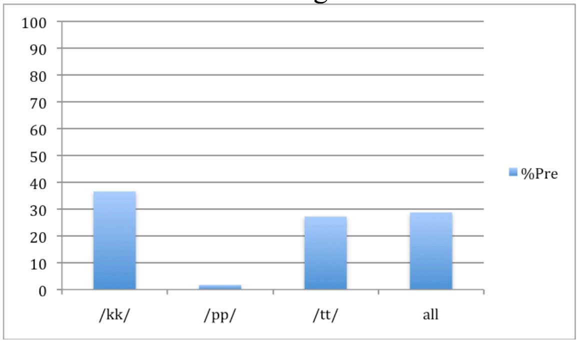


Figure 8. Frequency of preaspiration according to consonant place of articulation (as a percentage of total tokens for each C-place, 465 velar, 114 bilabial and 356 dental).

Preaspiration is most frequent for velar /k:/ (36.6%) and less frequent for dental /t:/ (27.2%), whereas only 1.75% of the /p:/ tokens analysed showed preaspiration. This overall pattern for consonant place (kk >> tt >> pp) is consistent with Sienese Italian spontaneous speech (Stevens & Hajek 2004).

The frequency of preaspiration was then examined according to vowel type in Figure 9 below. The sample is restricted because only three vowels /a e o/ preceded /p: t: k:/ in the data. Nonetheless, preaspiration is most frequent after low central /a/, a pattern that is consistent with the earlier corpus (Stevens & Hajek 2004).

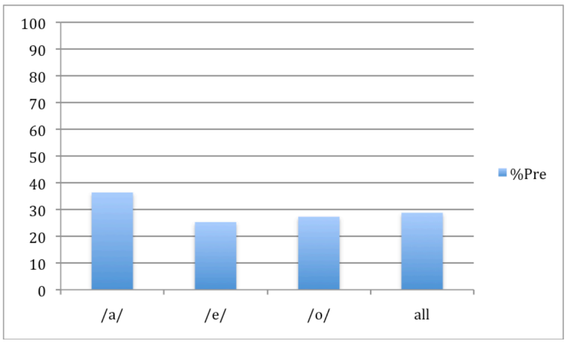


Figure 8. Frequency of preaspiration according to vowel type (values show the percentage of the 231 /a/, 356 /e/ and 348 /o/ tokens that were followed by a preaspiration).

Finally the frequency of preaspiration was examined according to whether the vowel in the /VC:/ sequence was stressed (e.g. <u>oc</u>*chi*) or not (e.g. *bot<u>to</u>ni*). Here we refer to lexical stress, although phonetic destressing was not an issue as the words were read in isolation, and lexical stress was realised phonetically in every case. Figure 9 shows the frequency of preaspiration in unstressed (*VCC) and stressed (‘VCC) sequences, as well as for all sequences combined.

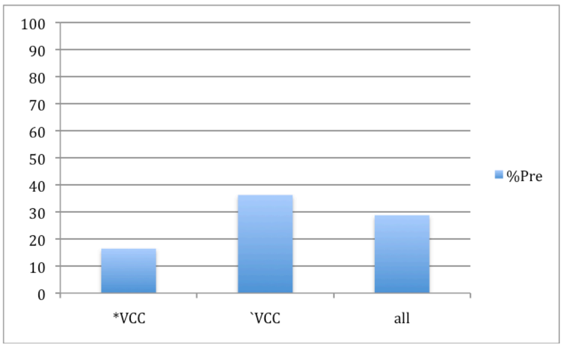


*Figure 9. Frequency of preaspiration in pre-tonic (*VCC) and post-tonic (‘VCC) positions.*

We can see that preaspiration was twice as frequent in post-tonic position (36.2%) as following unstressed vowels (16.4%). This is consistent with existing evidence for preaspiration in Swedish (Helgason & Ringen 2008) but not with evidence for Sienese Italian (Stevens & Hajek 2004). However that earlier study did not take phonetic destressing – prevalent in spontaneous speech – into account. The controlled speech data in Figure 9 are therefore more reliable and show a clear interaction with stress: preaspiration is more likely to occur following lexically stressed than unstressed vowels in standard Italian.

[1] Further research is needed into these categories and whether they should in fact be treated as part of the same overall phenomenon as they are here (indeed Andersen (2002) categorises preglottalization as weakening, but preaspiration as strengthening).

Further discussion

While not previously reported, results just presented show that preaspiration occurs in standard Italian, and with a similar frequency and distribution to patterns already documented for Sienese spontaneous speech (e.g. Stevens & Hajek 2004, 2007). Therefore preaspiration is not a specific regional feature of geminate production in Sienese Italian, but rather a more widespread feature in Italy, occurring as it does in controlled standard Italian in all regions sampled in the CLIPS corpus. Given these results, it is unclear why preaspiration has not been reported elsewhere for Italian.

The issue of more interest, however, is why preaspiration would arise in Italian, and what it may mean for sound patterns in Italy. There is not space here to adequately address these issues. Nontheless in terms of the 2 specific hypotheses raised earlier, results do not support a direct link with GT. Preaspiration occurred in the standard Italian of all 15 cities, and no more frequently for speakers from Florence, for whom GT is a regional feature, than in places without GT e.g. Bari or Turin. On the other hand while preaspiration appears to be a tendency across Italy, it is clearly more frequent in some regions than others, and perhaps it is regional allophonic processes that can explain these differences. The relative lack of preaspiration for speakers from Catanzaro, for example, is striking as Calabria is the only Italian region for which post-aspiration of /p: t: k:/ is reported (Sorianello 1996). At this point the interaction between pre- and post-aspiration remains unsolved, particularly as post-aspiration was in fact widespread in the present corpus of standard Italian.

Hypothesis 2 was also not supported by these results: just under one third of /pp tt kk/ tokens are preaspirated in the isolated words analysed here. Therefore preaspiration cannot be considered a kind of hypoarticulation associated with or even confined to spontaneous speech. On the contrary, duration values were longer for /VC:/ sequences with preaspiration. This could perhaps support its interpretation as a deliberate reinforcement of the geminate, although again, more detailed investigation is needed on this point.

Overall, while voiceless stops are described as unaspirated in standard Italian (Bertinetto & Loporcaro 2005) this study has shown that both pre- and post-aspiration are prevalent. This preliminary result suggests that [spread glottis] plays a role in phonological stop contrasts in Italian, although phonetic research into voiced /b: d: g:/ is necessary here, as well as perceptual tests with native listeners to determine what impact, if any, preaspiration has on these contrasts (i.e. singleton-geminate & voiced-voiceless). Now that preaspiration has been broadly documented here for standard Italian, further dialect specific research (e.g. involving more natural speech styles) is also clearly necessary in order to shed light on the possible interaction between regional allophonic processes and the rise and distribution of preaspiration in Italy.

Acknowledgments

This research was supported by the Alexander von Humboldt foundation through a postdoctoral research award. Help from Francesco Cutugno with access to the CLIPS corpus is greatly appreciated.

References

Andersen H (2002). Preglottalization in English and a North Germanic bifurcation. In D Restle & D Zaefferer eds, *Sounds and systems: studies in structure and change.* Berlin/New York, Mouton de Gruyter, 15-34.

Bertinetto & Loporcaro (2005) The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *JIPA* 35, 2: 131-151.

CLIPS *Corpora e Lessici di Italiano Parlato e Scritto*. http://www.clips.unina.it/it/index.jsp. Accessed 31 March 2010.

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

Pickett E, S Blumstein & M Burton (1999). Effects of speaking rate on the singleton/geminate contrast in Italian. *Phonetica* 56, 135-157.

Sobrero A (2006) Definizione delle caratteristiche generali del corpus: informatori, località. Available at: http://www.clips.unina.it/it/documenti.jsp.

Sorianello P (1996). Indici fonetici delle occlusive sorde nel Cosentino. *Rivista italiana di dialettologia* 20: 123-159.

Stevens M & J Hajek (2007) Towards a phonetic conspectus of preaspiration: acoustic evidence from Sienese Italian. *Proc. ICPhS XVI.* Saarbrücken, 429-432.

Stevens M & J Hajek (2004) Preaspiration in Sienese Italian & its interaction with stress in /VC:/ sequences. In: *Proc. Speech Prosody 2004*. Nara, 57-60.

Villafaña-Dalcher C (2008). Consonant weakening in Florentine Italian: a cross-disciplinary approach to gradient and variable sound change. *Language Variation and Change* 20, 2, 275-316.

Durations of phonologically long segments in native and foreign accented Swedish

Bosse Thorén
Akademin för humaniora och medier, Högskolan Dalarna

Abstract

Phonological features in a specific language could be expected to be reflected in realizations of second language speech. Swedish is known to have a quantity distinction involving duration in vowels, postvocalic consonants and vowel spectrum in stressed syllables.

Three speakers with respectively Swedish, Spanish and Estonian as their first language were recorded when telling a short story in Swedish. Durations in phonologically long segments were measured. The native Estonian speaker showed mostly longer durations than the other two, and the native Spanish speaker showed mostly shorter durations than the other two. The shorter durations of the Spanish speaker were expected, since Spanish does not have a quantity distinction, and it seems as though the Estonian speaker exaggerated the duration feature in her Swedish.

Introduction

The background of the present study is an attempt to evaluate the usefulness of a simplified prosodic strategy for teaching Swedish pronunciation to immigrants. The strategy focuses on “stress induced increase of segment duration” and is thoroughly described in Thorén (2008). The proper aim of the present study is to test the effect of manipulated durations in foreign accented Swedish and the present paper reports some preliminary production data from the recordings intended for manipulation.

Swedish is often described as having distinctive word stress and a vowel quantity distinction as main temporal prosodic contrasts. The vowel quantity distinction is known to involve vowel duration, spectral differences between long and short vowel allophones and complementary consonant length. E.g. Elert (1964) has shown that the quantity distinction is consequently associated with durational differences in both vowels and consonants. Fant & Kruckenberg (1994) concluded furthermore that duration is a main phonetic correlate of perceived stress in Swedish. The simplified pedagogic description ranks vowel and consonant lengths equally and all phonological length is seen as a consequence of stress. Every syllable regarded as “strong” must in the educational setting be lengthened and the length must be associated with either the vowel or the post-vocalic consonant. Further motivation for taking an interest in stress induced increase of duration is given in Thorén (2008).

Can we measure “how well” a speaker of e.g. Swedish as a second language realizes the temporal patterns of Swedish without checking every segment and the expected duration of the segment as a consequence of speech rate, voicing, openness, stress, phonological length, pre-pausal position etc? The lengthening of phonologically long sounds – by giving them increased duration, is the pedagogical formula, but how can we treat it in experiments? Traunmüller & Bigestans (1988) found a neat connection between vowel duration and utterance duration for perceived Swedish quantity categories, but if we want to elicit spontaneous speech, most utterances are incomparable due to different length.

One of the intentions of the study is to find out whether the temporal patterns of Estonian – a language which utilizes duration to a greater extent than Swedish, having a three grade quantity distinction – and Spanish – a language without any phonological length distinction – are transferred into their pronunciation of Swedish.

The question is: Are there typical properties in the Swedish L2-speech of specific L1-speakers that are reflected in absolute or relative durations of phonologically long sounds in Swedish?

Measurements

Considering e.g. speech rate as a factor that influences segment duration (Fant et al. 1991) it is assumed that measures of absolute durations would not give robust information of how the speakers treat the segments that should receive extra duration according to Swedish phonological rules. First it must be decided what words are suitable to be regarded as “stressed”. Prominence degrees “accented” and “focused” are mostly clear on linguistic basis, but there is “secondary stress” in a sentence perspective (Anward & Linell 1976) occurring typically in verbs followed by a stressed particle or similar types of lexicalized verb phrases. These words are on the borderline between stressed and unstressed categories and can lose their quantity signals at fast speaking rates (Thorén 2008: 32).

Choice of speakers

A native Spanish and a native Estonian speaker were chosen. The Spanish speaker sounded very “choppy” and staccato-like and gave an impression of having equal shortness of all syllables, although sentence stress and word stress were signalled unambiguously. The Estonian speaker seemed to have a temporal pattern in her Swedish that resembled a native one. Both had unmistakable foreign accent and none of them was a beginner in Swedish. These two speakers seemed ideal to record and have rated for intelligibility and listener friendliness. Their recorded speech was also assumed to provide good raw material for improved and deteriorated duration patterns.

Method

In order to elicit productions of many identical words, three native speakers of respectively Swedish, Estonian and Spanish were recorded when telling a short story inspired by a short cartoon showing a boy playing football in the road, being hit by a car, going by ambulance to the hospital and then coming back home with crutches and his leg in plaster, and in this condition playing with his ball again. The Swedish speaker was the last to be recorded and he was explicitly asked to include as many words as possible from a list elicited from the two other speakers, in order to obtain many common words for comparison.. Each recording lasts between 1 and 2 minutes. Words that were unambiguously stressed were measured with respect to durations of phonologically long sound and the duration of the entire word. Relative durations are defined as segment duration divided by (word duration minus duration of the segment of interest). Words that were unambiguously stressed were nouns, verbs and an adjective, which are all typical content words. Measurements were made in Praat (Boersma & Weenink 2001).

Result

Figure 1 below shows absolute durations of phonologically long segments. The typical case is that the Spanish speaker has durations shorter than the other two, and that the Estonian speaker has durations equal to or mostly higher than the native Swedish speaker. The pattern is similar for both vowels and consonants. The segments produced by the Swedish speaker are in three words (*bil* ‘car’, *tappar* ‘looses/drops’, *kommer,* ‘comes’) longer than the corresponding segment produced by the Estonian speaker.

The Spanish speaker shows a remarkably stable shortness of his vowels, whereas the Estonian speaker shows the greatest variation in absolute vowel durations. The native Swedish speaker has an intermediate degree of variation.

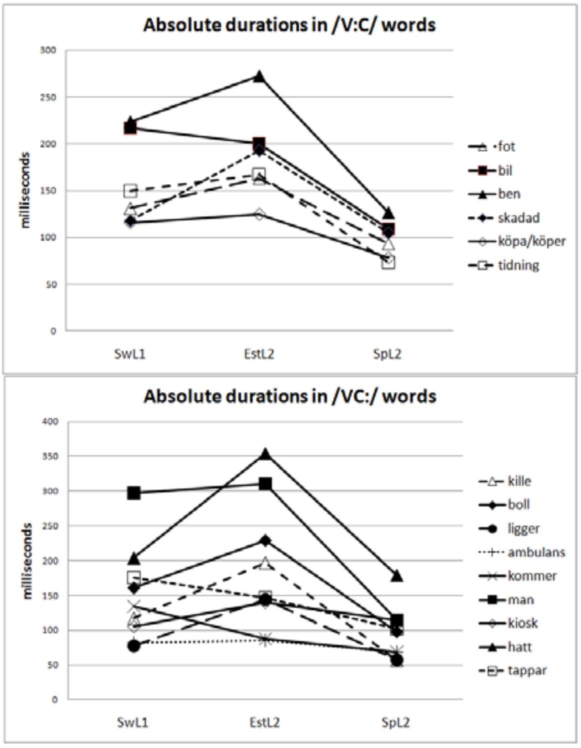


Figure 1.Upper panel: Absolute durations of phonologically long vowels. Lower panel: Absolute durations of phonologically long consonants.

This pattern comes back in the consonant durations, but with greater variations for all three speakers.

Figure 2 shows relative durations for all phonologically long segments. The pattern from the absolute durations are reflected also in this measure.

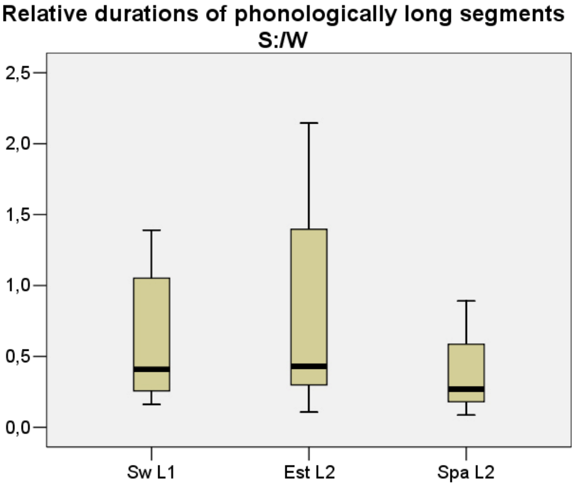


*Figure 2. Duration ratios for all phonologically long vowels and consonants in common words. Segment duration is divided by word duration minus the duration of the long seg*ment.

Discussion

The results confirm the expectation that a native speaker of Spanish would use duration to a lesser extent than an Estonian speaker when speaking Swedish. Surprisingly, the Estonian speaker not only applies the temporal pattern of Swedish, but exaggerates it in so far as to give phonologically long segments – both vowels and consonants – even longer durations than the native Swedish speaker does. Had we only looked at absolute durations, one could suspect that a slower speaking rate was the main cause of the longer durations, but as the relative durations show the same pattern, the result can be assumed to reflect a true tendency.

It should be admitted that also the relative measures are problematic since words can be followed by a pause, which in turn can cause a final lengthening of e.g. a final unstressed syllable. This is frequent in the present recordings since the speakers hesitated rather often while looking at the pictures and figuring out how they would continue. Final lengthening in unstressed syllables would render the word a longer duration without increasing the duration of the segment of interest, thereby reducing the segment/word-ratio. The continued study will hopefully show whether increased durations in the Spanish speaker will influence the perception of his speech by native Swedish listeners.

The greater variation of the segments produced by the Estonian speaker could be explained by the three-degree quantity distinction in Estonian, involving “short”, “long” and “over-long” as phonological categories. These categories are mainly relying on segment duration, e.g. Engstrand & Krull (1994). The Estonian speaker can be assumed either to vary segment duration in Swedish as a result of great variation in her first language, or as a result of perceiving the variation in Swedish segment duration as corresponding to different distinctive length categories in her L1. The longer durations produced by the Estonian speaker could also be the result of a distinct speaking style (Fant et al. 1991).

References

Anward J & Linell P (1976). Om lexikaliserade fraser i svenskan. *Nysvenska studier 55- 56*. 77-119

Boersma P & Weenink D (2001). *Praat – a system for doing phonetics by computer*. http://www.fon.hum.uva.nl/praat/

Elert C-C (1964). *Phonological Studies of Quantity in Swedish*. Uppsala: Almqvist & Wiksell.

Fant G et al. (1991). Temporal organization and rhythm in Swedish, *Proceedings of the XIIth ICPhS*, Aix-en-Provence, 251-256.

Fant G & Kruckenberg A (1994). Notes on stress and word accent in Swedish *STL-QPSR* 2-3/1994

Thorén B (2008). *The priority of temporal aspects of L2-Swedish prosody – Studies in perception and production.* Doctoral dissertation. Department of Linguistics, Stockholm University.

Bertil Malmberg, first Chair of Phonetics in Sweden
Peer Evaluations

Paul Touati
Centre for Languages and Litterature, Lund University

Abstract

Following a royal act, a new academic discipline, Phonetics, was officially given birth in 1947. However, it was not until November 1950 that, at Lund University, Bertil Malmberg was appointed to the first Chair of Phonetics in Sweden.

Four Scandinavian experts were chosen to evaluate Malmberg's scientific competence. The peers produced four evaluations, which constitute the central corpus for the present study. In order to analyse these evaluations, I will make use of a hermeneutic approach combining three ways of interpretive positioning - three stances. The first stance fits into the field of the history of science. The second stance concerns the rhetorical and textual aspects of the peer evaluations. The last stance is more epistemological in character. It mainly concerns the relationship between scientific knowledge and the definition of an academic discipline.

First Chair of Phonetics and Peer Evaluations

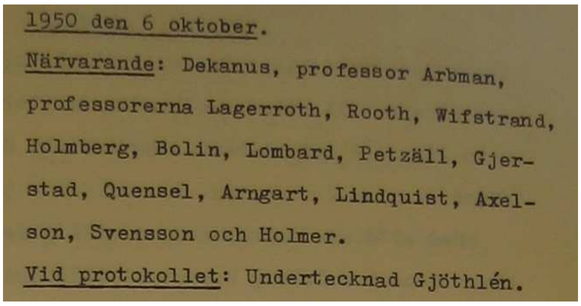
When in 1947, following the royal act no 272, the new academic discipline, Phonetics, was officially given birth, the two oldest and most famous universities in Sweden, Uppsala University (founded 1477) and Lund University (founded 1666) were designated to house the chairs (Gjerdman, 1959).[1] However, it was not until November 1950 that, at Lund University, Bertil Malmberg was appointed to the first Chair of Phonetics in Sweden (Gårding, 1994; Sigurd, 1995). The young romanist – Malmberg was born in 1913 and had defended his thesis on *Le roman du Comte de Poitiers* in 1940 – was "called" to apply; thus sole candidate to the position. Four experts were appointed to evaluate Malmberg's scientific competence. Their work consisted essentially in answering the question as to whether Malmberg was qualified enough for the task in view.

The peers produced four evaluations, which constitute the central corpus exploited for the present study[2]. In order to analyse these evaluations, I will make use of a hermeneutic approach combining three ways of interpretive positioning - three stances. The first stance fits into the field of the history of science. In fact, I consider the appointment of a candidate to a new Chair as a key event in the history of a discipline, just as the textual records that enforced that event, the evaluations and comments made by the scientific and academic actors involved, are key documents to understanding the state of the art in phonetic science at the moment that its status as an academic discipline was ratified politically. The second stance concerns the rhetorical and textual aspects of the peer evaluations. The questions asked are as follows: do the peer evaluations pertain to the rhetorical domain of the epideictic? Do these evaluations express consensus or controversy? What kind of texts are produced by the four peers? The last stance is more epistemological in character. It mainly concerns the relationship between scientific knowledge and the definition of an academic discipline. In the important historical moment when specific scientific knowledge is ratified by institutionalisation, peer evaluations are absolutely crucial in defining the discipline. For instance, in our particular case, the question as to whether Phonetics should be considered as part of the Humanities or as an Experimental science was repeatedly actualised. It should be noted that these three interpretive stances are, in my opinion, quite inseparable; only when taken together, do they capture the significant movement of science in action (Latour, 2005).

The section of Humanities, the peers and the candidate

The section of Humanites

Of the twenty-six professors at the section of the Humanities at Lund University in 1950, twenty were involved in the appointment of Malmberg to the first Chair of Phonetics. On October 6 1950, fifteen professors participated in the meeting during which the appointment was unanimously voted and immediately sent to the Grand Consistory. The list is as follows:



1950 den 6 oktober.

Närvarande: Dekanus, professor Arbman, professorerna Lagerroth, Rooth, Wifstrand, Holmberg, Bolin, Lombard, Petzäll, Gjerstad, Quensel, Arngart, Lindquist, Axelson, Svensson och Holmer.

Vid protokollet: Undertecknad Gjöthlén.

In: Protokoll 1950-10-06, Lunds universitet, Filosofiska Fakultetens Arkiv 1666-1956, Volym A1A: 122

The peers

They came from four different Scandinavian universities. The only national expert, Olof Gjerdman [1883-1965] came from Uppsala University. His Scandinavian fellow experts were, a Norwegian, Ernst Selmer [1890-1971] from the University of Oslo, a Dane, Louis Hjelmslev [1899-1965] from the University of Copenhagen and a Finn, Antti Sovijärvi [1912-1995] from the University of Helsinki. This pool was thus composed of experts not only coming from four different countries but also, as we will see, from four different rhetorical and academic traditions. The evaluations were written in three different languages: Swedish (Gjerdman and Sovijärvi), Norwegian (Selmer) and Danish (Hjelmslev). It should also be added that these experts belonged, by their age, to different generations of scholars. In 1950, Gjerdman, the oldest of the four, was aged 67, Selmer, 60, and Hjelmslev, 51. Sovijärvi was the youngest. With his 38 years, he was the only one that belonged to the same generation as the candidate (Malmberg was 37 at the time).

The candidate

Bertil Malmberg was described in the following terms by Gjerdman, the Swedish expert:

“The sole candidate, Docent Bertil Malmberg, who was born in 1913, has a BA embracing German, English and Romance languages, Scandinavian languages and pedagogy. In 1940, he defended his thesis on *Le roman du Comte de Poitiers* [...]”; the same year he became a lecturer in Romance languages at Lund University [...]. In 1938, he obtained a diploma in phonetics at the Institute of Phonetics in Paris. Among the study trips made by Malmberg, I mention three: a journey to South America (Argentina, Paraguay, Uruguay) made in 1946, a journey to Norway, Denmark, Netherlands, Belgium, France and Italy in 1947 and a journey to Finland undertaken in 1948. Funded by Lund University, the aim of the last two was to gain knowledge of how research (instrumental) and teaching in phonetics was effectuated in these countries“.[3]

Peer evaluations rhetorical and textual aspects

Rhetorical aspects

According to a list presented by Malmberg, the number of his publications amounts to 75, of which 59 were submitted to peer evaluation (see also “Bertil Malmberg Bibliography” by Gullberg, 1993). The experts were supposed to use the publications submitted by Malmberg to evaluate the candidate's scientific potential and, at the same time, estimate the actual scientific strength of his production in the light of advanced contemporary research in phonetics. The fundamental scientific positioning adopted by each expert was in fact “staged” in a text that was built along a dichotomised dimension constituted by the span between the poles “praise-vs-blame”, the power inherent to such arguments are well known to epideictic rhetoric. The section of Humanities at Lund fully understood the importance of the kind of arguments chosen. In assessing Malmberg’s scientific contribution in the rhetorical terms of “praise or blame”, the Scandinavian peers did not only judge a candidate but also a potential discipline. Praising Malmberg's work meant considering him “the right person at the right place”, something that, in its turn, assured that consensus and social legitimacy was bestowed upon the new discipline. On the other hand, blame causes controversy and involves putting not only the candidate but also the scientific content at stake. It seems as though each of the four peers took a different stand along the continuum “praise-blame” and consequently also that of “consensus-controversy”. Each positioning had consequences. “If praise implies consensus, [it] engages a vision of ethical preference involving the whole theory of *decision* [...]”(Dominicy and Frédéric, 2001:15). It is evident that blame also leads to decision. Modern epideictic theory (Dominicy, 2001) stresses that the use of the span “Praise-Blame” easily permits the transformation of spontaneous preference into “rational” preference, something that may be perceived in the peers' different manner of evaluating Malmberg's contribution.

Textual aspects.

The texts produced by the experts are of different length. Selmer produced the longest text (31 pages). It is followed in descending order by the text of Gjerdman (27 pages), of Hjelmselv (18 pages) and finally of Sovijärvi (9 pages). The overall textual structure is highly similar. It consists of an opening, a body where some chosen publications are discussed and finally a closure in which each expert expresses his overall evaluation and makes a decision. If the classical tripartite compositional structure is similar, the fact remains that content and style differ markedly. In the following, text openings are used as illustrations of the diverging styles adopted. As we have seen before in the quote from Gjerdman, his account started off with a *biographical* account, intended to put the life and writings of Malmberg, the scientist, into perspective. Sovijärvi, however, began with an *enumeration*. He lists Malmberg’s areas of competence: “I ranked the candidate's publications which can be used to assess his competence in 5 groups: 1) phonetics, 2) experimental phonetics, 3) phonology, 4) phonetics and historical linguistics […] and 5) phonetic and pedagogy”. It may be noted that Sovijärvi did not include Malmberg's publications in philology, with the one exception of an article entitled “A child changes language”. In contrast again to Gjerdman, Selmer chose to integrate Malmberg in a line of prominent phoneticians. He started with a *list* of names of renowned phoneticians belonging to a vast panorama of countries and working under different definitions of their discipline: “In Norway, the chair is described as chair of “general” phonetics which I think is a vague qualification that gives rise to many interpretations. In Finland [...] if we based it on representatives of the discipline [Pipping, Alma, Sovijärvi] we can say that the orientation is rather that of a natural science. Naming people like D. Jones, Fouché, Menzerath, Calzia, N. Tomàs, Meriggi as all more or less representatives of “general” phonetics [points to] widely divergent directions [...]”. As a further argument in favour of the diversity of profiles in phonetics, Selmer expanded his list to names like E.A. Meyer, Millardet, Grammont, Duraffour and Rousselot. Finally, Hjelmslev began his evaluation by advancing a *definition* of phonetics: “As the chair (to fill) has been described as a chair of phonetics without further explanation, I permit myself to assume that the definition of phonetics applied to this Chair is to be understood in the broadest sense, meaning that we think not only about classical phonetics, that is physiological phonetics reclining on an auditory basis, but as experimental phonetics, on one side the most recent, mainly acoustic oriented development, on the other, the theory of phonemes, […]”.[4] Hjelmselv continued his definition by saying that phonetics is not only the field of synchronic descriptions but also one of diachronic analysis, and added that it should be seen as an advantage if the holder of the Chair also had some knowledge also in the areas of general linguistics, psycholinguistics as well as in philological method. From the beginning, Hjelmslev was playing his cards by being very positive towards Malmberg, whom he saw as “very active” in all areas included in defining phonetics.

Scientific Knowledge and Academic Discipline.

To confer institutional legitimacy to scientific knowledge by giving it the status of an academic discipline and by creating a Chair that is appointed for the first time was an ambitious goal and not an easy task for any party involved. In our case, the goal was reached by means of presenting the competence of the sole candidate, Bertil Malmberg, who was officially Docent in Romance languages as a phonetician. In order to decide if Malmberg could meet the terms, it was first necessary to advance a definition of phonetics, then to investigate the competence of the candidate. As it might be expected, the experts’ definitions vary. Hjelmselv gave a broad definition of phonetics as a discipline that covers areas as diverse as linguistics, philology and experimental phonetics. According to him, the work done by Malmberg was undoubtedly pertinent. In his evaluation Malmberg's work is described as “a synthesis of perspectives and methods of linguistics, experimental phonetics and classical phonetics”. There is therefore absolutely no doubt that Hjelmslev considered Malmberg as a “phonetician [who] meets the demands that nowadays are put on a scholar and a teacher [in the field of phonetics]”[5]. In his peer evaluation, Hjelmslev adopted the tone of “praise and consensus”. The peer evaluation produced by Sovijärvi is different in tone. He emitted strong reservations about Malmberg as a phonetician; obviously arousing from the narrow definition that the young Finnish phonetician had adopted. He considered phonetics essentially as an experimental science and concluded that although Malmberg’s merits were largely adequate for a Chair of “General Linguistics and Phonetics”, it was not perfectly clear that they were sufficient for a Chair of “Phonetics”.[6] He also pronounced some unwillingness to accept certain experimental methods used by Malmberg. The tone adopted by Sovijärvi tended to be the one of blame and controversy. The positioning of the other two peers may be situated in between these two poles, with a tendency towards a positive view on the pertinence of the work achieved by Malmberg. With all required lucidity, the evaluations of Hjelmselv and Sovijärvi already embody the opposing trends that were to divide the discipline twenty years later, working as a paragon examples of the uneasiness of the Humanities at large.

References

Domincy M and Fréderic M (2001). *La mise en scène des valeurs. La rhétorique de l’éloge et du blâme*. Lausanne : Delachaux et Niestlé.

Domincy M (2001). L’épidictique et la théorie de la décision. In : Domincy M and Fréderic M, éds, *La mise en scène des valeurs. La rhétorique de l’éloge et du blâme*. Lausanne : Delachaux et Niestlé, 49-77.

Gjerdman O (1959). *En Fonetiker som lärling och gesäll vid Uppsala Universitet (1903-195*3). Tierp.

Gullberg M (1993). Bertil Malmberg Bibliography, *Working Papers*, 40: 5-24.

Gårding E (1994). Prosody in Lund. *Speech Communication*, 15: 59-67.

Latour B (2005 [1989]). *La science en action*. Paris : Éditions de la Découverte.

Sigurd B (1995). Bertil Malmberg in memoriam. *Working Papers*, 44: 1-4.

Touati P (2009a). De la construction discursive et rhétorique du savoir phonétique en Suède : Bertil Malmberg, phonéticien (1939-1969). In : Bernardini P, Egerland V & Grandfeldt J, eds, *Mélanges plurilingues offerts à Suzanne Schlyter à l’occasion de son 65éme anniversaire*, Lund : *Études romanes de Lund*, 85 : 417-439.

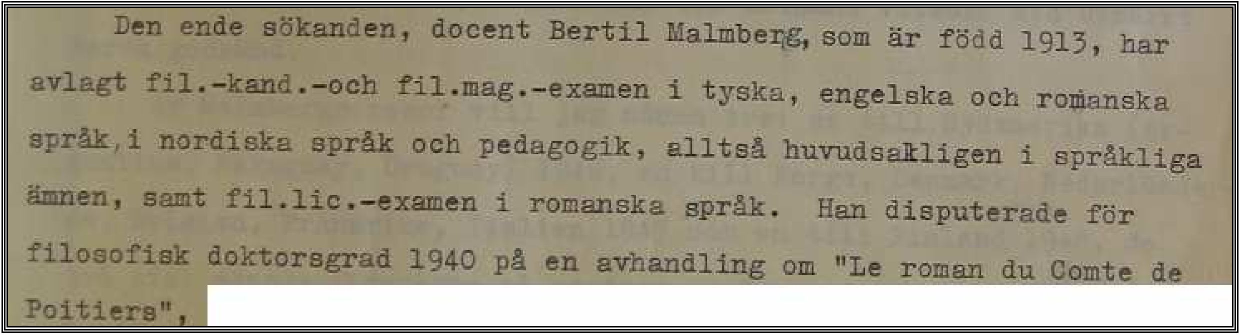
Touati P (2009b). Swedish phonetics 1939-1969. *Proceedings, FONETIK 2009*, Stockholm: 208-212.

Touati P (Forthcoming). De la médiation épistolaire dans la construction du savoir scientifique. Le cas d’une correspondance entre phonéticiens. *Revue d’anthropologie des connaissances*.

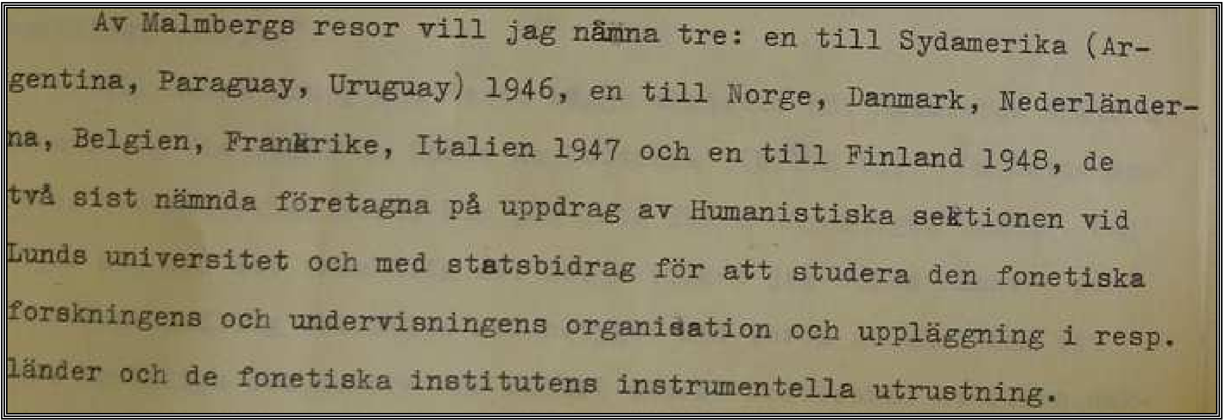
[1] Many thanks to professor Olle Engstrand for sending me Olof Gjerdman’s interesting autobiographical notes concerning his career as a phonetician (Gjerdman 1959).

[2] The present study is part of a research in progress concerning the historical, social, discursive, and rhetoric conditions that determined the emergence of phonetic science in Sweden between 1939 and 1969 (see Touati, 2009a, Touati, 2009b and Touati, Forhtcoming)

[3] Utlåtande Gjerdman, In: Lunds universitet, Filosofiska Fakultetens Arkiv 1666-1956, Volym A1A: 122:

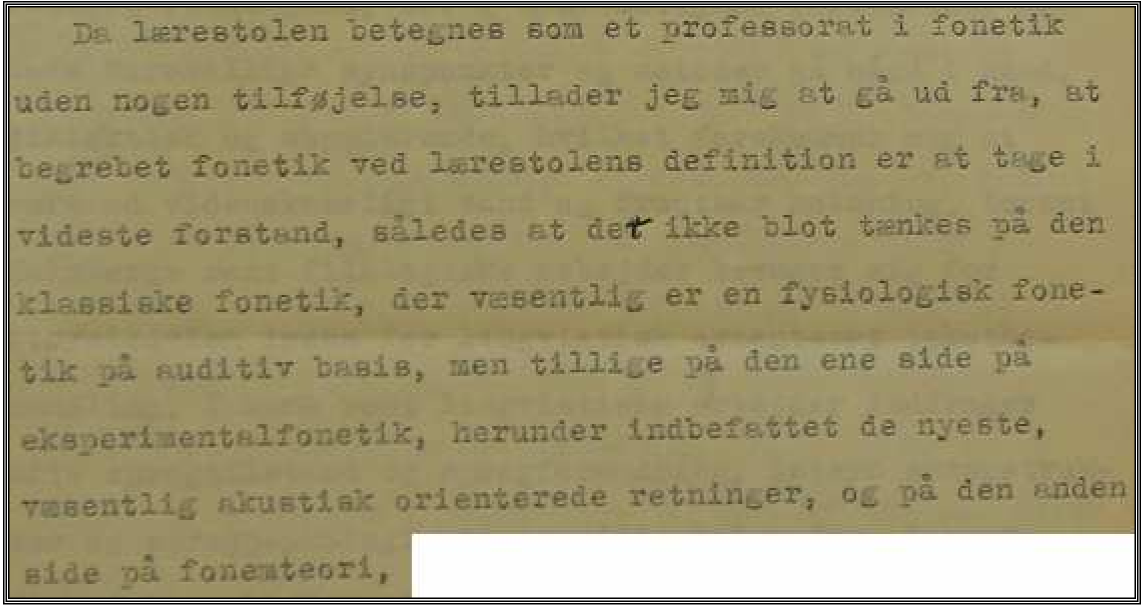


Den ende sökanden, docent Bertil Malmberg, som är född 1913, har avlagt fil.-kand.-och fil.mag.-examen i tyska, engelska och romanska språk,i nordiska språk och pedagogik, alltså huvudsakligen i språkliga ämnen, samt fil.lic.-examen i romanska språk. Han disputerade för filosofisk doktorsgrad 1940 på en avhandling om "Le roman du Comte de Poitiers",



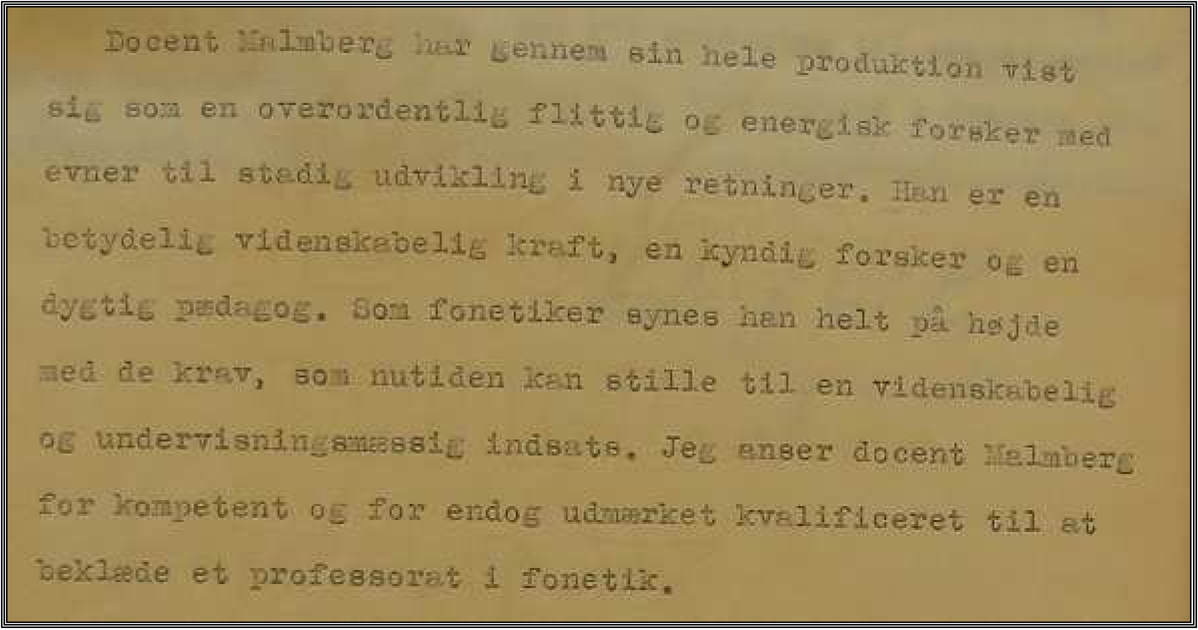
Av Malmbergs resor vill jag nämna tre: en till Sydamerika (Argentina, Paraguay, Uruguay) 1946, en till Norge, Danmark, Nederländerna, Belgien, Frankrike, Italien 1947 och en till Finland 1948, de två sist nämnda företagna på uppdrag av Humanistiska sektionen vid Lunds universitet och med statsbidrag för att studera den fonetiska forskningens och undervisningens organisation och uppläggning i resp. länder och de fonetiska institutens instrumentella utrustning.

[4]Utlåtande Hjelmselv, In: Lunds universitet, Filosofiska Fakultetens Arkiv 1666-1956, Volym A1A: 122:



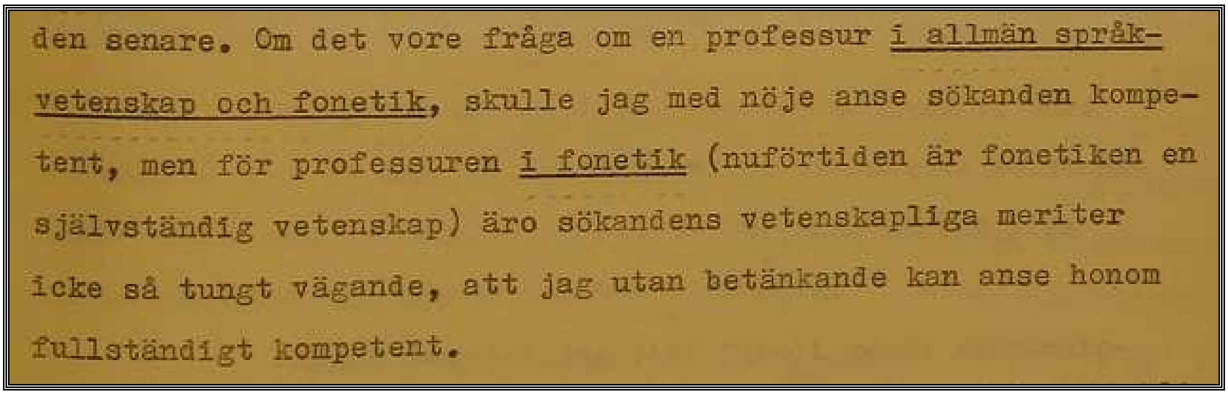
Da lærestolen betegnes som et professorat i fonetik uden nogen tilføjelse, tillader jeg mig at gå ud fra, at begrebet fonetik ved lærestolens definition er at tage i videste forstand, således at det ikke blot tænkes på den klassiske fonetik, der væsentlig er en fysiologisk fonetik på auditiv basis, men tillige på den ene side på eksperimentalfonetik, herunder indbefattet de nyeste, væsentlig akustisk orienterede retninger, og på den anden side på fonemteori,

[5] Utlåtande Hjelmselv, In: Lunds universitet, Filosofiska Fakultetens Arkiv 1666-1956, Volym A1A: 122:



Docent Malmberg har gennem sin hele produktion vist sig som en overordentlig flittig og energisk forsker med evner til stadig udvikling i nye retninger. Han er en betydelig videnskabelig kraft, en kyndig forsker og en dygtig pædagog. Som fonetiker synes han helt på højde med de krav, som nutiden kan stille til en videnskabelig og undervisningsmæssig indsats. Jeg anser docent Malmberg for kompetent og for endog udmærket kvalificeret til at beklæde et professorat i fonetik.

[6] Utlåtande Sovijärvi, In: Lunds universitet, Filosofiska Fakultetens Arkiv 1666-1956, Volym A1A: 122:



den senare. Om det vore fråga om en professur i allmän språkvetenskap och fonetik, skulle jag med nöje anse sökanden kompetent, men för professuren i fonetik (nuförtiden är fonetiken en självständig vetenskap) äro sökandens vetenskapliga meriter icke så tungt vägande, att jag utan betänkande kan anse honom fullständigt kompetent.

Earwitnesses: The effect of type of voice lineup in identification accuracy and the realism in confidence judgments

Elisabeth Zetterholm[1], Farhan Sarwar[2] and Carl Martin Allwood[3]
[1]School of Language and Literature, Linnaeus University, Växjö
[2]Department of Psychology, Lund University
[3]Department of Psychology, University of Gothenburg

Abstract

This contribution is a partial report from a study of the identification accuracy and realism in the confidence judgments of the correctness in the identification reports in two kinds of target-present voice lineup. 24 men and 54 women were asked to identify a voice that they had heard previously in a dialogue context that simulated the planning of a burglary by two males 22 and 27 years old. The voice lineup either consisted of recordings of each of six male speakers reading a text from a book (text-lineup condition) or each of the same six speakers having a spontaneous dialogue with another male speaker (dialogue-lineup condition). Each recording lasted 30 seconds. The results showed a tendency (p<.06) for better accuracy and better ability to separate correct from incorrect identification responses by means of ones' confidence judgments for the text-lineup condition compared with the dialogue-lineup condition. The text-lineup condition also showed a tendency for lower overconfidence. These results deviate from expections following from the encoding specificity principle in memory psychology (Tulving & Thomson, 1973), maybe because text reading provides a more varied representation of the features of the human voice compared to dialogues.

Introduction

A speaker with a distinctive and characteristic voice and speech is easier to recognize compared to a speaker with less distinctive features referring to different studies presented in Hollien (2002). An earlier study, focusing on characteristic features in the speech and voice, reported that a foil with characteristic voice features, close to the target speaker, is confusing for listeners in a voice lineup (Zetterholm et al, 2009). This paper focuses on the effect of different speaking styles, namely read text and spontaneous speech, in a voice lineup. It is known that speakers with an unfamiliar dialect or accent sound more similar to a listener than speakers with a familiar dialect or accent (Yarmey, 1995). In the present study, as in the previous study (Zetterholm et al, 2009), all speakers, both the target speaker and the foils, had the same Swedish dialect in order to avoid confusion about the dialect.

When preparing an earwitness lineup the samples should be equal in length and Hollien (2002) points out that it might be good to use different types of voice samples, e.g. natural speech, text-dependent words, phrases and sentences as well as read text. Hollien also recommends between five to eight foils of approximately the same age as the suspect, or target voice, as well as speaking the same accent or dialect in general.

Previous research about speaking styles have used different data and the material has been classified in two main groups. *Connected speech* consisting of read speech and *spontaneous speech* consisting of speech produced in more or less unprepared situations with both professional and non-professional speakers (Llisterri, 1992). Some of the studies show that average F0 is higher in reading than in spontaneous speech. This is not confirmed in present study, see Table 1 and 2. However, the speech samples used in this study is not longer than 30 s. All speakers are non-professional.

The task presented to the participants in this study involves retrieval from memory. In this context it is relevant that Tulving and Thomson (1973) formulated a central conclusion from psychological research on memory: the *encodning specificity principle*. In this context retrieval from memory is seen to be occasioned by cues to (or in) memory which activate similar memory elements as the content in the cues. For example, the content in a question will act as a cue to memory and will help to activate a possible answer to the question posed.

According to the *encodning specificity principle* we tend to remember better the more similar the cues provided are at the time of recall are to the encoded material that we want to retrieve. This means, for example, that the more similar the context of the occasion when the memory was encoded is to the context of retrieval, the more successful the retrieval is expected to be. The context is here taken to be the total situation that the memory retriever is in, especially aspects that are at the focus of the person's attention.

For example, in a recognition task situation such as a lineup, the more similar the features of one (or more) of the persons in a target present lineup are to the voice features of the culprit the greater is the chance for recognition. For this reason, in the present study we speculated that if the people in the lineup are presented in dialogue form when the originally encoded experience took the form of a dialogue then the chances for correct identification would be better compared with when the people in the lineup are presented in some other format, such as reading from a book.

In brief, the aim of this study thus was to analyze if there is a difference in the ability to recognize a voice in a voice lineup depending on the speaking style of the people being presented in the lineup.

Method

Participants

Seventy-eight students (54, females, 24 males) from Lund University participated in the study. The mean age was 26.5 years (range 20 - 65 years). Each participant received a lottery ticket worth 25 SEK. Two participants were removed from the data because they choose more then one option in the voice lineup.

Design

A between-subjects design with two conditions was used. The conditions differed with respet to the format of the presentations of the voices in the lineup. In the first condition, the *Text-lineup condition* (n = 38), each of the people in the lineup read the same passage from a book. In the second condition, the *Dialogue-lineup condition* (n = 43), each of the voices in the lineup was engaged in a dialogue.

Material

Original event dialogue. A 2 minutes long dialogue between two male speakers was recorded. The dialogue simulated the planning to break into a house by two burglars. The speakers were 27 and 22 year old respectively and both spoke with a Scanian dialect.

Voice lineups. Two kinds of voice lineups were used. A "text" voice lineup which consisted of six recordings of six male speakers reading a text, the target speaker and five foils. They all read the same text. A "dialogue" voice lineup which consisted of six recordings of the same six male speakers having a spontaneous dialogue with another male speaker. The discussion partner was the same for all the six speakers. The voice of the conversation partner was not audible in the lineups. They all engaged in a dialogue about the same news article. The duration of each recording, in both conditions, was about 30 seconds. All the speakers in both lineups had a Scanian dialect (i.e., the dialect of the southern-most part of Sweden) and were of almost the same age. The target speaker was present in the test. The results of the acoustic measurements of the mean fundamental frequency (F0) and the standard deviations (SD.) for both lineups are shown in Table 1 and 2 respectively. It is obvious that the there is almost no difference in mean F0 or SD. between the recordings with the different speaking styles.

Table 1. Age, F0 mean and standard deviations (SDs) for the target speaker and the five foils, in the text-lineup condition.

	Age	F0, mean	SD.
Target	22	125 Hz	21 Hz
Foil 1	19	99 Hz	15 Hz
Foil 2	21	121 Hz	21 Hz
Foil 3	23	86 Hz	12 Hz
Foil 4	23	123 Hz	26 Hz
Foil 5	22	90 Hz	16 Hz

Table 2. Age, F0 mean and standard deviations (SDs) for the target speaker and the five foils, in the dialogue-lineup condition.

	Age	F0, mean	SD.
Target	22	132 Hz	23 Hz
Foil 1	19	100 Hz	21 Hz
Foil 2	21	126 Hz	30 Hz
Foil 3	23	87 Hz	16 Hz
Foil 4	23	114 Hz	23 Hz
Foil 5	22	87 Hz	17 Hz

Confidence judgments. Participants rated their confidence in their lineup decision on an 11-point scale beginning at 0% ("Completely sure that I remember wrong") and ending at 100% ("Completely sure that I remember correct").

Procedure

The experiment was run in small groups with 2 to 5 participants in each group. Each group was randomized into one of the two conditions. The participants were received in the lab where they first signed the consent form. Next, the participants were told that they were going to listen to a dialouge between two men. They were instructed to just listen to the dialouge and that they would recieve further instructions afterwards. Participants listened to the 2 minutes long dialogue between two men planning to brake into somebody's home. After this, the participants participated in another experiment for about 15 minutes as a filler task.

The participants were then told that their task was to identify the speaker that that they had heard most in the original dialogue. Then the participants listened to the voice lineup relevant for their condition and each lineup was played twice. After listening to the voice lineup the listeners answered the question: "Do you recognize if any of these six voices is from the person from the dialouge that you listened to earlier?" If they could not identify anyone in the lineup as the culprit they could choose the option "do not recognize any of the voices". The participants were also told to be aware of that the voice they had heard in original dialogue might not be present at all in the lineup. After this, the participants gave their confidence judgement about the correctness of their idenfication decision on the 11-point confidence scale described above. Finally, the participants were given a lottery ticket, debriefed and thanked before leaving.

Measures

Apart from the accurcy in the identification responses (proportion correct of all identification responses) and the participants' confidence in the correctness of their identification responses, we also calculated three measures of the realism in the participants' confidence judgments. By the *realism* in the participants' confidence judgments we mean how veridical the confidence judgments were with respect to the correctness of the identification responses (this is sometimes called the participants' metamemory realism).

Two of these measures, *calibration* and *over-/underconfidence*, concerned the relation between the level of the participants' confidence judgments and the proportion correct identification responses.

Calibration is calculated by first dividing a person's confidence judgments into different confidence classes based on the level of confidence (11 confidence classes were used, since the participants were allowed to use 11 confidence levels, that is, 0%, 10%, 20%, etc). Calibration is computed by the following formula:

$$\text{Calibration} = 1/n \sum_{t=1}^{T} n_t (r_{tm} - c_t)^2$$

Here *n* is the total number of responses rated, *T* is the number of confidence classes used and n_t is the number of responses within confidence class r_t , r_{tm} is the confidence level of the confidence class r_t, and c_t is the percent of correct responses within the confidence class *t*. For each confidence class the percent of correct answers within that class was thus subtracted from the mean level of confidence within that class. This difference was squared and multiplied with the number of times this confidence class was used by the listners. The resulting product was then summed over the corresponding products for the other confidence classes and finally this sum was divided by the total number of responses (for further details see e.g., Yates, 1994).

Over-/underconfidence is computed by subtracting the listners' average proportion correct responses (in percentage) from their average confidence level for all responses (also in percent). Just as for calibration, the result zero expresses perfect realism, in the sense that there is no over-/underconfidence. A negative value indicates underconfidence and a positive value indicates overconfidence.

We also used a measure of the listners' ability to discriminate correct from incorrect identifications by means of their confidence judgments. For this purpose we used a measure called *resolution* which is computed as:

$$\text{Resolution} = 1/n \sum_{t=1}^{T} n_t (c_t - c)^2$$

Here, c is the proportion of all items for which the correct identification response was given. A higher value reflects better resolution than a lower.

Results

Calibration curves

Figure 1 shows that calibration curves for the text-lineup condition and the dialogue-lineup condition. The x-axis shows the eleven different confidence levels (from 0 to 100%) and the y-axis shows the percent of correct answers, but in the graph the data-points have been reduced to five (0 %, 10 - 40 %, 50 %, 60 - 90% and 100 %). This reduction was done in order to smooth the calibration curves since the number of listners was small. The numbers inside the graph give the number of answers for each of the five reported confidence level in each condition. The diagonal shows perfect calibration. As can be seen in Figure 1, the calibration curves for the text-lineup and dialogue-lineup conditions show a difference in that the text-lineup condition evidence less overconfidence compared with the dialogue condition, However, at the 50% level there was no difference between the conditions and at the confidence levels below 50 % there is even a hint of underconfidence for the text-lineup condition.

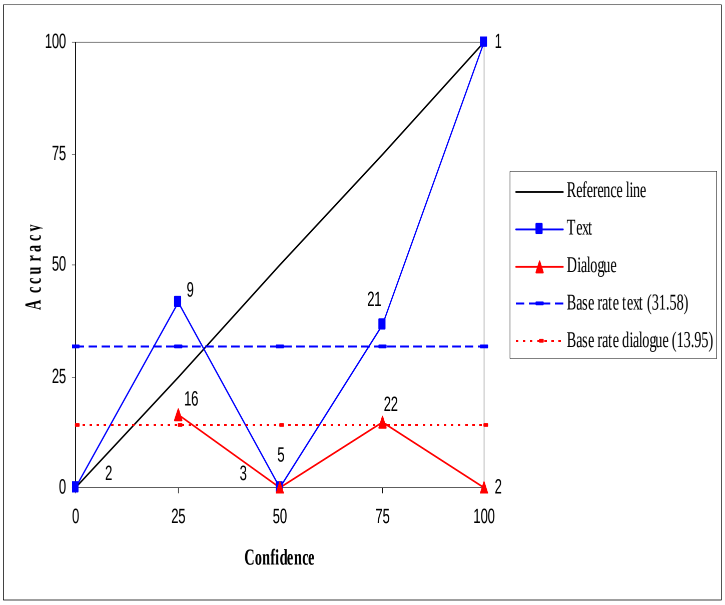


Figure 1. Calibration curves for text-lineup condition (squares) and dialogue-lineup condition (triangles). Digits at each point show the number of listners in the condition that used this confidence level.

Only 17 listners selected the target in the lineup, which is 21 % of all listeners. The average confidence level of these 17 listners was 64 %, which can be compared with the average confidence level 53 % for the remaining 79 % listners who selected the foils in the lineup. There is no significant difference between the mean confidence level of the listners who selected the target and those listners who selected the foil.

To study the relationship between age and accuracy a point-biserial correlation was computed. There was a very weak correlation between the two variables (r = .12, n.s, n = 76). In order to study the relationship between age and confidence a Pearson correlation was computed that similarly showed a very weak relationship between age and confidence (r = -.03, n.s., n = 76).

To test the importance of gender in voice identification we used t-tests to compare the males and females on their accuracy, confidence, calibration, over-/underconfidence and resolution scores but no differences were found.

The listners' scores in the two conditions were also submitted to independent-sample t-tests. The results are shown in Table 3. It can be seen that there was a trend towards a significant difference between the conditions for accuracy t(79) = 1.92, p < .06 and for resolution t(79) = 1.92, p < .06, meaning that the participants in the text-lineup condition showed better performance than the participants in the dialogue-lineup condition for accuracy and resolution. Furthermore, a trend towards a significant difference between the two conditions was also found for over-/underconfidence. The participants in the dialogue-lineup condition showed a tendency towards greater overconfidence compared with the participants in the text-lineup condition. No significant differences were found between the text-lineup condition and the dialogue-lineup condition for confidence and calibration. For accuracy and resolution the Levene's test showed that equal variance for these two variables cannot be assumed, but the t-value for accuracy and resolution was same both with and without assuming equal variance and both t-values were significant at p < .06 level, as shown in Table 3.

Table 3. Accuracy, confidence, calibration, over/underconfidence, and resolution for the text-lineup condition and the dialogue-lineup condition.

	Text-Lineup	**Dialogue-lineup**	**t**	**p**
Acc.	31.58(47.11)	13.95(35.06)	1.92	.06
Conf.	55.26(27.09)	54.88(26.58)	.06	.95
Calib.	.28(.25)	.36(.28)	-1.42	.16
O-/u.	.24(.48)	.41(.44)	-1.68	.09
Res.	.32(.47)	.14(.35)	1.92	.06

Note. Acc. = Accuracy, Conf. = Confidence, Calib. = Calibration, O-/u. = Over-/under-confidence, Res. = Resolution.

Discussion and conclusions

In the present study we tested whether voice-lineup recordings resulted in better identification performance when the lineups were presented in the form of the voices reading from a book (the text-line condition) or voices participating in a dialogue (the dialogue-lineup condition). Our expectation was that the lineup-condition using the dialogue presentations whould result in the better performance. This expectation was based on the encoding specificity principle presented by Tulving and Thomson (1973) which says that the accuracy in memory retrieval should be better when the there is a better match between the features of the encoding situation (including the focussed "object") and the features of the retrieval situation (included the attended-to "object").

In contrast to our expectation, the results showed that it was the listners in the text-lineup condition that showed the best performance of the two conditions. However, this is not the first time that the encoding specificity principle has been challenged by empirical data (e.g., Bower & Mayer, 1989; Higham 2002). For example, Bower and Mayer reported six experiments that failed to show stable evidence for mood-dependent retrieval.

In line with previous research on voice lineups (e.g., Olsson, Juslin, & Winman, 1998; Yarmey, 2007) the level of overconfidence evidenced by the listners in this study was quite high (especially in the dialogue condition) if compared to what is reported in research on eyewitness lineups. It remains a task for future research to provide information as to why this is the case.

As noted above, the present contribution reports the results from the first 78 listners of the at least 200 listners that we plan will participate in the study when it is completed. A further limitation of the results in the present report is that it is not clear what the effect was of the fact that the recordings in the dialogue-lineup condition did not include the conversation partner, only the person being part of the lineup. This was done so that the listner should be able to concentrate on the voice of the person in the lineup and so that the listner should not be confused about which voice we wanted them to respond to. Moreover, our study only used target-present voice lineups. In future research also target-absent voice lineups should be investigated with the two conditions used in the present study.

Given that our results hold up in our remaining data-collection and in future research it is of great interest to investigate why the presentation of text recordings in voice lineups lead to better indentification performance, including better meta-memory performance. One speculations is that this effect, if it is real, is due to that text reading provides more varied and representative information about a speaker's voice compared with a voice participating in a dialogue.

Acknowledgements

This work was supported by a grant from Crafoordska stiftelsen, Lund.

References

Bower GH, Mayer JD (1989). In search of mood-dependent retrieval. In D. Kuiken (Ed.), Mood and memory: Theory research and applications. *Special issue of Journal of Social Behavior and Personality, 4*, 121-156.

Higham PA (2002)/. /Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30*(1), 67-80.

Hollien H (2002). *Forensic voice identification*. San Diego, CA: Academic Press.

Llisterri J (1992). Speaking styles in speech research. *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, Ireland, 15-17 July 1992. http://liceu.uab.es/~joaquim/publications/Speaking Styles_92.pdf.

Olsson N, Juslin P, Winman A (1998) Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied, 4*, 101–118.

Tulving E, & Thomson DM (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352-373.

Yarmey AD (1995) Earwitness speaker identification. Psychology, Public Policy, and Law 1, 792-816.

Yarmey AD (2007) The psychology of speaker identification and earwitness memory. In R.C. Lindsay, D.F. Ross, J. Don Read & M.P. Toglia (Eds.), *Handbook of eyewitness psychology, Volume 2, Memory for people* (pp. 101-136). Mahwah, N.J.: Lawrence Erlbaum Associates.

Yates JF (1994). Subjective probability accuracy analysis. In G Wright & P Ayton (Eds.), *Subjective probability* (pp. 381-410). New York: John Wiley & Sons.

Zetterholm E, Sarwar F, Allwood CM (2009). Earwitnesses: The effect of voice differences in identification accuracy and the realism in confidence judgments. *Proceedings, Fonetik 2009,* Dept. of Linguistics, Stockholm University.

Post Surgery Effects on VOT for Parkinson Disease STN/DBS Patients

Joel Åkesson[1,2]*, Jonas Lindh*[1]*, Lena Hartelius*[2]
[1]*Dept. of Philosophy, Linguistics and Theory of Science,* [2]*Institute of Neuroscience and Physiology, University of Gothenburg*

Abstract

In this paper we discuss and analyse voice onset time (VOT) pre and post surgical treatment with deep brain stimulation (DBS) in 17 patients diagnosed with Parkinson's disease (PD) at Sahlgrenska University Hospital in Gothenburg, Sweden. The patients were all at different stages of the disease but with the common denominator they have all undergone surgery to enhance synaptic responses through bilateral electrode implants in the subthalamic nucleus (STN) region of the brain, also known as Deep Brain Stimulation (DBS).The main focal point of the paper is to compare the pre and post surgery VOT data to see if there were any effects stemming from the STN surgery. Preliminary results for Mean VOT, Standard deviation VOT and percent of unsuccessfully produced/unmeasureable diadochokinetic syllable repetitions are presented and discussed. We found that the standard deviation decreased significantly for the consonant /p/ and this is discussed in the perspective of the ease of articulation of the different plosives.

Background and Introduction

Today there is a relatively large amount of literature on the effectiveness of surgical treatment with deep brain stimulation (DBS) to the subthalamic nucleus (STN) in Parkinson's disease (PD) patients. Several studies have documented the effectiveness of STN-DBS surgery concerning improved motor limb functions (Dromey et al., 2000), but studies investigating improvement of PD articulatory and/or phonatory speech symptoms are still small in numbers (Iulianella, 2007). The effects of deep brain stimulation of the subthalamic nucleus (STN-DBS) on speech are varied and inconclusive. Dysarthric symptoms frequently appear as side-effects and pre-existing dysarthria can be worsened. Dysarthria is reported as an adverse side effect in as many as 4-17% of patients (Benabid AL et al, 2009). The studies available on speech-related symptoms in PD in context with neurostimulatory operations show varying results in their reports where some conclude positive effects and reduced symptoms on specific motor functions, (Obeso et al., 2001) whereas others report negative effects: "With stimulation, precision of the glottal and supraglottal articulation as well as the phonatory function is reduced for some individuals, whereas for other individuals an improvement is observed." (Pützer et al., 2008). These studies emphasise the importance of individual case evaluation and substantial differences on the effects of neurostimulatory surgery on different speech functions and they are cautious determining or stating any general global effectiveness resulting from the surgical procedure. Positive reports on STN effectiveness on certain areas of speech can in some cases possibly be linked to the selection criteria itself (e.g. Gentil et al, 2003, Hoffman-Ruddy et al., 2001) where the selection of participants are made "on the basis of a significant speech impairment" (Iulianella, 2007). Apart from small patient groups, the variability in reported effects of STN-DBS on speech may be accounted for by a number of factors, including disease-specific variables, type and degree of dysarthria before surgery, and stimulation-related variables such as location of electrodes, amplitude and frequency of stimulation. Recent findings suggest that other stimulation sites may be more promising in terms of speech effects (Plaha et al., 2006).

Nearly 90% of individuals with Parkinson's disease (PD) develop speech disorders during the course of their illness but only 2–3% receive speech treatment. Speech disorders associated with PD are characterised by reduced voice volume (with a tendency for voice volume to decay over time), poor voice quality, reduced pitch variations, and reduced range of articulatory movements. Many patients also develop a tendency for speech articulation to festinate or rush as well as hesitant and/or dysfluent speech. Furthermore, PD exercises a strong influence on communication even before alterations in intelligibility or motor status become apparent.

The two main types of speech assessments, included in speech-language pathology practice, are auditory–perceptual analysis and instrumental analysis. An auditory–perceptual analysis is based on recordings and can be done blindly, but the method has questionable reliability. Instrumental analysis, e.g. acoustic analysis, is a more reliable measure and gives more precise measurements. For example, it permits measuring aspects of speech that are not yet audible or measurable in other parts of the body, and has therefore been described as a potential biomarker of early disease progression (Harel, et al. 1994).

The aim of the present project is to explore the effects of surgical treatment with DBS on a selected relevant acoustic aspect, voice onset time (VOT), defined as "…the time between the release of the oral constriction for plosive production and the onset of vocal fold vibrations…" (Özsancak et al. 2001). This aspect is thought to reflect a timing component of the motor speech process, particularly sensitive to various types of treatment.

Method

Participants

The patient group was not selected according to any specific age criteria and patient age ranges between 49-75 years. In total there are 8 men and 9 women represented. The 17 participants have all undergone bilateral subthalamic nucleus implant operations and are subject to so called deep brain stimulation.

STN-surgery

This specific operating procedure involves implanting electrodes in the subthalamic nucleus region of the brain as to enhance synaptic response. Patients are selected for surgery based on specific inclusion and exclusion criteria and are well documented pre and post surgery, regarding motor ability, cognition, depression and speech. Stimulators are turned on 1-2 days post surgery. Several post-surgical investigations are conducted, at 6 months, 1, 2, 3, 5, 7 and 10 years.

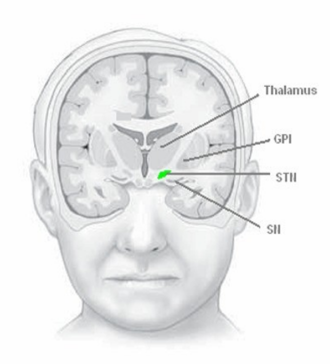


Figure 1. The surgical region of the operation.

Recordings

A majority of the recordings have been conducted at somewhat set times after surgery, 6 months, 1 year and 3 years post operation. There are however post-operation recording material for some participants dating as far as up to ten years, but the emphasis will be put on discussing the post-operative data nearest in time to the surgery.

The material investigated here are syllable repetitions, both SMR (sequential motion rates, i.e. papapapa) and AMR (alternating motion rates, i.e. patakapataka). The participants were asked to:

1. Repeat the individual nonsense syllables /pa/, /ta/ and /ka/ as fast and steadily as they can, i.e. /papapa/, /tatata/ and /kakaka/…

2. Repeat the above mentioned nonsense syllables in a given sequential order, /pataka/, as fast and steadily as they can, i.e. /patakapatakapataka…/

This test is well established in clinical work among speech-language pathologists, referred to and known as a diadochokinetic test for assessing speech.

Having digitalised the original recordings from DAT- and CD-recordings with the software Audacity to an uncompressed format (wave PCM mono) we measured Voice Onset Time (VOT) for each of the produced nonsense repetitions in the different recordings. This was done by using Praat software (Boersma and Weenink, 2009) where the designated syllable repetitions were segmented using Praat TextGrid together with a wideband spectrogram and an oscillogram to carefully study the onset of voicing. The procedure employed for the measuring of VOT was done in accordance with Lisker and Abramson's recommendations, i.e. from the initial burst of the release of the plosive to the first noticeable and regularly occuring period relating to the vocal fold onset vibration. (Lisker L, Abramson AS, 1964).

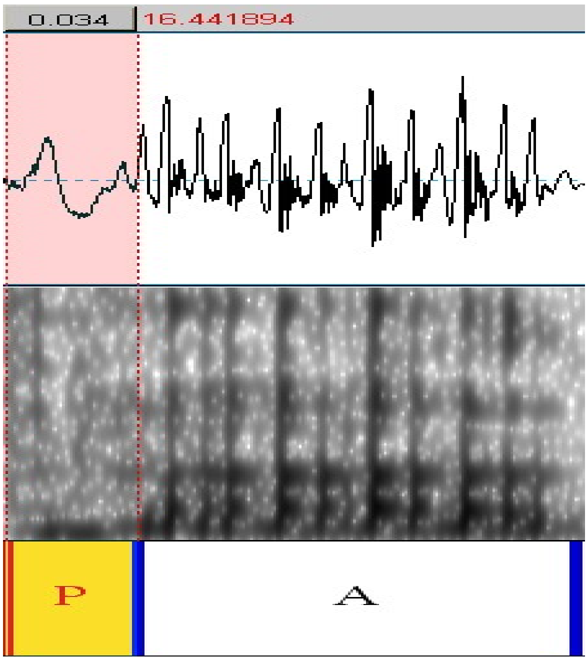


Figure 2. Illustration of VOT measurement.

When the segmentation process had been concluded, time values for VOT, calculated mean and standard deviation per consonant and recording were extracted using a Praat script. The following step included analysing the results to evaluate mean VOT, standard deviation VOT and percent unsuccessfully produced (or unmeasurable) syllables.

Measuring

In some cases the measuring of the initial burst had to be reduced to detecting visual patterns in the sound wave itself, i.e. they were inaudible bursts and had to be established visually by determining a starting point for the burst. This was achieved by looking at the energy levels in the spectrogram as well as by identifying any waveform anomalies or deviations from an expected pattern in the signal. These anomalies were then traced by the aid of the preceding utterances to find any systematic regularities occurring in the speech signal. In several other cases, as anticipated, there were problems with continuous voicing. The patient fails to realise the occlusion of the intended consonant whereby the voicing is maintained throughout the entire SMR- and AMR syllable repetition sequence. In many of these cases the only measurable initial burst was found in the very first of the many CV-segments or when the patient paused to breathe in between segments.

Another problem in relation to measurement issues for [t] was that there were some instances where the patient failed to realise the [t] and instead produced the voiceless apiko dental fricative [θ]. In these instances it was decided to exclude aforementioned segments due to lack of initial burst and also because the intended target sound was ill realised. In some cases of the production of [k] there was a pre-sequence of [x] before the realisation of the explosion itself in which case the measurement was still included, starting from the explosion as intended, ie. the pre-frication was simply ignored. In yet another case it was difficult figuring out if the actual plosive was bilabial or dental as the patient failed to produce the sequential utterance in the right order. Most of the times the patient fell towards producing [t] instead of the intended [p] or [k], which made it extremely hard to determine whether this particular instance, together with it being ill pronounced, was a [p] or [t]. The intended sound however was [p] why the instance was excluded. An interesting observation in another case was when the subject was to repeat the SMR /ta/-repetition and it was done at such a rate that the aural impression was that the target sound requirements were met. However, looking closer at the individual instances it was obvious that the [t] was really produced as an [s]. The rate at which the repetitions were made was so high that on a perceptual level it seemed like the target sound was realised but closer inspection showed that the [t] really lacked an initial burst, making it immeasurable.

Hypotheses

The question for this preliminary study was to investigate whether the surgery itself had an obvious effect on VOT. To get an overview of the data, a fit between VOT mean and time was made in order to represent the progressivity of the disease. This fit included the entirety of data for all consonants and syllable repetitions, both SMR and AMR.

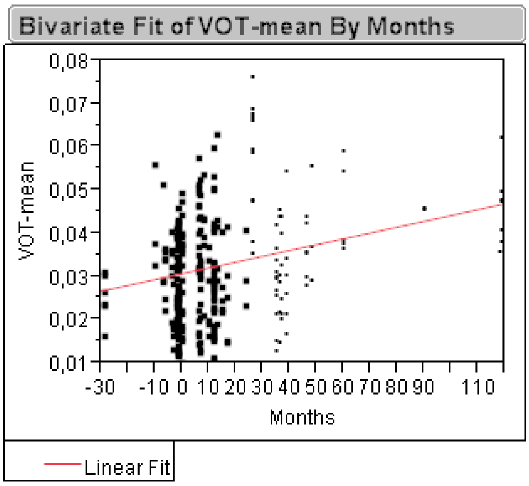


Figure 3. Progressivity of mean VOT (Y) by Months (X).

The figure shows an increase in mean VOT as a general effect of the disease. This gives us a general hypothesis on the development of speech symptoms in this progressive disorder. We can now further create the hypothesis that if the surgery has a positive effect on the VOT, the mean or standard deviation of the VOT should decrease. As we currently have no control group we cannot predict the general progressive increase in mean and standard deviation for VOT for Parkinson patients that have not been subjects for the surgery described here.

Results

Before presenting the following data we wish to point out that the results are preliminary and that the material will be subject to further studies including more detailed analyses.

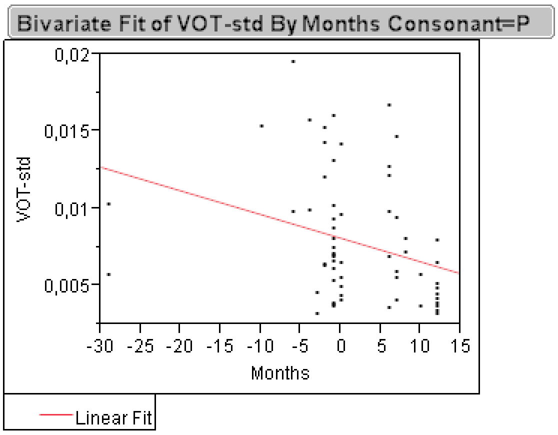


Figure 4. Bivariate fit of VOT standard deviation for /p/ one year post operation.

The figure shows the development of VOT standard deviation for the consonant /p/ calculated post-operatively over one year. VOT standard deviation shows a significant decrease (<0.05) with an explanation value of approximately 9% over the twelve month period post surgery. Apart from the significant decrease represented above in Figure 4 we could see that VOT mean for /p/ also showed a decrease although without any significant explanation value. In contrast, a slight increase for the percentage of unsuccessful productions of /p/ could be seen but similar to VOT mean without any significant changes.

Bivariate fit of VOT mean, VOT std and percent unsuccessful production of /t/ and /k/ one year post operation showed no significant changes. There was however a slight increase for VOT mean and a slight decrease for VOT standard deviation and there was also a noticeable drop in percentage of unsuccessful productions of the consonant /t/. For /k/ we saw an increase in VOT mean and VOT standard deviation but the percentage of unsuccessful productions of /k/ showed a decrease.

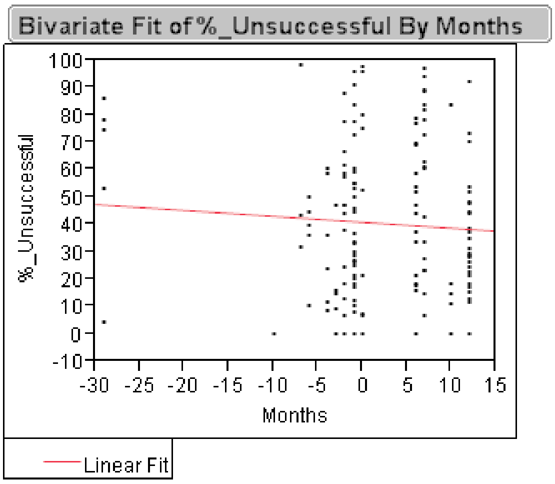


Figure 5. Development of % unsuccessful SMR and AMR syllable productions for the first post-operative year.

The figure shows a decrease in percentage of unsuccessful or unmeasurable SMR and AMR syllable repetitions calculated for the first twelve months after the surgical treatment. The percentage of unsuccessfully produced (or unmeasureable) bursts remain more or less on the same level. No significant changes were discovered post-surgically.

Discussion

These preliminary results seem to point towards no obvious significant effects for the first post-operative year other than a VOT standard deviation decrease for the production of /p/. This seems to correlate well with our hypothesis of a decrease. The fact that the STN/DBS treatment seems to have a significant positive effect on VOT standard deviation for the production of /p/ can be explained by the place of articulation. It is plausible to argue that the articulatory motor functions are affected in some ways positively as in the case of improved motor limb control. Perhaps not to the same degree as motor limb movements, however, speech production requires and is dependent on more fine-tuned and detailed muscular control such as tongue movements. In the light of this it seems reasonable to assume that the bilabial movement would be easier to articulate than /t/ or /k/ where more precise articulatory movements involving the tongue tip are included. Hence we would likely see the positive effects of STN surgery through decreased standard deviation, especially for /p/.

Conclusions

These findings suggest that there should be a decrease in VOT standard deviation for /p/ resulting from STN neurosurgical operation. However, no other significant changes were observed.

Future work

It is important to bear in mind the individual differences for PD patients' varying speech symptoms, which will be further addressed and looked into in future studies. It is also imperative we look at this data on an individual level, including such variables as patient history, age, gender and soforth. Besides looking deeper into the here employed conventional VOT measurement, it would also be beneficial to include VOT ratio measurements in future studies, i.e. where the effect of rate is removed (Fischer et al., 2010). These things combined will no doubt present us with a more detailed and descriptive view of the effectiveness of neurostimulatory STN surgery on VOT for patients with Parkinson's disease.

Acknowledgements

The financial support from ALF-funding and Hjärnfonden is hereby gratefully acknowledged.

References

Benabid, A.L. et al. (2009) Deep brain stimulation of the subthalamic nucleus for the treatmen of Parkinson's disease. Lancet Neurology 2009, 8, 67-81.

Boersma, Paul & Weenink, David (2010). Praat: doing phonetics by computer [Computer program]. Version 5.1.31, retrieved 4 April 2010 from http://www.praat.org/.

Dromey, C., R. Kumar, et al. (2001). An investigation of the effects of subthalamic nucleus stimulation on acoustic measures of voice. 15: John Wiley & Sons--1138.

Gentil M, Pinto S, Pollak P, Benabid AL. (2003). Effect of bilateral stimulation of the subthalamic nucleus on parkinsonian dysarthria. Brain Lang. 2003 May;85(2):190-6. Harel BT et al 2004a. J of Neurolinguistics 17, 439-453.

Harel BT et al 2004b. Brain and Cognition 56, 24-29.

Hoffman-Ruddy B., Schulz G., Vitek, J. and Evatt, M. (2001). A preliminary study of the effects of sub thalamic nucleus (STN) deep brain stimulation (DBS) on voice and speech characteristics in Parkinson's Disease (PD). Vol. 15, No. 1-2, Pages 97-101.

Iulianella, I. (2007). Critical Review: Can sub-thalamic deep brain stimulation (STN-DBS) improve speech output in patients with Parkinson's Disease?.

Fischer, E. and A. M. Goberman (2009). Voice onset time in Parkinson disease: Elsevier.

Lisker, L. and A. S. Abramson (1964). A cross-language study of voicing in initial stops: acoustical measurements. 20: -422.

Plaha, P et al (2006). Brain 129, 1732-1747

Pützer, M. B. W. J. and J. R. Moringlane (2008). Effect of bilateral stimulation of the subthalamic nucleus on different speech subsystems in patients with Parkinson's disease. 22: 973.

Özsancak, C., P. Auzou, et al. (2000). Measurement of voice onset time in dysarthric patients: methodological considerations. 53: Karger Publishers--57.

The Deep-Brain Stimulation for Parkinson's Disease Study Group (2001). Deep-Brain Stimulation of the Subthalamic Nucleus or the Pars Interna of the Globus Pallidus in Parkinson's Disease. 345: 963.

A preliminary study of audiovisual integration of roundedness in front vowels: limitation due to discrepancy in jaw depression

Niklas Öhrström
Department of Linguistics, Stockholm University

Abstract

Audiovisual integration of vowel roundedness was investigated, as the discrepancy in jaw depression increased. The results show that the relative visual impact on perceived roundedness decreases at larger discrepancies. The results may suggest that this tendency would be stronger among acoustically presented [i] than [y]. To verify this, more research with talking heads may be required. The results confirm earlier findings that audiovisual integration doesn't require unconsciousness among subjects about the dubbing procedure.

Introduction

Information about events and object can be mediated through various channels. An approaching car conveys optic information of an object whose imprint on the retina is magnifying. The simultaneous acoustic information is characterized by magnifying amplitude of motor noise as well as relatively high frequency due to Doppler shift. When estimating the position of parts of one's body, one can rely on interacting visual information (if visible), and proprioceptive information.

Perceiving our surroundings involves taking into account information from these various channels and integrate them to match a single concept according to fig. 1. This is highlighted in speech perception, since visual information about the speaker's speech gestures enhances the intelligibility at low S/N ratios (Sumby and Pollack, 1954; Erber, 1969). The integration can take place at an early level of processing, which means that integration precedes the percept, or later, which means that integration merely is part of the categorization into concepts. Fig. 1 does not take any position on the controversy of late or early integration.

Integration of information, as mentioned above, does not require information, mediated through the various channels, to originate from the same source. The ventriloquist effect is one example, where the source of the sound appears to be the same as mediated through the visual signal, although they are different. However, the effect can be reduced if sound and vision are asynchronous, especially in cases where sound precedes the vision (Slutsky and Recanzone, 2001). One further example of illusions is when proprioception is influenced by a discrepant visual signal. This holds as long as the discrepancy is not too large (Warren and Cleaves, 1971).

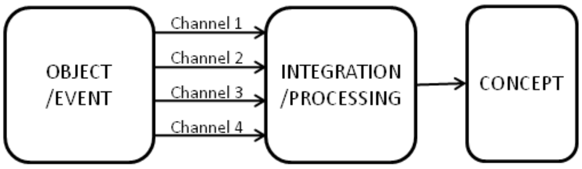


Figure 1. The box to the left represents an object or event. Information can be mediated through various channels. This information is detected by the suitable sense and processed/integrated to match a mental concept.

Another example of cross modal illusions is due to Shams et al. (2000): A number of short presented beeps altered the perceived number of optically presented flashes.

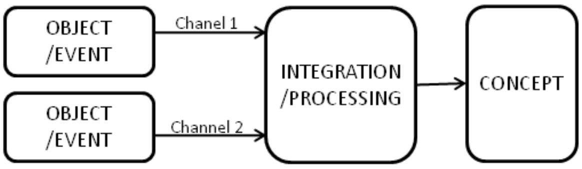


Figure 2. The boxes to the the left represent two different objects or events. Information is mediated through different channels. Information is detected by a suitable sense and processed/integrated to match a mental concept.

Illusions have been useful since they may reveal how the brain interprets and processes information about our surroundings. This process is shown in fig. 2. As remarked earlier, there seems to be a precondition, that objects/events must be similar enough to be integrated.

A well known illusion, relevant to speech perception, is the McGurk effect, (McGurk and MacDonald, 1976): An auditory /baba/ synchronized with a face pronouncing /gaga/ was perceived as /dada/, i.e. a fused percept which neither agreed with the information conveyed acoustically, nor optically. In the reversed situation: An auditory /gaga/ together with visual /baba/, the subjects reported having heard combinations from the two modalities, such as /gabga/ or /gaba/.

The McGurk effect seems to be robust since it persists asynchronies quite well (Munhall et al., 1996). Audiovisual integration of incongruent stimuli does not require the listener to be unaware of the discrepancy (Massaro, 1987). Further proof of the robustness is due to the fact that the McGurk effect still appears, even when a male voice is dubbed onto a female face (Green et al. 1991).

Recent studies have shown that McGurk effect also appears in audiovisual perception of vowels: Traunmüller and Öhrström (2007) showed that an auditory /gyg/ presented together with a visual /geg/ was perceived as /gig/. An auditory /gig/ together with a visual /gøg/ evoked the percept of /gyg/. An auditory syllable /geg/ paired with a visual /gyg/ evoked the percept of /gøg/. Briefly, information about lip rounding is captured by the visual modality, while information about openness is captured by the auditory modality.

The present study aims at investigating the robustness of audiovisual integration of vowel roundedness as the discrepancy in jaw depression increases. It can be assumed that integration will be impeded, at least at large amounts of discrepancy. The perceivers' consciousness of dubbing will also be controlled.

Method

Subjects

16 (8 female and 8 male) subjects volunteered as perceivers. They reported normal hearing and had normal or corrected-to-normal vision. Average age among the subjects was 33.9 years. They were all native speakers of standard Swedish.

Speech material

One female (aged 34 years) served as speaker. The speech material consisted of the open syllables [hV] (V represents a rounded or unrounded front vowel), where jaw depression was to be varied in small steps (i.e. ranging between [i] and [æ]; [y] and [œ]). In order to produce vowels in these small steps in jaw depression, the speaker used different levels of vocal effort (Schulman, 1989). The speaker was asked to produce vowels such as the teeth were clearly visible. During production, the speaker was recorded on video and audio.

Table 1. Auditory, visual and audiovisual stimuli.

Aud	Vis	Aud	Vis	Aud	Vis
[hi]	-	[hi]	[hy]-[œ]	-	[hy]-[œ]
[hy]	-	[hy]	[hi]-[hæ]	-	[hi]-[hæ]
[hi]	[hi]				
[hy]	[hy]				

Each token was thereafter measured and tagged in terms of maximal distance (in mm) between the upper and lower teeth and distance between the corners of the mouth. The video recordings were subsequently focused around the mouth part and dubbed such as an auditory syllable [hi] was combined with a visual mouth pronouncing [hV] (V represents a front rounded vowel). An auditory syllable [hy] was combined with a visual [hV] (V represents a front unrounded vowel). There was a total of 56 unimodal visual stimuli, 4 unimodal auditory stimuli (each presented twice), 4 congruent audiovisual stimuli (each presented twice) and 112 incongruent audiovisual stimuli according to table 1, thus making up a total of 176 stimuli.

Procedure

The subject sat at an arm's length from a computer screen. The session was divided into two parts. In the first part subjects were presented unimodal visual stimuli only. They were asked which one of the nine Swedish long vowels they had perceived through lip reading.

In the second part of the session, the subjects were presented auditory and audiovisual (congruent and incongruent) stimuli. They were asked which one of the nine Swedish long vowels they had heard. For each audiovisual stimulus, the subjects were asked to judge the stimulus as dubbed or not.

Results

Visual presentation

Stimuli presented in unimodal fashion was to be identified as one of the Swedish long vowels. For further analysis, a criterion was set up: The intended roundedness had to be identified correctly in at least 70% of the cases. Two of the visual stimuli didn't meet this criterion and was therefore excluded.

Auditory presentation

Auditory stimuli were presented together with audiovisual stimuli. In total there were four different auditory stimuli, of which two were intended /i/ and two were intended /y/. The roundedness of /i/ was correctly identified in 93.8% and 90.6% respectively. The roundedness of /y/ was correctly perceived in 71.9% and 68.8% respectively.

Audiovisual presentation

The audiovisual stimuli were either congruent or incongruent. The congruent stimuli were all correctly identified.

The incongruent stimuli were analysed based on the relative impact of visually presented rounding. Fig. 3a and 3b show the relative visual impact on perceived roundedness as a function of discrepancy in jaw depression. As can be seen, the relative visual impact on perceived vowel roundedness is impeded, in cases of larger discrepancies in jaw depression. This tendency seems to be stronger for an acoustically presented [i] than for [y].

Due to the fact that natural stimuli were used, distances in jaw depression were not independent from the distance between mouth corners, as can be seen in fig. 4a and 4b. A lower jaw correlated negatively with distance between the mouth corners.

The subjects' ability to detect whether an audiovisual stimulus was congruent or not was investigated. As can be seen in fig. 5, many of the stimuli, where the incongruent visual signal had an impact on perceived rounding, were also perceived as incongruent/dubbed.

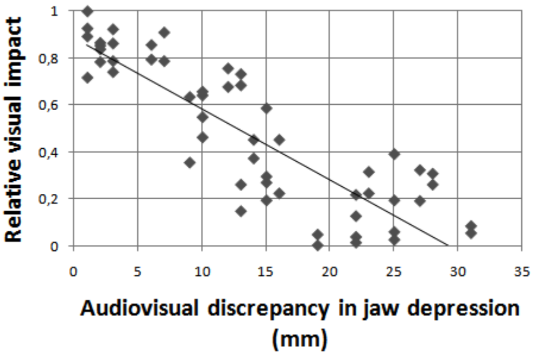


Figure 3a. Acoustically presented [hi], optically presented [hV] (V represents a front rounded vowel). The abscissa relates to the audiovisual discrepancy in jaw depression (in mm). The ordinate relates to the relative visual impact on perception of vowel roundedness.

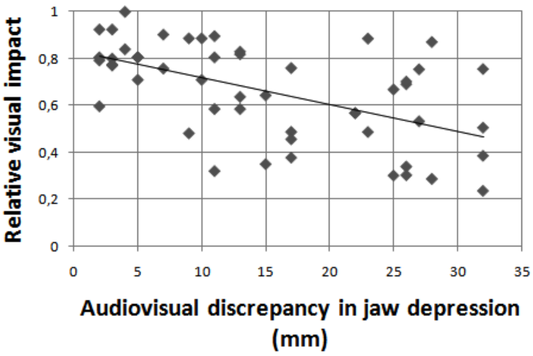


Figure 3b. Acoustically presented [hy], optically presented [hV] (V represents a front unrounded vowel). The abscissa relates to the audiovisual discrepancy in jaw depression (in mm). The ordinate relates to the relative visual impact on perception of vowel roundedness.

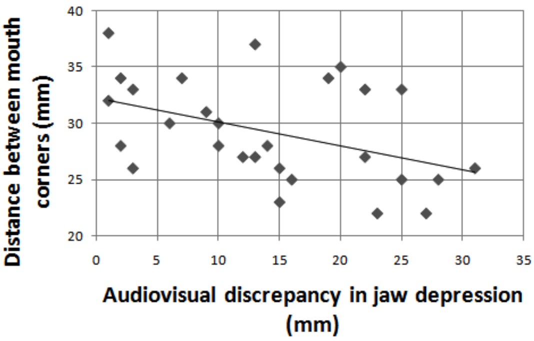


Figure 4a. Dependency between jaw depression and vertical distance between the mouth corners among visually rounded stimuli.

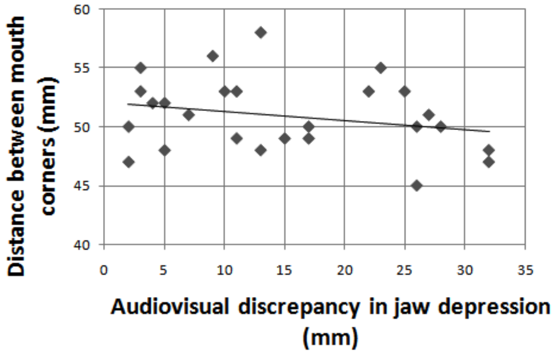


Figure 4b. Dependency between jaw depression and vertical distance between the mouth corners among visually unrounded stimuli.

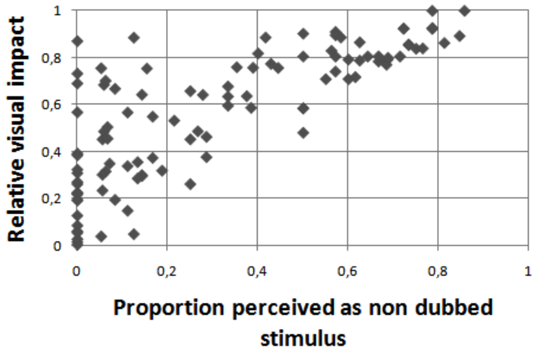


Figure 5. Relation between visual impact on perceived roundedness and detectability of incongruency

Discussion

Analogous to many other illusions, this study show that the visual impact on roundedness is reduced when discrepancy in jaw depression increases. The results may also suggest that this tendency would be stronger among acoustically presented [i] than [y]. This may be true, since the relative visual impact take into account the visual and auditory perception of roundedness in single mode, but such a conclusion may nevertheless be rash, since the two variables, *distance between mouth corners* and *jaw depression* are not independent from each other. Thus, further research is needed. A suggestion would be to proceed with artificial stimuli, using talking heads, in order to fully control these variables.

In this study, the acoustically (unimodally) presented [y] had a perceptual bias towards [i], although it was not reflected in the congruent stimuli. This could be explained by the fact that these auditory stimuli were presented in the same block as the audiovisual ones, of which a majority were incongruent regarding roundedness. However, this potential contextual effect did not have any impact on the auditory perception of /i/. In a future continuation of this study, effort should be put to make /y/ as auditorily rounded as /i/ is auditorily unrounded.

As expected, the subjects' unconsciousness of the dubbing procedure was not a prerequisite for audiovisual integration to occur. This is in line with earlier results by Massaro (1987). This also raises question about visual and auditory separation in other modes of speech perception. As Traunmüller (2006) shows, there are two percepts in speech perception, one auditory (vocal) and one visual (gestural). Will the gestural perception behave in the same way as vocal concerning separation due to discrepancies?

References

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

Green KP, Kuhl PK, Meltzoff AN and Stevens EB (1991). Intergrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Percept. Psychophys.,* 50.9: 524-536.

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

Massaro DW (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Lawrence Erlbaum Associates.

Munhall KG, Gribble P, Sacco L and Ward M (1996). Temporal constraints on the McGurk effect. *Percept. Psychophys.,* 58.3: 351-362.

Schulman R (1989). Articulatory dynamics of loud and normal speech. *J. Acoust. Soc. Am.,* 85.1: 295-312.

Shams L, Kimitani Y and Shimojo S (2000). What you see is what you hear. *Nature.,* 408: 788.

Slutsky D A and Recanzone G H (2001). Temporal and spatial dependency of the ventriloquist effect. *Neuroreport*, 12.1: 7-10.

Sumby WH and Pollack I (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.,* 26: 212-215.

Traunmüller H and Öhrström N (2007). Audiovisual perception of openness and lip rounding in front vowels. *J. Phonet.,* 35.2: 244-258.

Traunmüller H (2006). Cross-modal interaction in visual as opposed to auditory perception of vowels., *Working Papers.* 52. 137-140. Dept. Lingustics, Lund University.

Warren DH and Cleaves WT (1971). Visual-proprioceptive interaction under large amounts of conflict. *J. Exp. Psychol.,* 92.2: 206-214.