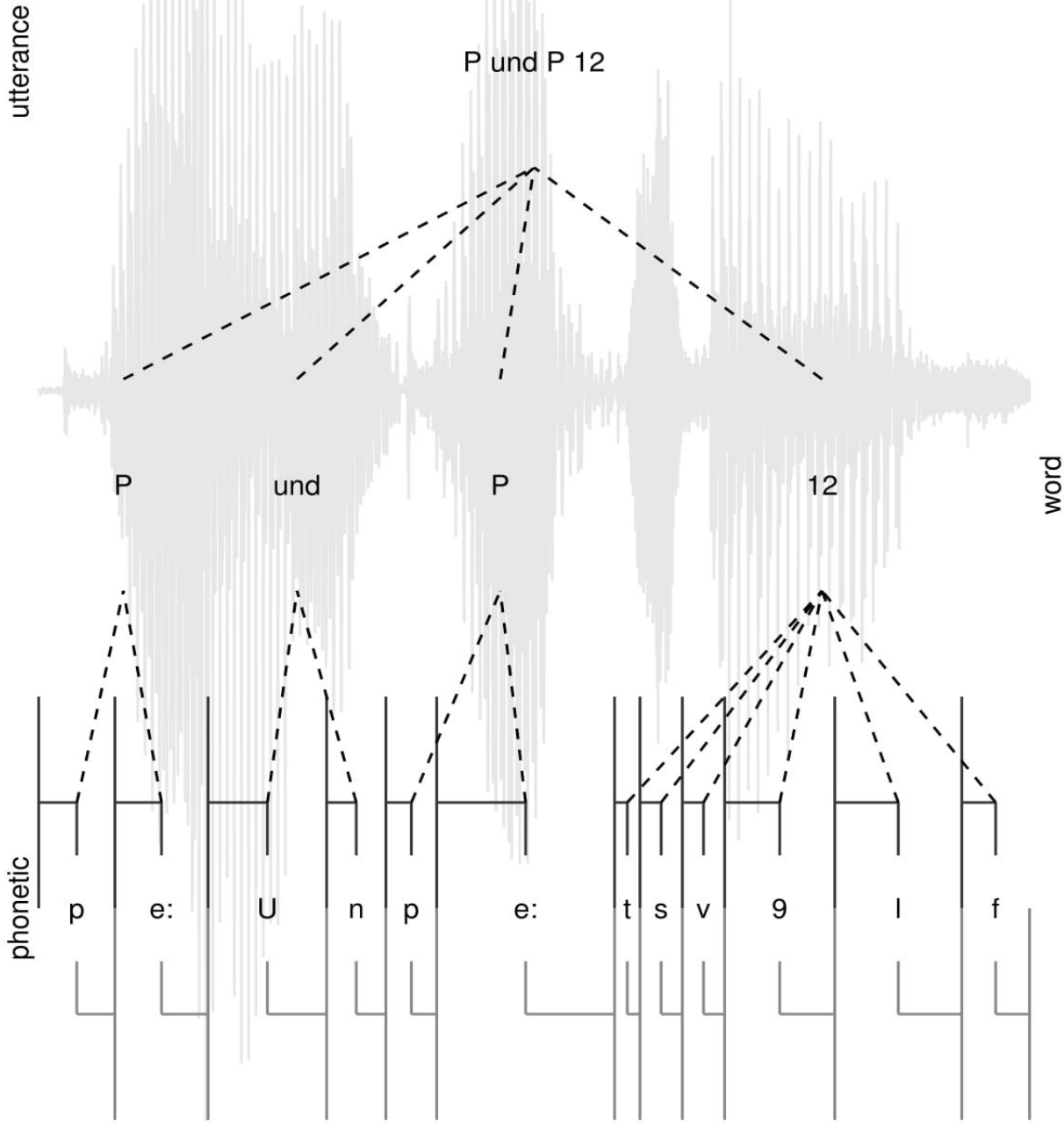
Tagungsband

12. Tagung Phonetik und Phonologie im deutschsprachigen Raum



12. - 14. Oktober 2016

München, Deutschland

Herausgegeben von
Christoph Draxler
Felicitas Kleber

urn:nbn:de:bvb:19-epub-29405-2

Erratum
Die Formatvorlage für die Beiträge im Tagungsband war enthielt einen Fehler. Aus diesem Grund ist die Überschrift für das Literaturverzeichnis auch in englischsprachigen Beiträgen „Bibliographie“.

The author’s kit for the proceedings contained an error. This error caused the heading for the references section to be printed as „Bibliographie“ even in English papers.

Tagungsinformationen:
http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/pundp12/

Redaktion des Tagungsbandes
Katharina Schmidt

VORWORT

[sɛɐvʊs gɾɪas di]

Wir freuen uns sehr Sie in diesem Jahr als Gastgeber der 12. Tagung „Phonetik und Phonologie im deutschsprachigen Raum“ an der LMU München und dem Institut für Phonetik und Sprachverarbeitung begrüßen zu dürfen.

Die P&P hat sich seit dem ersten Treffen 2004 in Potsdam als wichtige Plattform für ForscherInnen und NachwuchswissenschaftlerInnen etabliert, die vor allem in Deutschland, Österreich und der Schweiz an der Schnittstelle zwischen Linguistik und Phonetik oder aber direkt an Fragen zur Phonetik und Phonologie des Deutschen und seiner regionalen Varietäten und Dialekte arbeiten.

Neben Beiträgen aus allen Bereichen der Phonetik und Phonologie wird auf der diesjährigen Tagung der thematische Schwerpunkt "Datenbanken, Korpora und Big Data" mit einer eigenen Session aufgegriffen. Mit diesem Motto sollen gezielt auch technische und informationsverarbeitende Fachbereiche angesprochen (Stichwort *Digital Humanities*) und der Bedeutung von Empirie in Phonetik und Phonologie Rechnung getragen werden. Darüberhinaus spiegelt der Fokus auf zunehmend größere Datenbanken und empirische Methoden die Lehr- und Forschungsausrichtung des Instituts für Phonetik und Sprachverarbeitung der LMU München als auch die aktuellen Forschungsansätze in der Wissenschaft wider.

Der vorliegende *Tagungsband* enthält die vollständigen Konferenzartikel zu vielen auf der Tagung präsentierten Vorträgen und Postern. Die Reihenfolge der Beiträge entspricht nicht der Präsentationsreihenfolge auf der Tagung, sondern erfolgt alphabetisch nach dem Nachnamen des Erstautors.

Wir bedanken uns herzlich bei allen AutorInnen für die zahlreich eingegangenen Beiträge, freuen uns über die große Teilnehmerzahl und wünschen uns allen eine interessante und anregende Tagung.

Felicitas Kleber und Christoph Draxler
München, Oktober 2016

GRUSSWORT

It is a great pleasure to welcome delegates to this P&P Conference in Munich. It is also very pleasing to see how this conference series has flourished since the first P&P in Potsdam in 2004.

This visionary idea of holding these yearly conferences has now culminated in the 12th P&P in Munich with 3 keynote papers, oral presentations spanning two days, as well as three workshops with over 50 poster presentations across so many areas: human speech processing, speech corpora and tools, models of speech production and perception, prosody, first and second language acquisition - indeed many of the major sections that we typically find in the international congress of phonetic sciences.

Perhaps most importantly, the P&P - in a way that is similar to the regular Acoustical Society of America meetings - provides such a valuable opportunity to present research that is still in progress and to discuss and to develop ideas with the help of colleagues that now attend these conferences from such a range of different disciplinary perspectives.

The great progress that has been made in P&P over more than a decade provides a very firm foundation for extending its international reach in the years to come - thereby reflecting the very high standards of scientific research in phonetics and phonology that is carried out in Germany.

I wish everyone a productive and enjoyable time here in Munich.

Jonathan Harrington
Chair of Phonetics and Speech Processing
LMU Munich

SPONSOREN

amazon

münchener universitätsgesellschaft

Gesellschaft von Freunden und Förderern der Ludwig-Maximilians-Universität München e.V.

IPS

INSTITUT FÜR PHONETIK UND SPRACHVERARBEITUNG

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INHALTSVERZEICHNIS

The Karl Eberhards Corpus of spontaneously spoken southern German in dialogues – audio and articulatory recordings

Denis Arnold, Fabian Tomaschek

University of Tübingen, Germany

denis.arnold@uni-tuebingen.de, fabian.tomaschek@uni-tuebingen.de

Abstract

The current paper presents a corpus containing 35 dialogues of spontaneously spoken southern German, including half an hour of articulography for 13 of the speakers. Speakers were seated in separate recording chambers, mimicking a telephone call, and recorded on individual audio channels. The corpus provides manually corrected word boundaries and automatically aligned segment boundaries. Annotations are provided in the Praat format. In addition to audio recordings, speakers filled out a detailed questionnaire, assessing among others their audio-visual consumption habits.

Index Terms: corpus, spontaneous speech, conversation, articulography, German.

The authors contributed equally to the paper and its content.

1. Introduction

Recently, Wagner, Trouvain and Zimmerer [11] have shown that phonetic studies mainly rely on 'scripted' speaking styles, i.e. speech which is recorded in a highly controlled environment in the phonetic lab ($\approx$ 70%). By contrast, 'unscripted' speaking styles i.e. styles without any a priori control, are in the minority ($\approx$ 15%). This is not only the case in acoustic analyses but also in articulatory analyses [10, 9, 12, 13]. This imbalance in speaking styles affects our models about speech production, as large amounts of variations in casual speaking situations are neglected. One possibility to overcome this shortcoming is to record corpora of spontaneously spoken language.

There are plenty of corpora which contain spontaneously elicited speech (see for a extensive list [2]). However, to our knowledge, none of them contain spontaneously elicited speech in dialogues, as is the case in the Kiel Corpus [4, 5] and the GECO corpus [7]. In the Kiel Corpus, spontaneous speech was elicited by mimicking a telephone call as well as discussing non-matching videos with a dialogue partner; in the GECO corpus, speakers talked spontaneously about different topics. However, they did not know each other before the conversation.

The current paper presents the Karl Eberhards Corpus (KEC) of spontaneously spoken southern German elicited in dialogues. Dialogue partners were not instructed on the topic of their conversation, nor did the experimenters interfere with the speakers during recording. In contrast to the GECO corpus, dialogue partners were well acquainted friends. In addition, participants had a conversation of one hour.

At the time of publication, we have recorded 35 one hour long dialogues between two speakers. In addition to pure audio recordings, the corpus contains 13 speakers for which half hour long recordings of articulography were recorded, amounting to roughly 2 hours of speech without pauses. Finally, all speakers provided detailed personal information in the form of a questionnaire.

2. Recordings

We targeted speakers in their mid twenties and early thirties. Most of the speakers were students of the University of Tübingen. We recorded 11 male and 28 female speakers. Their median age was 25 years, with a range from 19 to 33 years. Speakers were compensated for participation either by course credits or €10/hour. We insisted that speakers could only take part in the recording if they were well acquainted with their partners and frequently spoke with him. In this way we ensured that 1. speakers were capable to chat for at least an hour and 2. that they found a common topic to discuss.

2.1. Audio

Recordings were performed in two separated sound-treated recording booths at the *Seminar für Sprachwissenschaften in Tübingen* from 2014 till 2016 and are still going on. During the one hour of recording, speakers were not interrupted. Every speaker was recorded onto an individual audio channel, allowing the investigation of e.g. interruptions and turn taking. The format of the recording for every speaker is:

- Six ten minutes long wave files in the *.wav format (Sampling rate: 44100 Hz, 32-bit float; in case of articulography: 22050 Hz, 16-bit float).
- Six Praat TextGrids for the respective wave files (UTF-8 Encoding, [1]).
- One questionnaire in *.txt format (UTF-8 Encoding)

2.2. Articulography

In a portion of the dialogues we recorded electromagnetic articulography of one of the speakers for one half of an hour. Articulatory recordings were performed by means of the NDI wave articulograph at a sample rate of 400 Hz. The audio sample rate was 22050 Hz, 32-bit float. Figure 1 illustrates the recorded sensor locations on tongue back (TB), tongue mid (TM), tongue tip (TT), upper teeth (UT), lower teeth (LT), upper lip (UL), lower lip (LoL), left lip edge (LL), jaw (J). Apart from the jaw and LL sensor, all sensors were attached along the midsagittal plane. We used three head positions (nasion (N), left/right mastoid (LM/RM)) as reference sensors for correction of head movements. We also recorded a bite plate recording in order to centralize sensor positions.

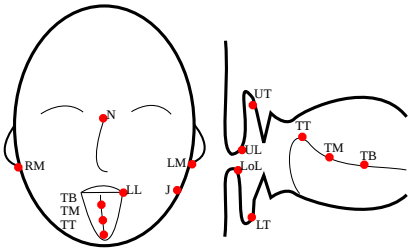


Figure 1: *Illustration of sensor positions. Left: frontal illustration. Right: midsagittal cut through the mouth. See section 2.2 for details on sensors.*

Table 1: *20 most common words and their absolute and relative frequencies.*

Word	Raw Freq.	Rel. Freq.	Word	Raw Freq.	Rel. Freq.
ja	10885	0.043	also	2859	0.011
ich	9650	0.038	der	2805	0.011
und	6507	0.026	halt	2700	0.011
so	6109	0.024	ist	2655	0.011
das	5586	0.022	nicht	2431	0.010
die	4968	0.020	du	2287	0.009
dann	4131	0.016	war	2034	0.008
auch	3874	0.015	was	1887	0.008
da	3155	0.013	hat	1845	0.007
aber	3123	0.012	'ne	1749	0.007

2.3. Annotation

Our focus was to provide precise annotations at the word level. For this, Praat TextGrids were corrected manually. Annotations at segment level were performed by means of a forced aligner [6] within the corrected word boundaries.

3. Questionnaire

After the recording, participants filled out a questionnaire. The questionnaire was in German. An English translation of the questionnaire is provided. Participants were allowed to skip questions, marking the skipped questions by 'keine Angabe' (*not specified*). We assessed answers to the following multiple types of questions, among others:

- **General questions:** Gender, year of birth; educational level; occupation; life situation;
- **Linguistic development:** Native language, proportion of language use.
- **Reading habits**.
- **Consumption** of audio-visual media.

4. Statistics

Table 3 compares the KEC in its current, preprocessed and annotated form with two existing corpora. The statistics for the Kiel Corpus cover the recordings of spontaneously spoken speech. KEC and GECO have roughly the same total number of words. But all corpora differ with respect to their number of unique and consequently in their total/unique word ratio, indicating how "often" a single token was used. In both, KEC and Kiel every token was used roughly 15 times, in the GECO it was used roughly 20 times. The corpora differ with respect of the number of words per minute, which can be regarded to be representative of average speaking rate. The duration of the

Table 2: *20 most common words in the SDEWAC corpus.*

Word	Rel. Freq.	Word	Rel. Freq.
die	0.037	des	0.008
der	0.035	nicht	0.008
und	0.029	für	0.008
in	0.018	auf	0.008
den	0.012	im	0.008
zu	0.011	sich	0.008
das	0.011	ein	0.007
von	0.010	eine	0.007
ist	0.009	es	0.007
mit	0.009	sie	0.007

Table 3: *Corpus statistics. See section 4 for details.*

	KEC	Kiel	GECO
Total words	240299	37257	246621
Unique tokens	15783	2241	12349
Ratio total/unique words	15.2	16.6	19.9
Duration in min.	996.5	214.4	1163.1
Words/min	241	174	212
% rare words	4.8	5.7	4.4

corpora was calculated by excluding the pauses. The Kiel Corpus has the lowest and the KEC has the highest. Furthermore, we calculated the percentage of rare words (frequency of occurrence in corpus < 10). In all the three corpora they are $\sim 5\%$ of the total number of words.

Table 1 shows the 20 most frequent words in the corpus. Note that *ja* is the most common word in the KEC, which is also the case in the GECO and the Kiel Corpus. This is especially striking when comparing these frequencies to frequencies in written corpora, e.g. like SDEWAC [3, 8] (cf. Table 2). Moreover, *'ne* is more frequent than its canonical form *eine* (ind. article. fem.), which is not present at all among the most frequent 20 words. Furthermore, the corpus contains ∼6050 hesitations and ∼3655 laughs, which is interesting for researchers of hesitations and interruptions [14].

5. Distribution

The KEC is planned to be submitted to Clarin-D [2]. In addition to wave files, articulography recordings and Praat TextGrids, the current distribution of the KEC contains R scripts to process the corpus, such as reading in Praat TextGrids, reading in articulography, tagging articulography. Furthermore, a lexicon of frequently reduced forms and their canonical equivalents is provided.

6. Acknowledgments

The authors thank Deniz Cevher, Jan Hoffmann, Ronaldo Rodrigues, Gina Hermann, Mareike Vermehren, Rachel Dockweiler, Verena Heusser and Jessica Viertel for their help creating the corpus. The paper and the corpus were funded by the Alexander von Humboldt Chair awarded to R. H. Baayen.

References

[1] Paul Boersma and David Weenink. *Praat: doing phonetics by computer [Computer program], Version 5.3.41, retrieved from http://www.praat.org/.*

[2] *Clarin-D*. URL: `http://www.clarin-d.de/`.

[3] Gertrud Faaß and Kerstin Eckart. "SdeWaC - A Corpus of Parsable Sentences from the Web". In: *Language Processing and Knowledge in the Web*. Ed. by Iryna Gurevych, Chriss Biemann, and Torsten Zesch. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 61–68.

[4] Klaus J. Kohler. *Labelled data bank of spoken standard German – The Kiel Corpus of read/spontaneous speech*.

[5] Benno Peters. *Die Datenbasis The Kiel Corpus*.

[6] S. Rapp. "Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An Aligner for German". In: *Proceedings of ELSNET goes east and IMACS Workshop*. Moscow, 1995.

[7] A. Schweitzer and N. Lewandowski. "Convergence of Articulation Rate in Spontaneous Speech". In: *Proceedings of Interspeech 2013*. Lyon, 2013.

[8] Cyrus Shaoul and Fabian Tomaschek. *A phonological database based on CELEX and N-gram frequencies from the SDEWAC corpus*. 2013. URL: https://fabiantomaschek.files.wordpress.com/2016/07/tomaschek%5C_corpus%5C_readme.pdf.

[9] Fabian Tomaschek et al. "Vowel articulation affected by word frequency". In: *Proceedings of the 10th ISSP*. Cologne, 2014.

[10] Fabian Tomaschek et al. "Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience". In: *Proceedings of the Interspeech*. Lyon, 2013.

[11] Petra Wagner, Jürgen Trouvain, and Frank Zimmerer. "In defense of stylistic diversity in speech research". In: *Journal of Phonetics* 48 (2015), pp. 1–12.

[12] Martijn Wieling et al. "Investigating dialectal differences using articulography". In: *Proceedings of the 18th ICPHS*. Glasgow, 2015.

[13] Martijn Wieling et al. "Investigating dialectal differences using articulography". In: *Journal of Phonetics* (submitted).

[14] Martijn Wieling et al. "Variation and change in the use of hesitation markers in Germanic languages". In: *Language Dynamics and Change* (2016).

Acoustic correlates of word stress in German spontaneous speech

Grigorij Aronov, Antje Schweitzer

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

grigorij.aronov@ims.uni-stuttgart.de, antje.schweitzer@ims.uni-stuttgart.de

Abstract

The acoustic properties of word stress have been explored in a number of studies. However, there is little research on German word stress, and even less on its realization in spontaneous speech. This paper tests whether parameters that have been found to implement word stress in mostly laboratory speech are also employed in a corpus of German spontaneous speech. Specifically, we consider spectral tilt, syllable duration and pitch. While the results for syllable duration conform with the prevalent finding that stressed syllables have a higher duration, we find no significant effect of pitch. In the case of spectral tilt however, we observe contradicting results, depending on the way we quantify tilt.

Index Terms: word stress, spectral tilt, spectral balance, intensity, pitch, spontaneous speech, syllable duration

1. Introduction

In this paper we explore the effects of word stress on spectral tilt, syllable duration and pitch. Word stress (also called *lexical stress*) denotes a relation of prominence between emphasized and unemphasized syllables of a word. In fixed-stress languages, there are constraints regarding the position of stressed syllables in words; for example, Turkish is considered a language with word-final stress, and Hungarian one with initial stress. Some languages are claimed to have no word stress at all, for instance French [1] or Chinese [2]. German on the other hand, just like English, has variable word stress, the position of which has to be learned together with the pronunciation of a word. Thus, identifying the stressed syllable can aid in word recognition. Speakers are expected to mark stress in production acoustically, and listeners rely on these acoustic cues to detect stressed syllables.

Since the 1950's, a considerable body of research on word stress has identified parameters that are employed in speech perception and speech production to detect or to mark word stress, however with inconsistent results. [3, 4, 5] found that duration, F_0, vowel quality and intensity affect the perception of stress in English listeners. Duration proved to be a stronger cue than intensity. [6] for Dutch also investigated duration, vowel quality, and intensity; however, they calculated the intensity both as overall intensity over the whole spectrum and as the individual intensities in several frequency bands. They confirmed that duration is a strong correlate in the production of stress. In addition they found that increased overall intensity is a poor cue, while increased intensity in the higher frequency bands is a reliable cue. This established that stressed syllables are not characterized by overall greater amplitudes, but that there is a shift in what [6] call the "spectral balance" of stressed syllables, and they explain this shift by greater vocal effort. A follow-up perception study [7] confirmed the perceptual relevance of these parameters.

However, using a rather small English corpus [8] failed to reproduce the results of [6], who had used a Dutch corpus. While [8] provide further evidence for spectral balance differences in vowels produced with and without pitch accents, no difference could be found between stressed and unstressed syllables when pitch accent was not involved. Furthermore, measurements of duration were inconsistent.

[9] compared English dialects regarding correlates of word stress. They built a classifier to predict human judgments of stress in a large corpus of natural speech, comparing acoustic correlates by their predictive power. These correlates were loudness, aperiodicity, spectral slope, several features related to F_0, and a running measure of duration (measuring how long acoustic properties remain stable). While in the literature F_0 is often considered to be a strong cue, [9] provide evidence that F_0 is a weak cue for prominence as neither local F_0 changes, values nor variances were particularly predictive. Also, their results did not support the importance of spectral tilt. Instead, loudness and duration turned out to be the primary indicators with loudness being more important. However, [9]'s measure of spectral slope was different from the way spectral balance was quantified by [6].

In fact, in the literature there are several different approaches to measure relative intensities in the spectrum. Further alternatives include calculating the difference between the amplitude of the first harmonic and the third formant [10], or the difference between the first and the second harmonic [8]. In this paper we tested an alternative where we quantified the spectral balance as the slope of a linear regression line of a spectrum using Praat [11]. This is similar to the measure used by [9]. We will use the term *spectral tilt* to distinguish methods that make use of the regression line fitted to a spectrum from other methods that look at differences between specific regions or points of interest, such as frequency bands, or specific harmonics or formants. We will refer to the latter by the term *spectral balance*.

The aim of this paper is to investigate the implementation of word stress in German on a corpus of natural speech. Most earlier studies are concerned with the analysis of simple, separate, sometimes novel words [10] or distinct sentences. This facilitates research greatly as noise is reduced that way. However, the results mentioned above were not tested on real everyday speech. Moreover, little research can be found on word stress in German. In this paper we will use a large corpus of

spontaneous speech to address both problems. We will analyze distributions of syllable duration, vowel pitch, spectral tilt and spectral balance for stressed vs. unstressed syllables and compare our results to the results found for other languages in the literature.

2. Data and feature extraction

GECO (**GE**rman **CO**nversations) [12] is a database consisting of 46 fully spontaneous dialogs, each with a duration of approximately 25 minutes resulting in 20.7 hours of dialog (two channels), with $\sim$ 250,000 words, making it the largest German database of its kind to the best of our knowledge. It was annotated on the segment, syllable, and word levels by forced alignment. Word stress is annotated as part of the syllable annotation in the forced alignment process, and whether a syllable is stressed or not is determined by way of a lexicon look-up.

For each of the approx. 310,000 syllables in the corpus, we extracted its stress and its syllable duration from the annotations, as well as pitch values at mid vowel using get_f0 from the ESPS software package. For every vowel for which we had more than 5 voiced frames (approx. 41 % out of 310,000), we used Praat to calculate spectral tilt between 0 to 5,000 Hz using the "Report spectral tilt" function on a long-term average spectrum with frequency bins with bandwidths of 100 Hz. Since the values for tilt obtained in this way are usually negative (due to the overall falling tendency of the spectrum), we multiplied all values by -1 to obtain positive values if the spectrum is falling. Thus higher values represent higher (negative) tilt. We also calculated spectral balance for each vowel by taking the absolute difference between the mean intensity in two frequency bands B1 (0-0.5 kHz) and B2 (0.5-1.0 kHz), in analogy to the spectral balance measure employed by [6]. Spectral tilt and spectral balance values, syllable durations, and F0 values were scaled and centered using the *scale()* function in R [13] (packages used are: *plyr v1.8.3* and *lme4 v1.1.11*) to obtain means of 0 and standard deviations of 1 for all parameter distributions. We discarded syllable durations that exceeded 600 ms as potential alignment errors.

3. Statistical analysis

3.1. Spectral tilt

Figure 1 shows the density plots of stressed (green, solid line) vs. unstressed (blue, dashed line) syllables. The x-axis indicates (scaled and centered) tilt values; the y-axis indicates the likelihood of observing these values. It can be seen that stressed syllables are more likely to exhibit greater tilt values than unstressed syllables: the green, solid line is shifted to the right, relative to the blue, dashed line. This is the exact opposite of what the literature suggests: usually stressed syllables are assumed to have a flatter, less steep slope, which would indicate greater vocal effort. In order to test for statistical significance of the influence of stress on spectral tilt, we employ the following methodology: Following common practice (e.g. [14]), we fit two linear mixed models [15], each predicting the acoustic parameter in question. In one model, we include stress as a fixed factor, in the other, we do not. We include random by-speaker intercepts in both models in order to allow for individual means of spectral tilt for every speaker. We then compare the two models by way of an ANOVA. We consider one model to provide a significantly better fit than the other model if the ANOVA indicates that $p < \alpha$, and if in addition its AIC value

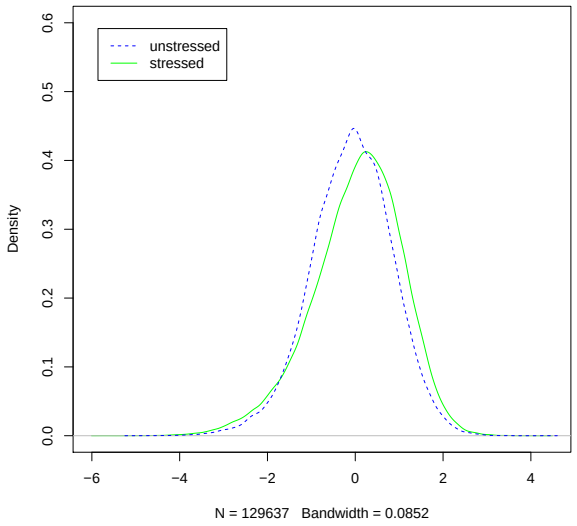


Figure 1: Normalized spectral tilt for stressed and unstressed vowels for all speakers. Stressed vowels have a higher tilt (= steeper slope).

is at least 2 points smaller than that of the competing model (cf. [14, 16]). We want to assume an $\alpha = 0.01$ for the present experiment. Since we will conduct this kind of analysis 4 times, once for each parameter, we use Bonferroni correction and set $\alpha = 0.0025$.

For the present parameter, spectral tilt, we thus compare models (1a) and (1b). The ANOVA determines that the model with stress in (1a) provides a significantly better fit ($\chi^2(1) = 662.72$, $p < 2.2e^{-16}$ and $\Delta AIC = 660$). Therefore we consider the effect of stress on spectral tilt significant.

$$tilt \sim stress + (1|talker) \quad (1a)$$

$$tilt \sim (1|talker) \quad (1b)$$

A potential confound in this analysis could be the impact of function words. It is well known that they are produced in a more reduced way than content words, possibly changing the quality of the vowels we explore for tilt. This way the syllables in function words could be weakened to such an extent that the intensity distribution in the frequency spectrum may become comparable to unstressed syllables (although they are marked as stressed in the lexicon and therefore in our analysis would be counted as stressed syllables). However, excluding all function words based on their part-of-speech tags did not change the graph in any noticeable way, so we omit the corresponding density distribution due to limited space.

3.2. Spectral balance

Figure 2 shows the density plots for the spectral balance of stressed (green, solid line) vs. unstressed (blue, dashed line) syllables. The x-axis indicates (scaled and centered) differences in intensity between bands B1 and B2; the y-axis indicates the likelihood of observing these values. It can be seen that unstressed syllables are more likely to exhibit greater differences between these two bands: the blue, dashed line is shifted to

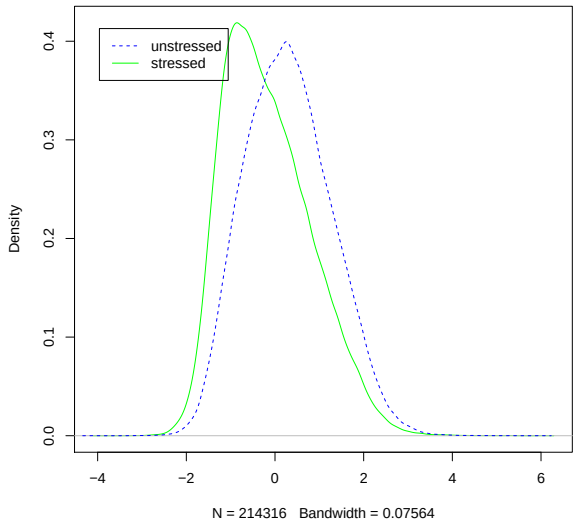


Figure 2: Normalized spectral balance for stressed and unstressed vowels for all speakers. Unstressed vowels (blue line) have greater differences in the intensities in B1 vs. B2 (i.e. unstressed vowels have a steeper spectral slope from B1 to B2).

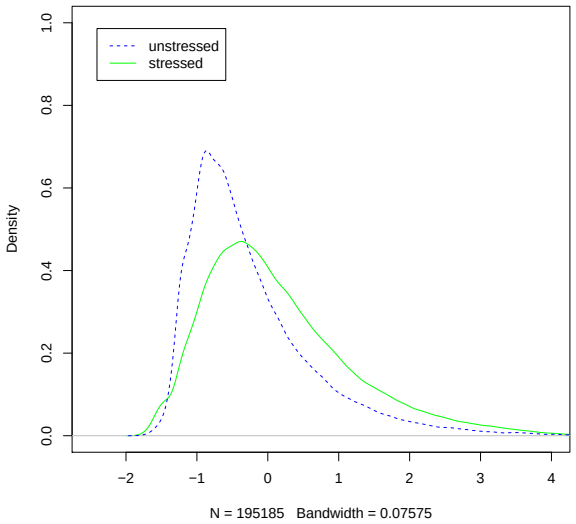


Figure 3: Normalized density plot for duration. Stressed syllables have higher duration and are more widely scattered than unstressed syllables

	Estimate	Std. Error	t value
(Intercept)	0.1727418	0.0050673	34.09
stress	0.0386189	0.0005092	75.85
syl_numphones	0.0436190	0.0002919	149.41

Table 1: Fixed effects of $syl\_dur \sim stress + syl\_numphones + (1|talker)$ with centered *syl_numphones*.

the right, relative to the green, solid line. This confirms the findings by [6]. We use the methodology described above to confirm that the effect is significant ($\chi^2(1) = 17552, p < 2.2e^{-16}, \Delta AIC = 17550$).

3.3. Duration

Regarding duration it is confirmed in Figure 3 that duration is an important cue for stress: Durations of stressed syllables (green, solid line) are shifted to the right compared to unstressed syllables (blue, dashed line). Comparing models with and without stress by an ANOVA confirms that stress affects syllable duration significantly ($\chi^2(1) = 6070.6, p < 2.2e^{-16}, \Delta AIC = 6069$). To make sure that the longer duration of stressed syllables is not simply an effect of different numbers of phones we fit a third model in which we included the (centered) number of phones as an additional fixed effect. This model was found to provide an even better fit ($\chi^2(1) = 21562, p < 2.2e^{-16}, \Delta AIC = 21559$).

The coefficients of that model are indicated in Table 1. They show that indeed the number of phones in a syllable is correlated with the duration of a word: duration increases by about 43 ms for every additional phone. The coefficient for stress of 0.0386 indicates that in addition, and independently of the number of phones, stressed syllables are approx. 39 ms longer on average than unstressed syllables. For the model without number of phones (not shown here), we had obtained a very similar coefficient of 0.0412, i.e. a difference of approx. 41 ms between stressed and unstressed syllables, thus we can conclude that the effect of stress on the duration is preserved even when integrating number of phones as another explaining factor. Similarly, removing function words did not affect the general results.

3.4. Pitch

Looking at the density plots for the pitch parameter we can see that the differences will hardly be significant (see Figure 4). In fact stressed and unstressed syllables have an extraordinarily similar distribution. There is almost no difference between the blue and the green line. An ANOVA using *stress* as a fixed effect and *talker* as a random effect shows that the difference between the models with and without stress is indeed not significant.

4. Discussion & Conclusion

We examined several acoustic parameters that have been suggested to be employed in marking word stress in speech production: For duration, we could fully confirm the wide-spread claim that duration is an important correlate of word stress. The effect was significant, and clearly visible in the density distributions of stressed vs. unstressed syllables.

Regarding spectral tilt, we specifically tried to reproduce the finding by [6] that stressed syllables have a lower tilt than unstressed syllables. On German data the finding could only be confirmed partially. When calculating what we called spectral *balance* by dividing the spectrum into frequency bands and comparing intensities in the lowest two bands, i.e. between 0 to 500 Hz and between 500 and 1,000 Hz, we could confirm that there is a less steep spectral slope for stressed vowels than for unstressed vowels. When calculating what we called spectral *tilt* by linear regression over the frequency spectrum, we found that stressed syllables have in fact a higher spectral tilt than unstressed syllables, which is a very unexpected finding, exactly

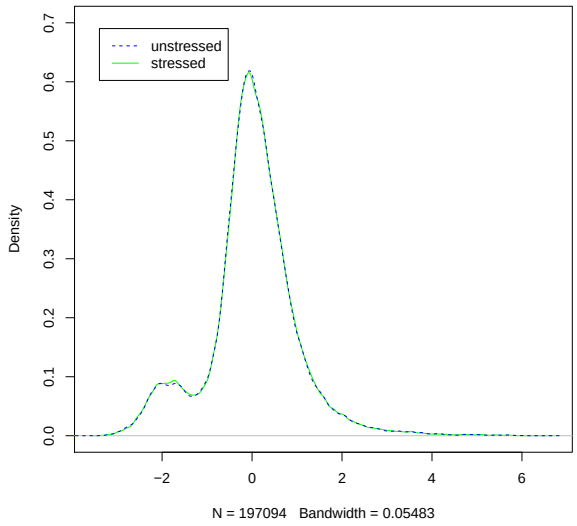


Figure 4: Normalized F_0 for stressed and unstressed vowels for all speakers with function words. Pitch for stressed and unstressed vowels is almost identical.

contradicting the findings by [6]. Several factors were different in our experiment. While [6] had a more artificial setup with participants producing very specific words, we analyzed natural speech, trying to create conditions that are as close to a real world setting as possible. Such scenarios cause a lot of uncontrolled variation in the data, which may be reflected in the seemingly inconsistent results. Also, we would like to point out that spectral balance as defined above looks only at the spectrum between 0 and 1,000 Hz. Preliminary results (not shown here) indicate that the differences are completely lost, and partially reversed in the higher frequency bands. Thus the contradictory results may arise simply as a consequence of looking at different ranges of frequency. This would imply that the wide-spread claim that unstressed vowels have a steeper spectral slope holds only for the lower frequencies.

Finally, while many authors consider pitch to be an important correlate of word stress, we did not find a significant effect. This might be due to the fact that we measured pitch in the middle of the vowel while prevalent literature [10] looks at the maximum F_0 of the vowel. Another possible reason is that other studies can control for pitch accent, due to their experimental design, while in our study we did not have information about the location of pitch accents.

5. Bibliographie

[1] D. C. Walker, *French sound structure*. University of Calgary Press, 2001, vol. 1.

[2] M. J. W. Yip, "The tonal phonology of chinese," Dissertation, Massachusetts Institute of Technology. Dept. of Linguistics and Philosophy, 1980.

[3] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *Journal of the Acoustical Society of America*, vol. 27, pp. 765–768, 1955.

[4] ——, "The dependence of stress judgements on vowel formant structure," *Language and Speech*, vol. 1, pp. 126–152, 1958.

[5] ——, "The Dependence of Stress Judgements on Vowel Formant Structure," in *Proceedings of the Fifth International Congress of Phonetic Science*, 1965, pp. 306–311.

[6] A. M. Sluijter and V. J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[7] A. M. Sluijter, V. J. van Heuven, and J. J. A. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 503–513, 1997.

[8] N. Campbell and M. Beckman, "Stress, prominence, and spectral tilt," in *Intonation: Theory, Models and Applications: Proceedings of an ESCA Workshop*, A. Botinis, G. Kouroupetroglou, and G. Carayiannis, Eds. ESCA ETRW, 1997.

[9] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, 2005.

[10] A. O. Okobi, "Acoustic correlates of word stress in american english," Dissertation, Cornell University, 2006.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2015, September 2015, Edition: 5.4.22. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[12] A. Schweitzer and N. Lewandowski, "Social Factors in Convergence of F1 and F2 in Spontaneous Speech," in *Proceedings of the 10th International Seminar on Speech Production, Cologne*, 2014.

[13] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2015, Edition: 3.2.3. [Online]. Available: https://www.R-project.org/

[14] B. Winter, "Linear models and linear mixed effects models in R with linguistic applications," *arXiv preprint arXiv:1308.5499*, 2013.

[15] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[16] K. P. Burnham and D. Anderson, "Model selection and multimodel inference," *A Practical information-theoretic approach. Springer*, 2003.

Schwa Elision in German Utterances of Bilingual Speakers with Different Ambient Languages during Speech Acquisition

Susanne Beinrucker

Ludwig-Maximilians-Universität München, Institute of Phonetics and Speech Processing, Germany

susanne.beinrucker©campus.lmu.de

Abstract

The aim of this study was to investigate how a reduction phenomenon like the schwa elision is realized by bilingual speakers if it exists only in one of the two languages of the bilinguals. Therefore two simultaneous bilingual speakers of German and Hungarian, who grew up with different ambient languages, were compared. These languages suit this question because only German features the [ə] sound and its elision.

Especially it should be tested whether the ambient language – either German or Hungarian – during the speech acquisition has an influence on the schwa elision in German utterances of the bilinguals. To extend the comparison, also a monolingual native German speaker and a monolingual native Hungarian speaker, who learnt German as a foreign language, were analysed. So the prediction was that the bilingual speaker, who grew up in a German environment, would be closer to the German monolingual in the realization of schwa elision than the bilingual speaker who grew up in Hungary.

The subjects had to read aloud test sentences as an answer to a visually presented question. These recordings were analysed by listening to them and visual inspection of the sonagram. It was observed that the realizations of the bilingual, who grew up in Germany, were almost identical to the ones of the German monolingual regarding schwa elision. This confirmed the prediction. In contrast, the other bilingual realized no complete schwa elision because the schwa was still slightly perceivable but it was not really identifiable as a vowel in the sonagram.
These findings offer a starting point for further research to design experiments especially with spontaneous speech and in consideration of a possible gradual schwa elision.

Index Terms: schwa elision, bilingualism, German, Hungarian

1. Introduction

Bilingual speakers are a very interesting research object. Many theories such as the "Speech Learning Model" [1] for example are based on very advanced foreign language learners or bilingual speakers and assume that languages mutually influence each other. Mostly it is researched how two slightly different phonemes of two languages influence each other or how the speaker is handling a phoneme gap in one language. According to [1] and others bilingual speakers differ from monolingual speakers. This is constrained for segmental elements and phonetic features such as the VOT [2]. Reduction phenomena such as the schwa elision are less investigated, especially in the research of bilingualism.

Generally foreign language learners tend to hyperarticulate and therefore produce mispronunciation (e.g. wrong vowel quality such as [ɛ] instead of [ə]) [3]. Native sound patterns also impact the realizations in a foreign language [4]. [5] found a difference in schwa elision between English learners of German and native German speakers because the learners elided it less often than the native speakers even though [ə] is a sound of the English phoneme inventory. Another study by [6] with Spanish learners of English showed that they rarely elide the schwa and if they elide, they have difficulties to do it in the correct context.

Two languages which suit an investigation of schwa elision in bilingual speakers are German and Hungarian. The [ə] sound is a part of the German phoneme inventory and occurs very often because it is not restricted to a position within a word and every vowel can be reduced to it in fast spoken speech [7]. In Hungarian the schwa [ə] and such a reduction is unknown, which results in difficulties in learning the pronunciation of a language like German. Therefore Hungarian speakers produce the vowel [ɛ] instead of [ə] to articulate the orthographic "e" [8, 9]. A further feature of the schwa is its elision which is another difficulty for learners of German [9]. The schwa elision is a deletion of the schwa sound in unstressed syllables, in which the sonorant carries henceforward the nucleus, e.g. laufen (to run): [laʊfən] → [laʊfn̩] [7, p. 107].

It was already stated that bilingual speakers differ from monolingual ones but it is not exactly proved in the previous research which influence the ambient language during speech acquisition has. It can be assumed that the ambient language is spoken more frequently. In learning a foreign language there are indications that the more frequent a language is used, the more native-like it sounds concerning vowel production [10].

Therefore the aim of this study was to find out whether bilingual speakers differ in schwa elision when they grew up with different ambient languages.

2. Predictions

In this study the languages German and Hungarian are investigated and the main prediction (P1) is that there is an influence of the ambient language during speech acquisition regarding the schwa elision. This means the bilingual speakers who grew up in a German speaking country elide the schwa more often than the bilingual speakers who grew up in Hungary.

Furthermore (P2) the bilingual speakers who grew up in a German speaking country are closer to monolingual German speakers in their realizations regarding schwa elision. As a last benchmark (P3) the bilingual speakers who grew up in

Hungary are closer to foreign language learners of German because they elide the schwa seldomly.

3. Method

In order to test these predictions recordings of speakers were chosen which were made in the project "Form and function of prosodic structure in Hungarian and German" by Felicitas Kleber and Katalin Mády.

3.1. Materials

The stimuli were short German sentences with the target word *Himbeeren* (/ˈhɪmˌbeːʁən/, 'raspberries'). There was a given context question which triggers an accented position of the target word in the answer with two possible word orders:

(Q) "Was hat {Melanie | Verena} {gegessen | gekauft}?"
('What did {Melanie | Verena} {eat | buy }?')

(1) "{Melanie | Verena} hat Himbeeren {gegessen | gekauft}."
('{Melanie | Verena} {ate | bought} Himbeeren.')

(2) "Himbeeren hat {Melanie | Verena}{gegessen | gekauft}."
('It was Himbeeren that {Melanie | Verena}{ate | bought}.')

The variation of the name (Melanie, Verena) and the content verb (essen, kaufen) was added in order to control the attention of the subjects.

One set of stimuli consists of 20 tokens (2 repetitions of the verb "essen" + 3 repetitions of the verb "kaufen" x 2 names x 2 target word positions). In total there were 20 recordings per speaker to analyse.

The recordings were made with the SpeechRecorder system [11]. The question and answer were presented visually.

3.2. Speakers

Four female speakers were chosen to analyse because of their different situations in regard to the acquisition of German. Two bilingual speakers who have acquired simultaneously German and Hungarian were recorded. They differ only in the ambient language during their speech acquisition. One has grown up in Germany (speaker BG; bilingual Germany) with German as ambient language and analogue the other one has grown up in Hungary (speaker BH). Speaker BH has got school education in German from her eighth year on and was living in Germany for 5.5 years at the moment of the recordings.

For comparison a German monolingual speaker was chosen (speaker MG) who grew up in the same area of Germany as speaker BG to control a potential influence of dialect. Additionally a female Hungarian learner of German (speaker MH) was analysed. She had a relatively short duration of learning German (4 years) and hasn't got any pronunciation training nor she has stayed in a German speaking country. So it could be assumed that she is mostly influenced by her first language Hungarian. She was recorded in Budapest, Hungary (the other speakers were recorded in Munich, Germany).

3.3. Labeling and data analysis

Firstly the recordings were automatically segmented with the online device „Multiple Web-MAUS“ [12] in regard to an orthographic and a phonetic (SAMPA) segmentation. With Praat [13] it was marked whether a schwa was produced or not. In doing so three markers were necessary: one for notation of a realized schwa, one of its elision and one for an auditorily perceivable schwa realization which is not clearly recognizable as vowel in the sonogram but changes the rhythmics of the word (with schwa elision the target word *Himbeeren* appears to contain of two syllables but when it is not elided completely, there are three syllables perceivable). The markers are shown in Figure 1.

The data analysis was made with R [14] and the package emuR [15] to be able to do database queries.

4. Observations

The first observation is that all speakers are very constant within their schwa (non-)realizations.

Speaker	Schwa elision labels		
	complete	perceivable	no elision
BG	19	1	0
BH	1	19	0
MG	19	1	0
MH	0	0	20

Table 1: Distribution of labels in the annotation of schwa elision (absolute frequency).

Regarding P1 both bilinguals differ: Speaker BG elided the schwa completely and speaker BH realized the schwa auditorily perceivable and did not elide it completely. This does not confirm P1 exactly even though they differ but there is no clear elision by speaker BH as by speaker BG. However the auditorily perceivable schwa shows a trend towards schwa elision. The bilingual speaker BG is quite identical in the schwa elision with the monolingual speaker MG because both clearly elided the schwa most of the time. This confirms P2. As expected in P3 the foreign language learner MH never elided the schwa. Despite of that P3 is not confirmed because the trend of schwa elision by speaker BH is not comparable to a full vowel, so speaker BH and speaker MH are not similar to each other in schwa elision.

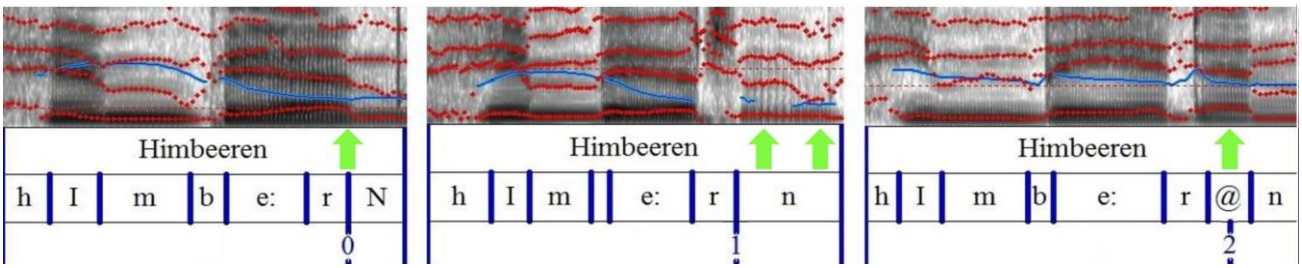


Figure 1: Annotation of schwa and its elision: 0 = schwa elision, 1 = auditorily perceivable schwa, 2 = no schwa elision.

Thus only speaker MH produced vowels at the position of the schwa in the target words, the formants of these vowels were measured with Praat [13]. Speaker MH deviates from the equidistant distances between the formants of a schwa [ə] - these are 500 Hz, 1500 Hz, 2500 Hz [16]. The first formant is located at 820 Hz at the average and quite constantly (standard deviation: 60 Hz). In contrast the second formant is located at 1630 Hz at the average but with a standard deviation of 420 Hz. The third formant has a similarly high deviation (average: 2460 Hz, standard deviation: 350 Hz).

5. Discussion

On the basis of the observations only one prediction can be clearly confirmed, namely that speaker BG and MG are very similar regarding the schwa elision. The other predictions can neither be verified nor falsified because of the gradual schwa elision (especially performed by speaker BH).

Speaker BH differs from the other bilingual speaker BG and the monolingual German speaker MG obviously because of the auditorily perceivable schwa realization which cannot be classified as a vowel in the sonagram but there is still something noticeable and also visible in the form of glottal stops in the sonagram. This is clearly a trend towards reduction but not a completely reduction or elision. Maybe the observations are due to the experimental setting which could result in some kind of hyperarticulation. This cannot be excluded because of the small number of speakers.

The observations do not fit in the "Speech Learning Model" [1] regarding the reduction phenomen because the speaker BG and speaker MG did not differ. But the observations can be seen as a cue of the influence of the ambient language during speech acquisition and are confirming the results of [10] in a broader sense.

The perceptual rhythmics of three syllables in the target word, when the schwa was not completely elided, could be due to the R-realization. As in Figure 1 the left and the middle realization differ in the R-realization: On the left the /R/ is vocalised, which leads to a diphthong but in the middle the /R/ is a fricative. The realization of the diphthong is possible because of a complete schwa elision [17].

The Hungarian learner of German (speaker MH) differs from speaker BH too. She never elided the schwa in the target word which was expected. There is not any equidistance of the formants in the "schwa" realizations and the first and second formant shows rather the vowel quality of a lowered central or front vowel like [ɐ] or [ɛ] especially because of the first formant [16]. This conforms with the finding of [4].

To be able to make evidential statements on the influence of the ambient language during speech acquisition of bilinguals in regard to schwa elision it is necessary to expand the research especially in the number of recorded and analysed speakers. Furthermore it is required to define schwa elision accurately, in particular to a gradual elision, so that there can be determined parameters of auditorily perceivable schwa sounds. The research should be extended to increase the number of target words and to look at different positions in which the schwa can occur but also be elided and how this is realized by bilingual speakers. Therefore a study with spontaneous speech would also be very interesting.

6. Acknowledgements

The author thanks everyone who was involved in the project "Form and function of prosodic structure in Hungarian and German" supported by DAAD and MÖB. Heartfelt thanks to the entire Institute of Phonetics and Speech Processing at the Ludwig-Maximilians-Universität for the great opportunities to study and learn, especially to Felicitas Kleber for her huge support and all the things I have learnt from her.

7. References

[1] J. Flege, "Second-language speech learning: Theory, findings and problems", in W. Strange, *Speech Perception and Linguistic Experience*. Baltimore: York Press, 1995.

[2] C. A. Fowler/V. Sramko/D. J. Ostry/S. A. Rowland/P. Halle, "Cross language phonetic influences on the speech of French–English bilinguals", in *Journal of Phonetics*, 36(4), 649-663, 2008.

[3] U. Gut, " Learner corpora in second language prosody research and teaching", in *Trends in Linguistics Studies and Monographs*, 186, 145-170, 2007.

[4] V. Nagy, "Phonetik im Fremdsprachenunterricht", in *Werkstatt 3*, Debrecen: Kossuth Egyetemi Kiado , 7-38, 2004.

[5] A. Lüdeling/S. Sauer/M. Belz/C. Mooshammer (2016, February 14). "Error Annotation in Spoken Learner Corpora", in *Interpretation*, 2(44) [Online]. URL: http://www.ifcasl.org/docs/L%C3%BCdeling_final.pdf

[6] E. Gómez Lacabex/M. L. Garcia Lecumberri/M. Cooke, "English vowel reduction by untrained Spanish learners: Perception and production", 2005.

[7] R. Wiese, "Phonetik und Phonologie". Paderborn: Wilhelm Fink UTB, 2011.

[8] K. Mády,"Kontrastive Phonetik Deutsch-Ungarisch in Hinblick auf zu erwartende Interferenzphänomene. Festschrift für György Hell", Piliscsaba, 30-51, 2001.

[9] S. Cohrs, "Wirkung und Akzeptanz prosodischer Interferenzen ungarischer Deutschlehrer auf deutsche Muttersprachler und ungarische Germanistikstudenten", in *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 12(2), 2007.

[10] J. E. Flege/C. Schirru/I. R. MacKay, "Interaction between the native and second language phonetic subsystems", in *Speech communication*, 40(4), 467-491, 2003.

[11] T. Kisler/F. Schiel/H. Sloetjes, "Signal processing via web services: the use case WebMAUS", in *Proceedings Digital Humanities 2012*, Hamburg, Germany, 30-34, 2012.

[12] P. Boersman/D. Weenink, "Praat: doing phonetics by computer". Version 5.4.06, 2015.

[13] The R Foundation for Statistical Computing (2015). Programmiersprache. Version 3.2.0, URL: https://cran.r-project.org/bin/windows/base/

[14] R. Winkelmann, "Managing speech databases with emuR and the EMU-webApp", in *Sixteenth Annual Conference of the International Speech Communication Association*,2015.

[15] C. Draxler/K. Jänsch, "SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software", in *Proceedings of LREC*, Lissabon, Portugal, 559-562, 2004. [Version 2.10.16]

[16] K. Machelett, "Das Lesen von Sonagrammen in der Phonetik", Hausarbeit zur Erlangung des Magistergrades an der Ludwig-Maximilians-Universität München, 1994. URL: https://www.phonetik.uni-muenchen.de/studium/skripten/SGL/SGLKap3.html

[17] K. J. Kohler, "Segmental reduction in connected speech in German: Phonological facts and phonetic explanations", in *Speech production and speech modelling*, Springer Netherlands, 69-92, 1990.

Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data

Simon Betz[1,2], *Jana Voße*[1,3], *Petra Wagner*[1,2]

[1]Bielefeld University, Phonetics and Phonology Workgroup
[2]Bielefeld University, Center of Excellence Cognitive Interaction Technology (CITEC)
[3]University of Gothenburg, Department of Philosophy, Linguistics and Theory of Science
simon.betz@uni-bielefeld.de

Abstract

Our overarching research project explores the usability of disfluencies in incremental spoken dialogue systems. This endeavor requires basic phonetic research on disfluencies in spontaneous speech corpora as to define strategies for synthesizing disfluencies in a meaningful way. In this paper, our current research focus lies in an investigation of disfluency-related lengthening as a promising time-buying strategy in synthesized dialogue [1][2]. We base our analyses on the results of a search tool aiming to automatically detect lengthening in spontaneous speech corpora occurring without adjacency to phrase boundaries or other disfluencies, i.e. standalone lengthening phenomena. We analyzed disfluency-related lengthening in the "monomodal" half of the GECO corpus [3], with regard to their context, word class, syllable position and phone type. We then postulate a disfluency insertion strategy for synthetic speech that prioritizes lengthening phenomena based on the results obtained in our study.

Keywords: Disfluencies, Hesitation, Lengthening, Spontaneous Speech, Speech Synthesis

1. Introduction

Disfluencies have become increasingly popular from a speech synthesis perspective [4][5][1]. Especially incremental spoken dialogue systems, that plan and prepare their responses while the interlocutor is speaking, are promising areas of their application [5][1]. One of the reasons for this development is that conversational speech phenomena such as disfluencies can buy valuable time to retrieve content, to facilitate the production of corrections and to signal complexity to the listener.

Disfluencies are manifold in structure and the terminology used to describe them is often ambiguous and varies depending on publication date and perspective. In general, we use the terminology established by [6] and [7] to describe the overarching macro-structure of disfluencies, and refer to the phonetic correlates in the speech signal, such as silent pauses, fillers, or lengthened words, as disfluency elements [1].

In this study, we focus on one particular disfluency element, namely *standalone lengthening*, which we define as a stretch of unexpectedly high segmental duration in an utterance that features no other disfluency elements. For a start, any elastic phone (i.e. one that is prolongable) in any syllable or word can carry the lengthening. However, we hypothesize that there are restrictions as to where lengthening manifests itself. To detect regularities of disfluent lengthening in German is one aim of this study. Are certain word classes, syllable types or phone types preferred?

Lengthening in general appears to be capable of buying valuable dialogue time without being detrimental for synthesis quality [1]. Lengthening occurs by default toward the end of syntactic and intonational phrases. Additionally, overt hesitations containing fillers such as "uhm" are regularly preceded by lengthening [2][8][9]. Standalone lengthening has gotten little to no attention so far, but our position is that if lengthening is both capable of buying time and can do so without being detrimental to synthesis quality, then it is worthwhile considering the synthesis of standalone lengthening. In order to do so, we examine in this study tokens of standalone lengthening extracted from human dialogue data.

We propose a general strategy for the synthesis of hesitation that does reflect human speech planning as described by [10], cited in [8] and provides a good testing environment for standalone lengthening:

1. **Lengthen if possible**
2. Silent pause if issue not solved
3. Insert filler if issue still not solved

When a speaker or dialogue system senses an upcoming production issue, such as end of available, pre-planned speech material, or the anticipation of upcoming complex information that needs more processing time, lengthening is applied if the articulatory buffer still contains suitable material [8]. If lengthening cannot be suitably applied or the planning issue has not been solved during the insertion of lengthening, measures with more severe acoustic impact, such as the insertion of silences or fillers can be taken. On the other hand, if the lengthening successfully bought enough time to solve the issue, fluency can be resumed, resulting in a standalone lengthening on the surface signal.

We hypothesize that standalone lengthening does not occur at arbitrary places and that certain rules have to be paid attention to when synthesizing them. In previous work, we conducted a corpus study based on spontaneous conversational German speech and automatically filtered out standalone lengthening [2]. For this study, the output of this search is annotated and analyzed with regard to its surrounding, word class, syllable position and phone type, thus providing an empirical basis for modeling synthetic hesitation.

2. Methods

2.1. Corpus Data and Lengthening Extraction

This study is based on the GECO corpus [3], a phonemically annotated corpus of spontaneous German speech. We used the first half of it, the "monomodal" condition, where speakers had

no visual contact. One file had to be omitted due to technical issues, yielding 43 files each containing 30 minutes of speech.

The method presented in [2] searches phonemically annotated corpora for places of markedly high phone duration of a z-score of 3 or more, that are not followed by fillers, silences or utterance endings, i.e. noticeable "standalone lengthenings" that are not caused by phrase finality. Z-scores were calculated per phone type and per speaker.

2.800 tokens of lengthening were extracted from this part of the corpus. These tokens fall mainly into three categories: (1) *Disfluent lengthening*, (2) *accentual lengthening*, and (3) *forced-alignment errors*. All tokens were hand-labeled by two annotators for further analyses.

2.2. Inter-annotator Agreement

The two annotators labeled the output phones according to the three main categories. Inter-labeler agreement was tested on a subset of 13 files of the corpus, after a training phase based on four different files from the same corpus.

Agreement was calculated on three categories. The most important one is the distinction between accentuation and disfluencies, where annotators agreed in 98.8% of cases. Furthermore, it was checked how many instances of accentuation or disfluency were only labeled by one annotator, i.e. instances where the other annotator labeled nothing. 92.2% of disfluencies were labeled by both annotators as well as 89.8% of accentuations.

It appears straightforward for listeners to identify disfluency and accent related lengthening. Agreement on the distinction between disfluency and accent related lengthening is also very high, yet it can be seen that not all instances of lengthening can clearly be defined as being of one type or the other. Overall it can be claimed that inter-annotator agreement is sufficient to base further analyses on these annotations.

2.3. Token Frequencies and Errors

In total, 1.000 tokens of lengthening, 75% of them disfluent and 25% accent, were extracted from the first half of the corpus. 1.800 tokens were discarded because of grave forced-alignment errors, or for reasons such as the lengthening being phrase-final and neither disfluent nor emphatic.

About 500 of the remaining 1.000 lengthening tokens still contain minor boundary errors, that are corrected for future analyses, but are not severe enough to discard the tokens.[1] This reveals that even where the search tool outputs the material we're after, forced-alignment shortcomings emerge. We suspect that the unusually high length of these phones troubles the language models the forced alignment works with.

3. Results and Discussion

3.1. Is there "pre-lengthening lengthening"?

Phrase-finality and disfluencies like filled pauses are regularly preceded by lengthening that extends and gradually increases over several phones [2]. The standalone lengthening examined here lacks such a feature. As can be seen in figure 1, no systematic durational variation can be observed in any phones preceding standalone lengthening. Normalized duration means cluster

[1]The 1800 tokens that were discarded for example were tokens labeled as /a:/ but were an entirely different phone. The 500 erroneous tokens that we kept in were ones that contain the right phone, but the boundaries are dislocated by < 20 ms.

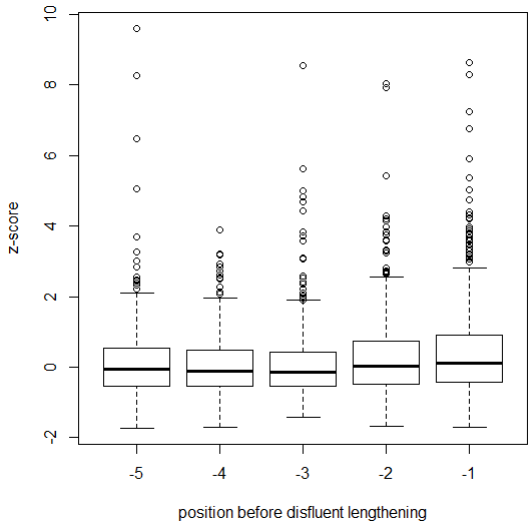


Figure 1: Normalized phone durations preceding lengthening. Positions are indicated in phones relative to position of lengthened phone ($= 0$).

around the mean (0), only the outliers directly before the disfluent lengthening (position -1) hint at a slight increase. Note that due to our filtering method, the phones that follow the -1 position have a z-score of 3 and more, which is a drastic increase from position -1. Pairwise t-tests were conducted on all pairs of adjacent positions, yielding no significant results ($p > 0.1$), thus supporting the hypothesis that there is no systematic increase in duration preceding a standalone lengthening disfluency.

This further supports the hypothesis that lengthening is the first signal of hesitation, i.e. the primary measure that speakers employ before using silent or filled pauses. These lengthening-only hesitations are not introduced by a slowing down of speech rate. Rather, they *are* the slowing down - but in case of successful time-buying, they appear without any further surface disfluency element following. The cases examined here are very likely ones where speakers are able to resume fluency after the lengthening.

3.2. Syllable positions and phone classes

The observation that hesitation begins with lengthening and has no apparent pre-planning beforehand is supported by the fact that disfluent lengthening manifests itself not only in the syllable nuclei but also to a considerable extent in the coda. In contrast, accent related lengthening manifests itself almost exclusively in nuclei (cf. table 2).

In fluent speech, speakers plan beforehand where they place their accent, so it is likely for them to choose vowel nuclei. In case of disfluencies, speakers often do not have the chance to time the "perfect phonotactic moment" to hesitate and resort to coda positions. One reason for doing so might be the vowel quality of the nucleus.

As can be seen in figure 3, the syllable position of the lengthening is related to the nucleus vowel being short, long or a diphthong. If disfluent lengthening occurs in the nucleus, it has a tendency to be realized on long vowels. Much more striking is that when disfluent lengthening happens in the coda, the preceding vowel is likely to be short. This could mean that

Disfluent Word	*English Transl.*	*Frequency*
und	and	61
die	the	35
so	so	27
dann	then	23
in	in	22
ich	I	19
das	the	16
ist	is	16
irgendwie	somehow	15
weil	because	14

Table 1: 10 most frequent words lengthened for disfluency

Function words	Content words
582 (77%)	173 (23%)
Total of words with freq. > 1	With freq. = 1
540 (71.5%)	215 (28.5%)
Content words with freq. > 1	With freq = 1
32 (5.9%)	141 (94.1%)

Table 2: Function and content word distribution within disfluencies

speakers, when they spontaneously have to find the best spot for placing a hesitation, they rather choose an elastic sonorant in the coda than a short vowel nucleus. For accentual lengthening in the nucleus, the vowel types are quite evenly distributed. Accentuation lengthening in the coda is rare, but even so, there is a slight majority of short vowels.

3.3. Word classes

3.3.1. Function words and content words

As noted by [11], lenghtening occurs mainly on function words, such as determiners, prepositions and conjunctions. This is confirmed by our data: we examined word frequencies of the 755 examined disfluencies and table 1 lists the 10 most frequent disfluent words. The same picture extends downward. Apart from auxiliary forms of *sein* "to be", there are no nouns, verbs or adjectives in the top 41 ranks, or in the top 59% of disfluent words. A preliminary word class-tagging was performed, showing that function words add up to 77% of the disfluencies. 28.5% of the words occur only once, and 81.5% of the content words fall into that region. To put it differently, 94.1% of the words that occur only once are content words, while only 5.9% of the more frequent words are content words (cf. table 2). It appears that hesitation indeed preferably manifests itself on function words. The fact that the great amount of lengthened content words occur only once in our data hints to an interpretation that a random target for hesitant lengthening is likely to be chosen, when no suitable function word is available in the articulatory buffer.

3.3.2. Conjunctions

The by far most frequent word on which disfluent lengthening occurs is the conjunction *und* "and". Conjunctions represent the default word class linking two parts of an utterance, so it makes sense for speakers hesitate at this point, in order to facilitate speech planning for the remainder of the utterance and to signal increased cognitive load to the listener, who can in turn infer that it is not the conjunction which is causing the trouble, but

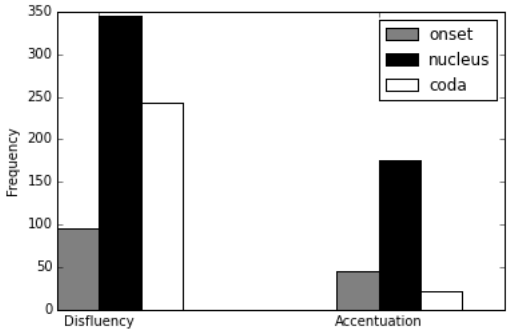


Figure 2: Syllable positions of lengthened phones

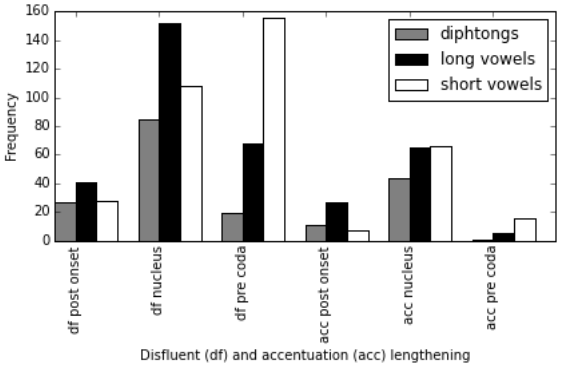


Figure 3: Vowel qualities related to syllable position of lengthening

the material that is about to follow.

3.3.3. Determiners

Quite remarkably, the distribution of the determiners of different gender is extremely skewed. As can be seen in table 1, the female (*die*, 35) and neutral (*das*, 16) determiners are quite frequent, while there are only three tokens of the male one, *der*. German word frequency studies predict these three words to be equally frequent. It can only be assumed that the long vowel in the open syllable of *die* is easiest or most suitable to sustain, whereas the diphtong in *der* might be less so.

4. Conclusions and Outlook

Our study set out to characterize naturally occuring standalone lengthenings in conversational speech as a blueprint for modeling hesitation in synthetic speech as a strategy for "buying time". Our reasoning based on the hypothesis that an unsystematic synthesis strategy to "lengthen anything anywhere whenever needed" may be detrimental for synthesis quality if natural conversational lengthening is characterized by a more specialized pattern, such as centering on function words and containing cues to differentiate between hestiation, accentuation and phrase-final lengthening. Our analyses strengthen this assumption, as annotators were consistently able to differentiate between accentual and disfluent lengthening and we assume that the annotator's ability to do so is at least partly due to the differ-

ent distributions of the two types of lengthenings with respect to phonotactic position, phone type and word class. Of course, other acoustic cues such as accent related pitch excursions may play an additional role and the examination of these cues will be future work.

At the moment, we cannot draw any conclusions with respect to listeners' ability to differentiate between phrase final and disfluency-related lengthening phenomena. For the time being, we assume that many of the lengthenings caused by disfluencies are interpreted as indicating phrase-finality. Many disfluency-related lengthenings occurred together with conjunctions, which can be seen as optimal syntactic position for placing an intonation phrase boundary. Our evidence thus points to a speaker strategy aiming to synchronize hesitation-related lengthening and places of naturally occurring phrase final lengthening. Still, speakers are not always able to match hesitations with such "ideal positions". From a synthesis perspective, it will be of future interest to find out whether hesitation-related lengthening interrupts the prosodic structure of the ongoing intonation phrase which is later resumed, or whether it initiates a new intonation phrase.

To conclude, we postulate that in order to model disfluencies in the synthetic conversational speech, a more sophisticated routine than random lengthening has to be developed. From the insights gained here, the following sequence of steps appears reasonable in order to determine the suitable place for lengthening insertion:

1. Is a function word available in the buffer, preferably a conjunction or determiner?
2. If yes, apply lenghtening[2] on long vowel nucleus of the final syllable.
3. If yes, but nucleus has no long vowel or diphthong, but coda contains a sonorant, lengthen coda instead.
4. If no, apply lengthening as described in the steps before to last syllable of last content word in the buffer.
5. If none of the above locations are available, don't lengthen but proceed to next step in disfluency insertion (silent pause)

[2]The extent of the lengthening will be determined on a follow-up study that tests the acceptability of various lengthening extents with respect to phone elasticity.

5. Bibliographie

[1] S. Betz, P. Wagner, and D. Schlangen, "Micro-structure of disfluencies: Basics for conversational speech synthesis," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015, Dresden)*, 2015, pp. 2222–2226.

[2] S. Betz and P. Wagner, "Disfluent Lengthening in Spontaneous Speech," in *Elektronische Sprachsignalverarbeitung (ESSV) 2016*, O. Jokisch, Ed. TUD Press, 2016.

[3] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pp. 525–529.

[4] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "Modelling filled pauses prosody to synthesise disfluent speech," 2010.

[5] G. Skantze and A. Hjalmarsson, "Towards incremental speech generation in conversational systems," *Computer Speech and Language 27*, 2013.

[6] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[7] E. Shriberg, "Preliminaries to a theory of speech disfluencies," *Ph D. thesis University of California*, 1994.

[8] J. Li and S. Tilsen, "Phonetic evidence for two types of disfluency," in *Proceedings of ICPhS 2015*, 2015.

[9] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "On the generation of synthetic disfluent speech: Local prosodic modifications caused by the insertion of editing terms," in *Proceedings of Interspeech*, 2008.

[10] E. Shriberg, "Toerrrr'is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–164, 2001.

[11] D. O'Shaughnessy, "Timing patterns in fluent and disfluent spontaneous speech," in *International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995*, vol. 1. IEEE, 1995, pp. 600–603.

A tonal analysis of the Limburgian Dialect spoken in Reuver

Sebastian Bredemann

Schießstraße 36a, 63486, Bruchköbel, Deutschland

basti.bredemann@yahoo.de

Abstract

This study aims at providing a tonal analysis of the Central Franconian dialect which is spoken in the Dutch Limburgian town of Reuver. It provides evidence for a binary lexical tone contrast, similar to that found in Roermond and Venlo, between which Reuver is geographically located. The tonal contrast, traditionally referred to as Accent I and Accent II, is only licensed in focused and final syllables. The contrast between both accents is resulting from a single lexical H-tone which is contained in accent II syllables. The interaction of this lexical tone and tones provided by intonation leads to different realizations of these accents relative to the prosodic and intonational context.

Both the dialect of Roermond and Venlo are well investigated, (for example see Gussenhoven 2004) but resemble very different patterns when it comes to the realization of Accent I and II, especially when it comes to lexical and intonational tones competing for the same prosodic edge. This study investigates in which way the dialect of Reuver differs from one or both dialects when it comes to tonal grammar, which leads to interesting insights, given that geographically Reuver is equally distant from both Roermond and Venlo.

As the theoretical framework under which the tonal contours are analyzed serves the tone-sequence model of intonation (Pierrehumbert, 1980). The tonal contours are analyzed in an Optimality theoretic framework and emphasis will be given to the arrangement of the tones from different sources (intonational and lexical) on one single tier.

Index Terms: Central Franconian Tone, Optimality Theory, Tone-Sequence-Model

1. Introduction

This paper aims to investigate the binary lexical tone contrast that is found in the Central Franconian variety which is spoken in Reuver, a town that is located between Venlo in the north and Roermond in the south. The varieties spoken in Roermond and Venlo have been cited and investigated in the literature as representatives of the lexical tone contrast found in the Limburgian area.

The theoretical framework for the analysis is the tone sequence model (Pierrehumbert, 1980). Both lexical tones and tones which are part of the intonation are seen as abstract, independent, grammatical, objects that have their own grammar that gives them a location relative to the prosodified text. The grammar arranges all tones on a single tier, e.g. a linear order for all tones is established in the grammatical output.

The lexical tone contrast (in the literature traditionally referred to as Accent I and Accent II) in the Central Franconian area is binary and restricted to syllables with two or more sonorant moras. I will follow the analysis of Gussenhoven (2000, 2004) and others who assume that syllables with Accent II carry a lexical tone and that the contrast with Accent I results from the fact that syllables with Accent I do not carry a lexical tone. The phonetic implementation of Accent I is derived from intonational tones only. A minimal pair showing this tonal contrast is given in figures (1) and (2). Minimal pairs generally differ in their lexical meaning or in number features, where the singular carries Accent II and the plural Accent I.

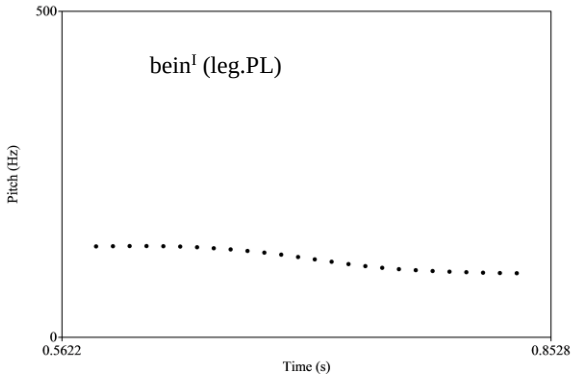


Figure 1: Accent I in Isolation

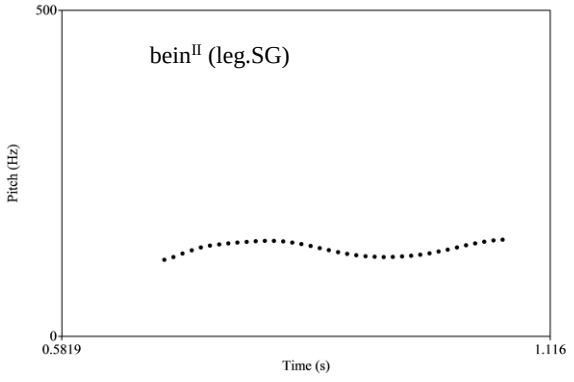


Figure 2: Accent II in Isolation

In the Central Franconian tonal systems the contrast is realized differently depending on the prosodic structure and the adjacent tones that are provided by the intonational component. One effects on the realization of the accents is a neutralization of the contrast in syllables that do not carry a

nuclear accent or are not at the left edge their intonational phrase. Another effect concerns the value of the lexical tone that tends to assimilate to adjacent boundary tones what is observed in the dialect of Cologne. Other properties of the lexical tone contrast will be explained in chapter 3 where the realizations of Accent I and II in different contexts are given a grammatical analysis and are compared with the patterns that have been attested for the dialects in Roermond and Venlo. Due to limitations in space emphasis will only be given to declarative contexts.

The rest of the paper is structured as follows: In section 2 I will explain the experimental design that I used to elicit the realizations of the tonal accents in different contexts. Section 3 illustrates the phonetic realizations of the accents under the tested conditions and will provide a grammatical analysis. Section 4 gives a conclusion.

2. Method

The group of participants included seven people, all of which were native speakers of the Limburgian dialect in the area around Reuver, ranging from ages 17 to 58. They were asked to speak out written Limburgian sentences containing members of tonal minimal pairs.

To elicit the required focal structure the target sentence was embedded in a small context story that provided the right focus-givenness structure to determine the new, focused element in the target sentence. Additionally the element that was supposed to be focused was given in capital letters.

Originally seven contexts were tested. All of them differed in the position of the focused element: The words carrying the Accent I or II were either focused or unfocused. Another condition was the position of the syllable carrying the lexical accent (final vs non-final in the intonational phrase) and the pragmatic context (interrogative vs declarative) resulting in different intonational contexts.

3. The Realizations of the Accents

In this section the different realizations of the accents depending on the tonal environment and prosodic context are illustrated. To illustrate the different shapes of the accents three prosodic and intonational contexts will be discussed. In 3.1 the neutralization of the accent will be illustrated. Section 3.2 discusses the realization of the lexical tone when it is associated to a mora inside a focused syllable in a declarative context. Section 3.3 discusses the linear order of the lexical tone and the ι-boundary tone in the final syllable of the intonational phrase.

3.1. Neutralization of the Contrast

A common property of the dialects in the Limburgian area is to allow the tonal contrast only in syllables that are aligned with the right edge of an intonational phrase or carry the nuclear accent of the sentence. In post-nuclear positions as illustrated in figures (3) and (4), the contour of F_0 does not significantly differ between both members of the minimal pair that shows a contrast when articulated in isolation as in figures (1) and (2).

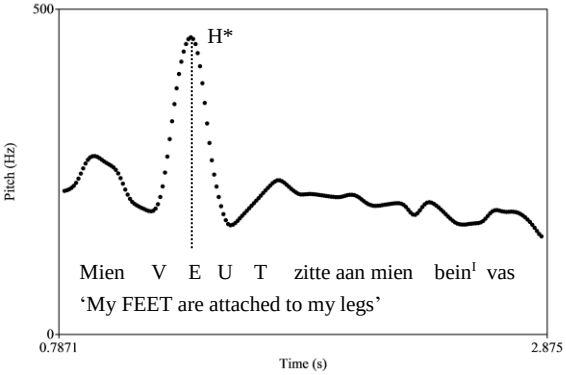


Figure 3: Accent I in non-final, non-focused

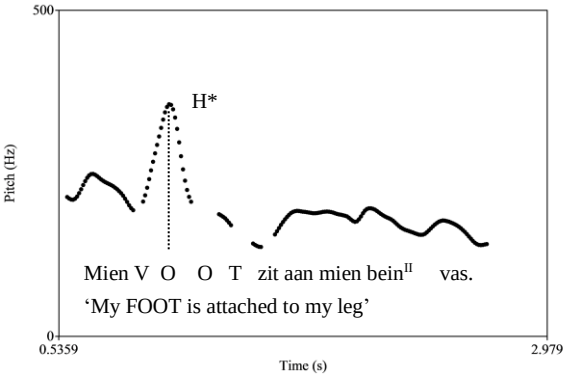


Figure 4: Accent II non-final, non-focused

The contours on the singular (Accent II) and plural (Accent I) of *bein* do not show a contrastive F_0 contour. The lexical tone that is assumed to be lexically represented in the singular form is therefore not realized, because it does not meet the prosodic requirements to be licensed. This deletion of the lexical tone is also found in the neighboring dialects spoken in Venlo and Roermond.

3.1.1. Positional Faithfulness and tonal Neutralization

To account for the implementation of this process in tonal grammar I follow the analysis of Gussenhoven (2000). He attributed the deletion process to a markedness constraint banning lexical tones. This constraint is only dominated by positional faithfulness constraints that demand an output correspondent for underlying lexical tones that are associated to focused or ι-final syllables. These constraints belong to a family of faithfulness constraints militating against the deletion of tones. A general formulation of that constraint family is given in (1):

(1) MAX(T): Tones in the input have correspondents in the output.

The two relevant positional faithfulness constraints, taken from Gussenhoven (2000), are illustrated in (2) and (3).

(2) MAX-FIN(T): Do not delete tones in the final syllable of the IP. (Gussenhoven 2000, 14)

(3) MAX-DTE(T): Do not delete tones in the main stressed syllable in the focused constituent. (Gussenhoven 2000, 14)

For the markedness constraint I assume a context sensitive markedness constraint that militates against adjacent lexical and intonational tones. The number of tonal contrasts in languages strongly depends on the richness of the intonational system. A language that features many intonational contrasts features less lexical contrasts and vice versa. Therefore a constraint as in (4) is assumed:

(4) NOMIX: lexical and post-lexical tones must not be adjacent.

The evaluation of the optimal candidate is shown in tableau (1). MAX-DTE(T) is ineffective and therefore a fatal violation is assigned to the second, which is faithful to the input by retaining the lexical tone.

Tableau 1: *Deletion of the lexical tone*

$(H^*L)_\iota$ H	MAX-DTE(T)	NOMIX	MAX(T)
☞ H* L_ι			*
H* H L_ι		*!	

3.2. The non-final, focused declarative Context

In the dialects spoken in Roermond and Venlo the presence of a lexical tone on the focused syllable leads to a contrast that displays the same properties in F_0- timing in both dialects (see Gussenhoven). For both accents the focused syllable begins on a high F_0-value that is followed by a drop towards a low target. The observed contrast between Accent I and Accent II is the timing of the low target. Syllables with Accent I display a steep fall where the low target is timed within the focused syllable. Accent II has a delayed fall of F_0 where the low target is reached outside the focused syllable. The F_0 contours attested for the variety of Reuver display the same pattern, as illustrated in (3) and (4).

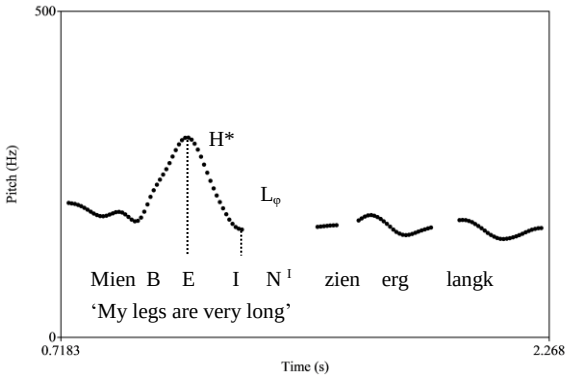


Figure 5: Accent I non-final, focused

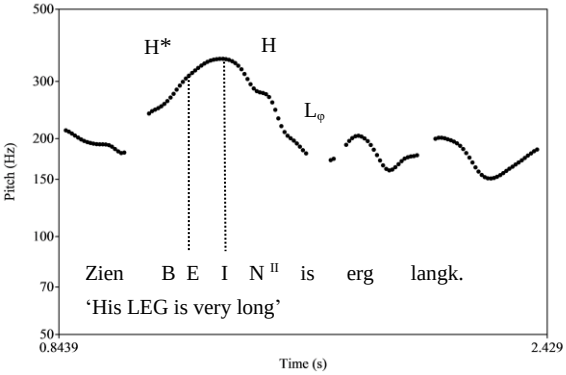


Figure 6: Accent II non-final, focused

On the phonological side the difference in timing results from the presence of the lexical H-tone in accent II syllables that generates an extension of the high F_0-contour in comparison to Accent I. The initial high value on either syllable carrying either accent I or II is assumed to be the target for a high pitch accent H*. In other words on Accent II syllables two tones receive a high target (the lexical H and the pitch accent H*) which leads to a delayed realization of the following low target compared to Accent I. The low F_0 target is the realization of a low boundary tone, most probably a phrasal boundary tone. The fact that this boundary tone receives a target inside focused Accent I syllables I assume it to spread to the second mora of the focused syllable. This is indicated by the line which connects L_φ with the focused word which is indicated in capital letters.

3.3. The ι-final Context

Another phenomenon concerning the interaction between lexical and post-lexical tones is the linear order of these tones in the final syllable of an intonational phrase. In this situation the lexical tone and the boundary tone compete for the same prosodic edge. The boundary tone L_ι seeks to be the final tone within the ι-phrase. The lexical tone on the other hand seeks to be the rightmost tone in the syllable in which it is associated to a mora. In order to be the final tone in their relevant domain no other tone may be specified to follow the lexical or the boundary tone. The final syllable shares its right edge with the right edge of the intonational phrase. In this configuration right alignment can only be satisfied for one of the two tones. The satisfaction of right-alignment of the lexical tone prevents the boundary tone to be aligned with the right edge of the intonational phrase and vice versa.

The graphs shown in figures (7) and (8) illustrate a clear contrast which is found for ι-final syllables between the two accents. The first observation that can be made is that a final syllable does not require nuclear stress for the contrast to be licensed. The realization of the lexical tone in that position leads to the contrast illustrated in figures (7) and (8). The second observation concerns the difference in the fundamental frequency between the two accents. In all the other conditions above the presence of the lexical tone led to a difference in timing of the tonal targets. None of the prosodic and tonal contexts led to a contrast that resulted in the presence or absence of a tonal peak. However in the final, non-focused context the difference between the accents comes from an additional, final rise in the final syllable for Accent II that is

absent in Accent I. The high target in Accent II is the realization of the lexical tone H, which surfaces with its input tonal value. Since the F_0 contour of Accent II ends on that high pitch, it can be assumed that in the dialect of Reuver the lexical tone follows the boundary tone.

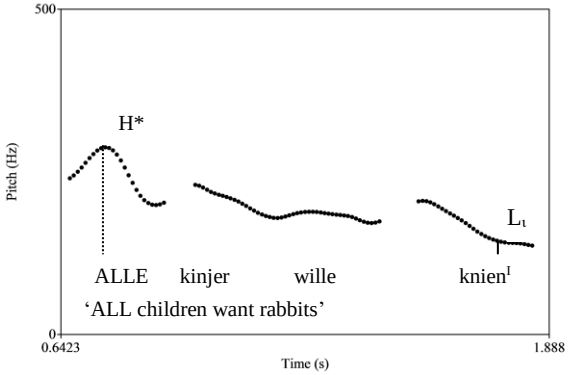


Figure 7: Accent I final, non-focused

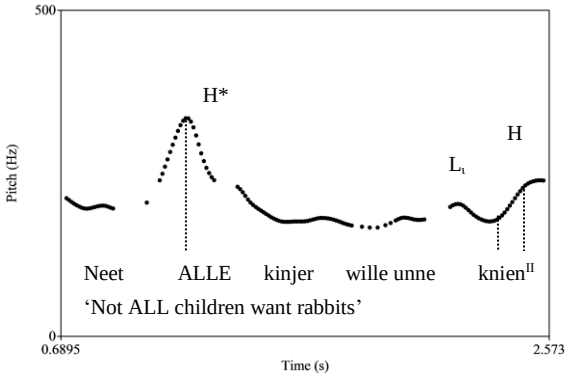


Figure 8: Accent II final, non-focused

3.3.1. Conflicting Alignment Constraints

The conflicting structural requirements for both the lexical tone and the boundary tone to be rightmost in their domain is assumed to be represented by two Alignment constraints, following the analysis from Gussenhoven (2000, 2004). The relevant constraints are shown in (5) and (6), taken from Gussenhoven (2000):

(5) ALIGN($T_ι$,RIGHT): The right edge of a boundary tone coincides with the right edge of the intonational phrase. (Gussenhoven 2000, 16)

(6) ALIGN(T_L,RIGHT): The right edge of lexical H coincides with the right edge of the syllable. (Gussenhoven 2000, 14)

The alignment constraints are responsible for the arrangement of tones in one linear order. The linear order of all tones is the result of the constraints and the property of Optimality Theory to exclude all candidates that violate constraints without satisfying conflicting constraints. Candidates without a linear order specified for their tones are ruled out by any possible grammar, since no constraint demands a non-linear tonal structure. Gussenhoven (2000, 2004) assumes alignment not only to be a family of constraints that make structural requirements based on the relation of ORDER between grammatical entities. He additionally assumes alignment to be a relation between tones that results in the timing of tonal targets relative to one another on the phonetic side. Since both tones (lexical and boundary tone) are assumed to be associated to the sonorant moras of the ι-final syllable,p it is sufficient to understand the alignment constraints as making demands on tonal order only.

The ranking of those constraints as well as the evaluation of the winning candidate is shown in Tableau 1. The first candidate is the most optimal candidate for a ranking in which ALIGN(T_LRIGHT) dominates ALIGN($T_ι$RIGHT). The first candidate violates ALIGN($T_ι$,RIGHT) but is optimal because the second candidate is assigned a fatal violation due to its violation of the higher ranked constraint ALIGN(T_L,RIGHT).

Tableau 2: *Establishment of a linear tonal order in ι-final contexts*

σ]ι L, H	ALIGN (T_L,RIGHT)	ALIGN ($T_ι$,RIGHT)
☞ Lι H]σ]ι		*
H Lι]σ]ι	*!	

4. Discussion

This paper provides evidence for a lexical tone contrast similar to that found in the dialects of Venlo and Roermond. The similarities with the dialect spoken in Roermond were expected and could be proven. This includes the restriction of the lexical tone to focused and final syllables as well as the property of the lexical tone to appear leftmost in its syllable in all contexts.

When uttered in a declarative context the realization of the accents in Reuver are in fact identical to those of the Roermond dialect. This results from the same tonal grammar featuring the same ranking of universal constraints. The constraints introduced in this paper can account for the output realizations of the accents in both a descriptive and explanatory way. Especially the constraint NOMIX can not only account for the deletion of the lexical tone but it also resembles the tendency observed across the phonology of languages that the lack of lexical tonal contrasts corresponds to a rich intonation. It can be discussed however if the relation between lexical and post-lexical tones can be accounted for as straightforwardly as by the formulation of NOMIX. It could very well be that the association to prosodic units also plays a role, in the sense that too many tones lead to too many associations to a prosodic unit which could be militated against by markedness constraints.

5. References

[1] C. Gussenhoven, *The phonology of Tone and Intonation*. Cambridge: Cambridge University Press, 2004.

[2] J. Pierrehumbert, *The phonology and phonetics of English intonation.* Cambridge: MIT dissertation, 1980.

[3] C. Gussenhoven, "The lexical tone contrast in Roermond Dutch in Optimality Theory," in *Prosody: Theory and Experiment*, p. 129-167. Amsterdam: Kluver, 2000. [Online]. Available: http://gep.ruhosting.nl/carlos/roermond.pdf

Investigating the communicative function of breathing and non-breathing "silent" pauses

Aleksandra Ćwiek, Sina Neueder, Petra Wagner

Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
{acwiek,sina.neueder,petra.wagner}@uni-bielefeld.de

Abstract

In this study we investigate the communicative function of two types of "silent" pauses according to breathing behaviour. Taking into account the hypothesis by [1], we expected breathing pauses to be interpreted as a turn-taking cue. A question-answer study in which participants were asked to react to a question as soon as possible was conducted to test this hypothesis. Subsequent analyses of the data revealed that in comparison to non-breathing pauses, breathing pauses are significantly more often interpreted as a turn-keeping signal, which contradicts the working hypothesis. Our results corroborate recent findings by [2].

Index Terms: respiration, breathing, conversation, reaction times.

1. Introduction

Respiration is not only a prerequisite of life, but also of speech itself. Despite its vital importance, breathing is often overlooked in everyday life. This importance and entailed complexity has, nevertheless, sparked a considerable amount of linguistic research on the topic of respiratory behaviour. Connections of respiratory behaviour with e.g. pausing [3], structure of prosodic boundaries [4] and utterance planning [5] have been previously shown (see [6] for an overview). A rather new trend in this field of study revolves around the question if one can assign a communicative function to respiratory behaviour in discourse.

1.1. Theoretical Framework

When it comes to respiratory behaviour during pauses, previous research has classified two types of "silent" pauses: (1) breathing and (2) non-breathing pauses. In [1], apart from making the distinction between the two pauses' types (there: 'holding' and 'trail-off'), the authors propose a possible existing communicative function of respiratory behaviour in "silent" pauses. It is hypothesised, that a 'glottal hold' in a pause indicates a turn-holding event and, likewise, a 'trailoff' a possible speaker turn-yielding event. Recent work addressing the possibility of an existing communicative function of respiratory behaviour has been done in [7, 8, 2, 9], [10] also investigated turn-management and its link to breathing.

1.2. Current Study

In this paper, we examine the effect speculated by [1] using the task developed by [11]: The authors conducted a perception experiment examining turn yielding based on phonetic cues in German. In this experimental design, participants were asked to make a short verbal response to resynthesized questions as soon they thought they were given the floor. In each such interactions there were 1-2 introductory sentences prior to the relevant target question to which a response was to be given. The target questions were sometimes followed by an optional alternative question starting with "or", thus each target question could be either turn-medial or turn-final. In our study, we assume respiratory behaviour to be a relevant cue for floor management, i.e. in turn organisation during discourse. Following the analysis by [1], we therefore expect a non-breathing "silent" pause or "glottal hold" to be turn-keeping and a breathing "silent" pause or "trail-off" to be turn-yielding.

2. Method

To elicit verbal reactions to different types respiratory behaviour, we used the general design by [11], but only manipulated the presence or absence of breathing after turn-medial or turn-final target questions.

2.1. Stimulus Recordings

Nine utterances spoken by one male native speaker of German were recorded. Each utterance consisted of an introductory context and two follow-up questions which are separated by an intermediate pause – symbolised by # – , e.g.:

1. *Ich hab richtig Lust mal wieder aus Deutschland rauszukommen. Was meinst du, würdest du mit mir reisen? # Oder bist du zu beschäftigt?*

   I would really like to get out of Germany for a while. What do you think, would you travel with me? # Or are you too busy?

2. *Der Termin unserer Präsentation wurde vorverlegt. Glaubst du, wir werden das noch zeitlich hinkriegen? # Oder müssen wir den Termin komplett neu legen?*

   The date of our presentation has been moved forward. Do you think we will manage this in time? # Or do we have to set a completely new date?

The recordings were made in two reading conditions. In the first, the speaker was asked to make a pause at the # boundary and to inhale while doing so (= breathing pause). In the second reading condition, he was asked to hold his breath during the pause (= non-breathing pause). The recordings were made in a sound proof booth and the respiratory behaviour was controlled using two respiratory inductance plethysmography (RIP) belts.

2.2. Question-Answer Study

2.2.1. Stimuli

The obtained recordings were segmented and the pauses between alternative questions were manipulated to be of equal length (1.5 seconds). The complete set of consisted of nine utterances in two breathing conditions. In order to prevent the listeners from expecting and waiting for a second question after the pause, the second question was removed for these utterances, thus obtaining an additional number of 18 stimuli. This resulted in a total of 36 stimuli being used in the final question-answer study.

2.2.2. Procedure

20 native German speaking students (undergraduates and graduates) of Bielefeld University participated in the question-answer study. Due to their bilingual background, two of them were excluded from the further analyses. The participants listened to the stimuli via headphones and were asked to respond to each stimulus verbally, i.e. with a short answer, as soon as possible, but making sure not to interrupt the speaker. A familiarisation phase of 10 stimuli was conducted with the participants before their reactions were measured in the main phase. Each session lasted for approximately 30 minutes and took place in a sound proof booth.

3. Results

3.1. Annotation and Measurements

In order to test our hypothesis, two dependent variables were examined: reaction (response) times (as measured from the end of the first question) and too-early-turns, i.e. a response resulting in an "interruption". A too-early turn could occur in a condition where the second question was asked. We expected the reaction times to be longer after non-breathing pauses. We also expected more too-early turns to occur after breathing pauses. The collected recordings were annotated for respose times and too-early turns[1] by two expert annotators. The measurements were subsequently analysed using the statistical software R [12].

3.2. Statistical Analysis

Reaction times were compared with a student's t-test. The results indicate a highly significant difference between response times after breathing and non-breathing pauses ($t(321) = -11.453, p < 0.001$), see Figure 1. However, the direction of the effect is contrary to our hypothesis, i.e. participants wait longer with a response after pauses that contain breathing.

A chi-squared-test was performed to examine potential differences in frequency of occurrence of too-early turns for the two respiratory conditions. For this analysis, only the stimuli containing the second question were taken into account, as interruptions could only occur in this condition. The test revealed a highly significant difference ($\chi^2(1, N = 132) = 66.939, p < 0.001$) between the two breathing conditions, i.e. there are significantly more interruptions or too-early turns after a non-breathing pause (cf. Table 1). These results again support the hypothesis that the presence or absence of breathing during silent pauses functions as a floor managing device. However, the effect is again in a direction contrary to our initial hypothesis.

[1] Like [11], we also annotated late turns, though we did not evaluate those in the analysis.

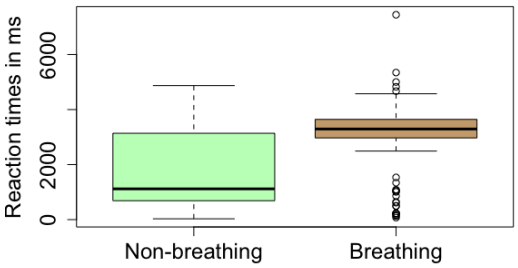


Figure 1: Reaction (response) times of participants after pauses containing breathing or breath-holds (non-breathing).

Response timing	Breathing pause	Non-breathing pause
In time	141	47
Too early	19	113
Sum	160	160

Table 1: Frequencies of occurrence of response timings in the two breathing conditions. Too early responses result in an interruption of the speaker.

4. Discussion

This study investigated the possible communicative function of respiratory behaviour in "silent" pauses as a floor management cue. Using a question-answer study, we analysed speakers' response times and interruptions after "silent" pauses either containing breathing noises or breath holds. We were able to identify differences in the response behaviors after these two different types of "silent" pauses. However, our data indicate an effect of the presence or absence of breathing noise opposite to our initially stated hypothesis, which expected breathing noise to result in quicker attempts to take the floor. Still, the results go hand in hand with recent findings by [2], where the authors found a difference in pause detection thresholds between breathing and non-breathing pauses: pauses containing breathing noise need to be longer to be perceived. In the present study, the participants appear to use breathing noise as a turn keeping cue, i.e. they interpret an audible inhalation as a signal that the utterance will be continued. An examination of the relations of breathing noise with other turn-taking cues is planned for future work.

5. Bibliography

[1] J. Local and J. Kelly, "Projection and 'silences': Notes on phonetic and conversational structure," *Human studies*, vol. 9, no. 2, pp. 185–204, 1986.

[2] M. Heldner and M. Włodarczak, "Is breathing silence?" *FONETIK*, 2016.

[3] F. Grosjean and M. Collins, "Breathing, pausing and reading," *Phonetica*, vol. 36, no. 2, pp. 98–114, 1979.

[4] J. Slifka, "Respiratory constraints on speech production at prosodic boundaries," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[5] S. Fuchs, C. Petrone, J. Krivokapić, and P. Hoole, "Acoustic and

respiratory evidence for utterance planning in German," *Journal of Phonetics*, vol. 41, no. 1, pp. 29–47, 2013.

[6] J. J. Ohala, "Respiratory activity in speech," in *Speech production and speech modelling*. Springer, 1990, pp. 23–53.

[7] J. Edlund, M. Heldner, and M. Włodarczak, "Catching wind of multiparty conversation," in *Proceedings of Multimodal Corpora 2014*, Reykjavik, Iceland, 2014.

[8] M. Włodarczak, M. Heldner, and J. Edlund, "Communicative needs and respiratory constraints," in *Proceedings of Interspeech 2015*. Dresden, Germany: ISCA, 2015.

[9] M. Włodarczak and M. Heldner, "Respiratory turn-taking cues," in *Proceedings of Interspeech 2016*. San Francisco, USA: ISCA, 2016.

[10] A. Rochet-Capellan and S. Fuchs, "Take a breath and take the turn: how breathing meets turns in spontaneous dialogue," *Phil. Trans. R. Soc. B*, vol. 369, no. 1658, p. 20130399, 2014.

[11] O. Niebuhr, K. Görs, and E. Graupe, "Speech reduction, intensity, and f0 shape are cues to turn-taking," in *Proceedings of SIDGIAL 2013 Conference*, Metz, France, 2013, pp. 261–269.

[12] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

PresenterPro: A tool for recording, indexing and processing prompted speech with Praat

Volker Dellwo

Department of Computational Linguistics
University of Zurich
volker.dellwo@uzh.ch

Abstract

Praat (www.praat.org) is a powerful tool for a wide variety of speech analysis and processing tasks. When it comes to recording speech, however, it lacks some fundamental functions that allow a user to prompt a reader with a written list of words or sentences (henceforth: speech prompts) on a screen and index the prompted recordings for further processing. *PresenterPro* - a Praat plug-in - fills this gap. It (a) prompts a reader to read utterances from a screen, (b) automatically indexes the recorded speech prompts in a Praat TextGrid and (c) extracts all recorded speech prompts into individual files. It thus offers an efficient solution for recording large lists of speech prompts. The present paper describes the plug-in and discusses in which situations it is particularly useful.

Index Terms: speech recordings, phonetic field work

1. Introduction

Recording speakers reading long lists of words or sentences (speech prompts) is a typical task that phoneticians or linguists need to perform frequently. Once a list of speech prompts is recorded, the post-recording editing work is usually considerable. Entire recording sessions might need to be listened to again to identify the correct version of a read prompt in case of multiple repetitions and to extract the individual utterances into files. *PresenterPro* is a Praat plug-in that makes this task easier. It presents speech prompts from a list one-by-one on a computer screen and prompts a speaker to read them. Misread items can be prompted again until produced satisfactorily. A single recording of the entire reading session is produced. Based on the time point at which a speech prompt is presented on the screen and the time point the screen is forwarded to the next prompt, a Praat TextGrid file is created in which the read versions of the speech prompts are indexed (start and end). *PresenterPro* further extracts all indexed items into individual files. It also creates a TextGrid for each extracted utterance containing the prompted text. This can be used for segment alignment with forced alignment processing, for example.

2. Why *PresenterPro*?

A variety of tools already exist which perform similar tasks compared to *PresenterPro*. Some of these tools are more sophisticated like *SpeechRecorder* (http://www.bas.uni-muenchen.de/Bas/software/speechrecorder/) or *ProRec* (http://www.phon.ucl.ac.uk/resource/prorec/) to name only two. However, *PresenterPro* hast some advantages that might not be neglected:

- **Platform independent:** *PresenterPro* runs on Praat which is a widely used speech processing and analysis tool that many phoneticians and linguists are familiar with and that is platform independent. Some good recoding tools are only available for particular platforms (e.g. ProRec only runs on Windows).
- **Free choice of recording device:** Speech recording tools typically require the use of an inbuilt recording software to carry out recordings. *PresenterPro* allows the recording to be made on any device. This makes *PresenterPro* very flexible as you can use your high-end portable recording equipment or any other preferred recording software on your computer.
- **Full back-up of entire recording session:** Some speech recording tools save the recording of a prompted utterance directly into a single file. In some cases it might be useful to be able to go back to the recording session where other helpful verbal comments or another version of read prompts can be found. *PresenterPro* leaves you with a recording of the entire recording session.
- **Easy to edit:** *PresenterPro* is written in the easily acquirable Praat scripting language. This means that changes can be applied and *PresenterPro* can be adapted to your own needs. The code does not require compilation.

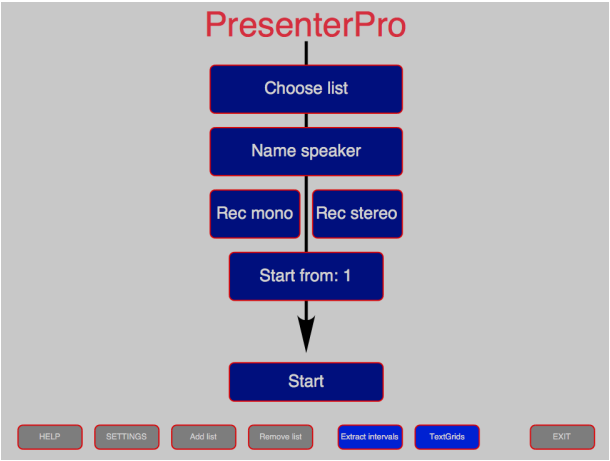


Figure 1: *Interface of PresenterPro. To carry out a recording, users work their way through the options from top to bottom along the arrow.*

3. How does *PresenterPro* work?

3.1. Obtain and install *PresenterPro*

To obtain *PresenterPro* please write an email to the author (volker.dellwo@uzh.ch). You will receive a zip directory which contains a directory named 'PresenterPro'. Use *PresenterPro* with Praat version 6.0 or higher. Install *PresenterPro* as a so called Praat plug-in. This can be done either manually by copying the directory 'plugin_PresenterPro' into your Praat preferences directory (see 'preferences directory' in the Praat help if you do not know where it is) or by executing the script 'install_deinstall_plugin.praat' that is inside the directory.

3.2. Prompting with *PresenterPro*

After installation, execute *PresenterPro* in Praat under 'New > PresenterPro'. The interface as in Fig. 1 appears in a Praat demo window. The basic idea for operating the tool is to follow the arrow along the vertical options starting under the red title 'PresenterPro' from top to bottom:

- **Choose list:** Choose a list of speech prompts from an installed set of lists. For your first usage you may want to use one of the preinstalled demo lists, for example 'demoSentences.txt'. This list contains nine speech prompts. As a side effect these prompts contain instructions about the use of PresenterPro. To add your own list of speech prompts, click on 'Add list' in the bottom menu. Your list must be a plain text file in which each speech prompt is placed in one line (do not include blank lines). To remove a list from your list selection use 'Remove list'. Lists are saved in the plug-in directory (content/lists) where they can also be added and removed manually. If you wish to see a template of a list, refer to the lists in this directory.
- **Name speaker:** Name the speaker or insert an ID. This name will be used to name the TextGrid that *PresenterPro* produces.
- **Choose your recording mode and start recording:** In case you record with Praat, click on one of the 'Rec' buttons to open the Praat Sound Recorder and start the recording inside the Sound Recorder. Choose whether you wish to record in mono or stereo mode. Stereo mode might be particularly useful if you want to record different types of signals simultaneously, like an acoustic and a laryngographic signal, for example. For single-speaker speech-only recordings, stereo is not necessary. Instead of using the Praat inbuilt sound recorder you may also use any other recording software on your computer. Instead or in addition to your computer can also start any other sound recording device at this point. All recordings will later be indexed by the Praat TextGrid that is created during the recording session.
- **Choose starting prompt:** If you do not wish to start your list from the beginning (e.g. after an interruption, see below) you can insert the number of the prompt in your list you wish to start with here.
- **Start prompting:** Start presenting your prompts by clicking on the 'Start' button. IMPORTANT: Immediately when you click the 'Start' button, *PresenterPro* will play a short calibration tone (500 Hz for 500 ms). It is very helpful to record this tone on your recording device as it will make it much easier to align your recording to the TextGrid that *PresenterPro* produces after you finish the recording. Some users reported, however, that they operate PresenterPro without the calibration tone. Bear in mind that it will make the operation more difficult.

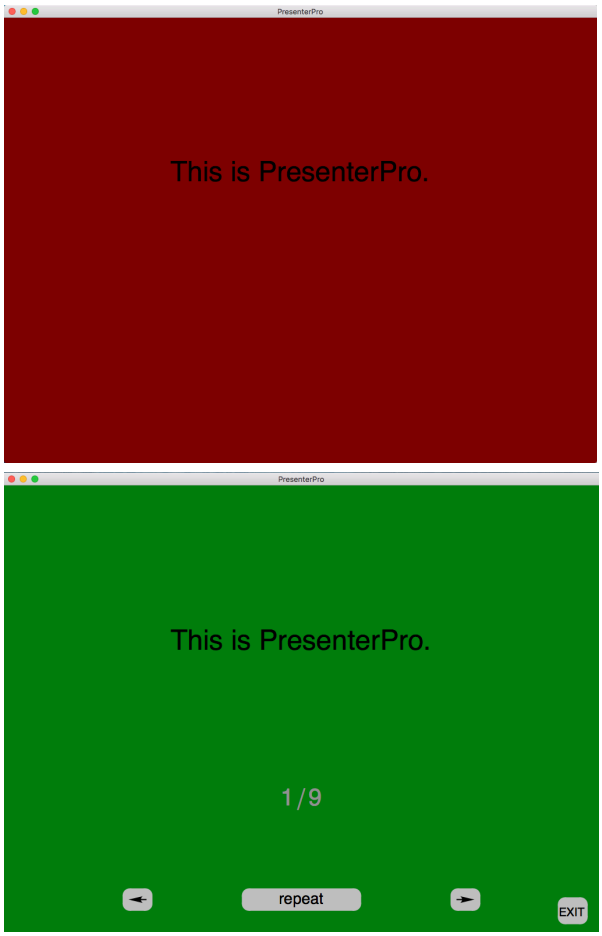


Figure 2: *A speech prompt in PresenterPro is first presented on a red screen (top) which turns to green after a short delay (bottom). The speaker's task is to read the speech prompt as soon as the screen turns green.*

- **Reading prompts:** Together with the calibration tone *PresenterPro* will also present the first speech prompt on the screen (Fig. 2). The screen will first be red and then turn green after about 1 second (delay can be changed under 'Settings'). *PresenterPro* will record the point in time at which the screen turns green and this point will be used to add a boundary in the Praat TextGrid to index the beginning of the prompted speech event (Fig. 3). It is thus essential that readers do not start reading the prompt before the screen turns green. In case the happens, press 'repeat' (see below).
- **Moving forward:** When a speaker has satisfactorily read a prompt, the forward button (arrow pointing to the right) should be pressed without much delay. This indexes the end of the reading process and places a boundary in the TextGrid respectively (Fig. 3). The next prompt will be presented immediately. If you wait for a long time before you press forward you will get long silences after you

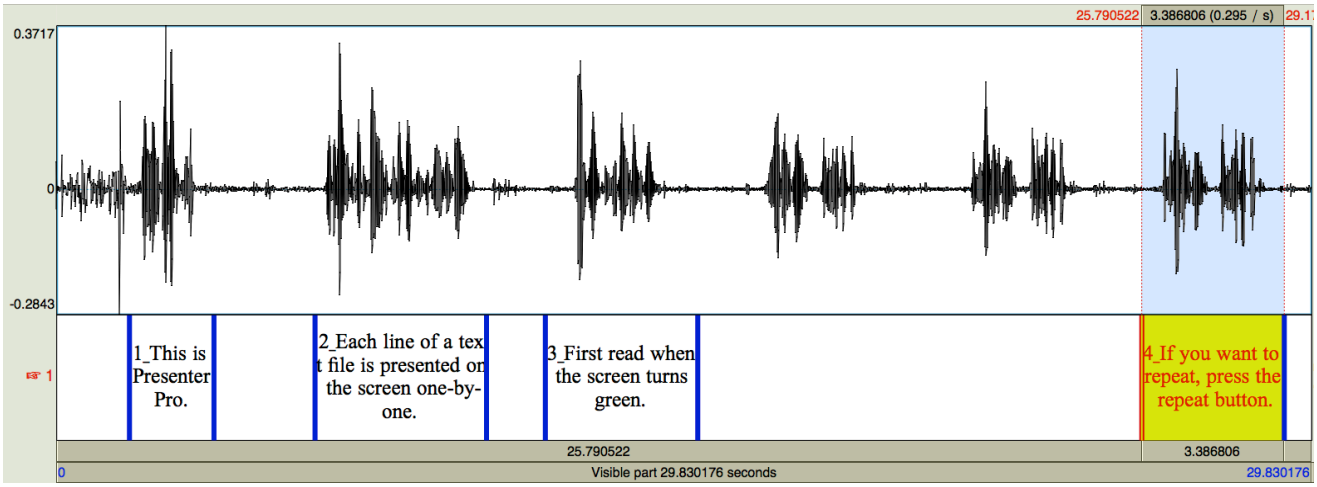


Figure 3: *Example of a recording (waveform) with a an indexed TextGrid. Filled intervals contain the prompted text. Start point of an interval is the point at which the screen turns green, end point is the point the screen is forwarded to the next prompt. Repeated speech prompts are not indexed in the TextGrid (prompt 4 was read three times in total).*

prompted utterance (or speech in case you talk). In this case just click on 'repeat' (see below) and record the speech prompt again.

- **Repeating prompts:** In case a speaker makes a mistake (e.g. reading a prompt incorrectly or reading before the screen turns green) or there is too much delay before forwarding to the next prompt, the prompt can be repeated by clicking on 'repeat'. The screen will turn red again and then green for the speaker to start reading. This process can be repeated as often as necessary. *PresenterPro* will only index the last reading of the prompt in the TextGrid. In Fig. 3 prompt number 4 was read three times in total before it was forwarded to the next prompt. Only the last version was indexed. However, all previously recorded readings of the prompt will be on the recording in case they are needed at a later point.
- **Moving backwards:** If prompting should be repeated from an earlier prompt the arrow pointing left can be used. This will delete all previously indexed boundaries up to the prompt you go back to. When moving forward, boundaries are placed again as usual.
- **Finishing prompting:** After the last prompt is presented or the 'Exit' button is pressed at any point during the prompting, a TextGrid with the intervals of the read prompts will be created and added to the Praat list. This TextGrid is also saved automatically inside the plug-in directory (content/TextGrids). When saved, the TextGrid automatically receives a time stamp. You can open any previously recorded TextGrid from this list. If you want to remove the TextGrids from that list, delete all TextGrids from your TextGrids directory inside the plug-in.
- **Calibrating your recording (with calibration tone):** Terminate your recording and load it in the Praat list of objects either from your external device or from the Praat Sound Recorder by clicking 'Add to list'. Since your recording might be rather long you may want to open it as a LongSound in Praat (see Praat help 'LongSound'). Open the recording in a Sound Editor window (View and Edit) and find the calibration tone in the beginning of the sound. Select the sound from time 0 to the end of the calibration tone as you see it in Fig. 4. Delete the selection (Edit -> Cut). Your sound file now matches your TextGrid as in Fig. 3. Save your sound under the same name as the TextGrid.
- **Calibrating your recording (without calibration tone):** In case you have not recorded a calibaration tone you need to find the starting time of your first prompted utterance in the sound and align it to the corresponding interval in your TextGrid. Measure the time of the onset of the first utterance in your Sound (utterance onset) and subtract the countdown duration from that value (by default roughly 1 sec; see settings). Measure the time of the left boundary of the corresponding interval in your TextGrid (interval onset). Subtract 'interval onset' from 'utterance onset'. The resulting value is the duration of the interval you need to remove from the start of your Sound. Make sure you backup your Sound (Save as WAV file...) in case you need several attempts for doing this (you might start to understand the value of the calibration tone at this point!).

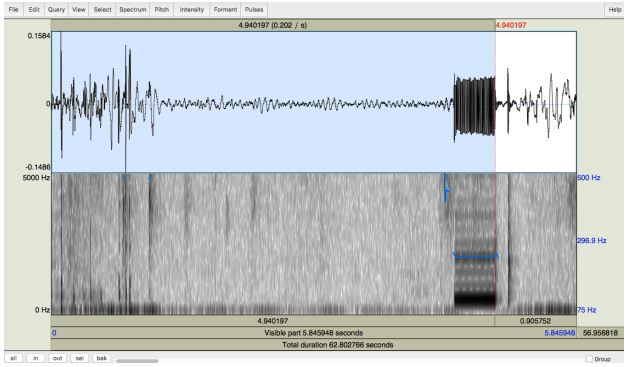


Figure 4: *Calibrating the recording by selecting the interval from time 0 to the end of the calibration tone and deleting the selection. The recording will then be aligned to the TextGrid.*

- **Extract intervals to sound files:** Once you have a Sound or LongSound with a corresponding TextGrid in your list

of objects, select both (and only these two) and click on 'Extract intervals'. Fill in the menu and continue. All your prompted utterances will now be saved to individual files. File names contain the item number. A TextGrid with the interval text will be created along with each sound file. With this TextGrid you can auto-align the segments in your files using Praat or other forced alignment tools.

Other helpful information:

- **Record it all:** After you have started your recording there is no need to rush as recording devices typically record for hours nowadays. If you use the Praat Sound Recorder for your recording, then make sure that your buffer size is big enough to allow long enough recordings (Praat > Preferences > Sound recording preferences...; see 'SoundRecorder' in Praat help menu for details).
- **Interrupting your recording sessions:** For short interruptions (e.g. speaker needs the bathroom), you can leave the recording running in *PresenterPro*. When the speaker is ready to continue you simply repeat the prompt where you left it off. For longer interruptions (e.g. overnight) you will have to stop your recording. You can exit your recording at any point with the 'Exit' button. A TextGrid will be created for the recoding up the prompt you left it off. When you restart the session you can use the 'Start from' option on the main screen (Fig. 1) and enter the number of the prompt after the one you left it off before. IMPORTANT: In case you are planning to interrupt your session you must not randomize your list of prompts (Settings>Randomize list). If you do that, both parts of your recording will contain a random selection of your list, which means that some prompts will be repeated and others will be missing.
- **Correct your alignment before extracting:** If you require start and end points that precisely align to the onset and offset of a recorded utterance you might want to manually align the boundaries in your Text Grid to your recording prior to extracting them.
- **Choose your settings:** Under 'Settings' on the main screen you can choose your line width, font size and line spacing to make sure that your sentences are presented well on the particular screen you are using (please try prior to recording). You can also choose your font and font colour. You can choose to present the prompt number and the total number of prompts under each prompt on the screen. You can also randomize your prompts for each speaker please but bear in mind that you cannot interrupt your recording anymore when you use randomization. You can also set your countdown duration here. Bear in mind that the numbers do only roughly correspond to seconds (depending on your Sound playing preferences).

4. Acknowledgements

This work was supported by the Swiss National Science Foundation (Grant number: 135287) and the Gebert Rüf Stiftung (Grant number: GRS-027/13). Many thanks to Sandra Schwab and Elisa Pellegrino for helpful comments on the draft.

Assessment of Prosodic Attributes in Codec-Compressed Speech

Johanna Dobbriner, Oliver Jokisch, Michael Maruschke

Hochschule für Telekommunikation Leipzig, Institut für Kommunikationstechnik, Germany

{johanna.dobbriner,jokisch,maruschke}@hft-leipzig.de

Abstract

This article deals with the representation of prosodic attributes in coded speech which is less-studied. Common models in speech coding assume that there is no relevant influence of prosodic variation on perceived quality and content of coded speech under suitable operating conditions. Our experiments included a listening test and the instrumental assessment of utterances from an especially constructed test database for the three categories *focus*, *type of sentence* and *situation*. Each category contained at least three different text phrases in several variants, and each original sample was compressed using the fullband-audio Opus codec and the narrowband G.711 codec for reference. The listeners evaluated the overall speech quality and processed a matching task to given prosodic categories. In general, the prosodic variations were well-recognised even when the coding degradation was significant. The overall assessments were comparably high, by achieving an MOS of 4.3 and above on the five-point scale. The hybrid Opus coding method seems to maintain the prosodic features of speech as given in the original reference.

Index Terms: speech coding, prosody, listening test, MOS, POLQA, Opus, G.711

1. Introduction

Audio encoding and decoding is frequently used for the efficient transmission and storage of speech or music data and may influence the perceptual audio or speech quality. It is common to analyse quality from several perspectives, e. g. with regard to certain specific features of speech. This contribution addresses the prosodic attributes of coded speech which are less-studied so far.

We started with a more general assumption from a former study [1] that there is no significant influence of prosodic variation on the quality of adequately coded speech (i. e. using a proper bandwidth/bit rate).

Our experiments incorporated listening tests and a quality assessment using the instrumental POLQA method [2]. The speech data in the prosodic part of these experiments [3] consist of multiple utterances of four speakers (two male and two female) for the three categories *focus*, *type of sentence* and *situation*. Each category contained at least three different phrases in two or more variants. Each original sample was compressed using the Internet-based Opus codec [4] as well as the G.711 a-law codec [5] – two widely used representatives of different coding algorithms, which are generally known to compress speech at high quality in their respective domains.

Our listeners were asked to evaluate the overall speech quality, to correctly match it to a given variant and finally to assess their own assessment difficulty. This test was presented online via the Web platform Percy [6] to 14 participants.

2. Speech Coding Methods

There are a multitude of codecs for speech and audio compression, and new algorithms are continually developed. In general, it is possible to categorise these codecs by the frequency band with four commonly used bandwidths for speech and audio signals. Narrowband (NB) includes speech signals up to 3.4 kHz and is common in landline and mobile telephony. In wideband (WB), audio signals up to 7 kHz are used which is known as high definition (HD) voice. Super-wideband (SWB) comprises an extension of wideband, including signals that contain frequencies up to 14 kHz. Beyond, there is fullband (FB), wherein the full audible frequency range up to 20 kHz is included. This frequency bandwidth is commonly used on CD and known as full HD voice. For each bandwidth there are many codecs available using vastly different methods of audio compression. Among the least complex algorithms is the pulse-code-modulation, a sample-based type of waveform coding wherein each audio signal is sampled at first, and every sample is then quantised and coded. This is a low-delay method with a constant and relatively high bitrate mostly used in landline communication. There are also algorithms which use linear prediction, transform coding or a combination of various coding methods to compress speech at a high quality with short delay and a low bitrate. Among the more frequently used codecs are the G.711 codec which sets a quality standard for NB coding and the Adaptive Multi Rate (AMR) wideband codec in the WB frequency domain (e. g. in modern smartphones).

More recently, the FB Opus codec has become popular used for Web browser-based Real Time Communication (WebRTC). It enables users to easily communicate in full HD voice quality via internet browsers (e. g. Google Chrome). Out of the available codecs we chose two typical ones for our experiments, mainly due to the reported high quality in their respective domains – the G.711 a-law codec and the Opus codec.

2.1. G.711 codec

One of the codecs that has been in use for a very long time, since the 1970s, is the G.711 codec [5]. It was standardised in 1988 by the International Telecommunication Union, Telecommunication Standardization Sector (ITU-T) and is customary for landline communication. The coding method for G.711 is PCM with two different quantisation schemes, the μ-law quantisation which is common for example in the US and the a-law used in European telephony. G.711 has proven to produce high quality speech for NB signals and has an extremely low delay that makes it especially practical for real-time communication, even in other application areas. Nevertheless, its high bitrate of 64 kbps is inconvenient.

2.2. Opus codec

In 2012, the Internet Engineering Taskforce (IETF) standardised the novel audio codec Opus in the RFC 6716 [4]. Opus was designed as a highly-adaptive codec usable for broad application scenarios from speech to music, like VoIP, video conferencing or online gaming. It can work in all four different frequency bands at varying bitrates from 510 kbps to 6 kbps – achieving high quality audio either mono or stereo. Even the coding delay remains relatively bearable – from 2.5 to 60 ms depending on the use case. This adaptability to virtually any scenario is achieved by using a combination of several existing coding algorithms like SILK based on linear predictive coding and the CELT which uses the Modified Discrete Cosine Transform to compress audio signals.

3. Speech Quality Measures

In order to evaluate how well a codec performs in terms of speech quality, it is necessary to perform several tests. Depending on the aim of this assessment, there are various methods to assess speech quality. Quality is generally very subjective – it is the human user who decides to either use or not to use an application, regardless of what a theoretical model may have predicted. Therefore we employed two kinds of speech quality measures. The more important one was the listening test with human participants who separately evaluated a number of speech samples. The second type consisted of instrumental measures. These are algorithms designed to simulate human perception and thereby to predict the quality assessment, a listening test would result in. The listening tests can be distinguished in category/numerical and intelligibility tests.

3.1. Category rating

In the category test, listeners will be exposed to the speech samples and then rate the perceived quality on a scale. There is Absolute Category Rating (ACR), wherein the proband listens to single audio samples and assesses the quality of the samples on a numerical scale like the five-point scale of the Mean Opinion Score which is frequently used – category 5 means excellent and 1 represents very poor speech quality. Afterwards the mean of the scores is determined for each codec and signifies the overall speech quality of that codec. Furthermore, there is Degradation Category Rating (DCR) in which each coded sample is directly compared to the original one, where the listeners assess how much the coded sample is degraded on a numerical scale. A third method is the Comparison Category Rating (CCR) which works similar to the DCR test but the two samples for comparison can be any of them – from two different codecs or codec and original – and the first sample heard is the reference, whereas the second must be rated as better, equal or worse to the reference on a numerical scale.

3.2. Intelligibility test

Intelligibility tests, on the other hand, are necessary in order to determine how clearly coded speech can be understood. The participant listens to a word or phrase and is asked to write down what he/she understood. Afterwards, the percentage of correctly understood speech samples is determined and interpreted as a measure of the intelligibility of speech coded with the tested algorithm. Typical representatives of this method are the Diagnostic Rhyme Test (DRT) and the Diagnostic Alliteration Test (DALT). In both tests, several pairs of words are given and the listeners have to decide which word from the current pair they heard. The DRT uses words with similar endings like “milk” and “silk” whereas the pairs in a DALT begin similarly e. g. “arm” and “art”. In this kind of test only trained listeners take part usually.

3.3. Instrumental assessment

As listening tests are usually time consuming and require much effort and organisation, many developers use instrumental quality measures to evaluate e. g. smaller developments in the coding algorithm or certain aspects and scenarios of speech coding. These instrumental methods use algorithms and perceptual models to approximate the likely results instead of the according listening test. The methods are more cost and time efficient, but also less accurate than real listening tests.

An established method for an instrumental speech quality assessment is POLQA, the Perceptual Objective Listening Quality Assessment defined in P.863 by ITU-T in 2007 [2]. It was designed as an improvement of its predecessor method P.862, also known as PESQ, the Perceptual Evaluation of Speech Quality. The POLQA algorithm requires all speech samples to fit into certain conditions, e. g. regarding the sampling rate. There are two different operating modes to assess NB and SWB signals – both resulting in a MOS-like quality measure.

4. Experiments

To conduct our experiments, it was necessary to focus on certain parameters in terms of the codec selection and the choice of prosodic attributes for the analysis. Furthermore, we needed to generate a number of speech samples representing these attributes and to decide on a certain test design.

4.1. Codec selection

We selected the Opus codec [4] as an example of frame-based hybrid coding, because it is an up-to-date standard published by the IETF, and high-quality audio signals can be encoded and transmitted at comparatively low bit rates. In the case of this study, we decided to encode our original speech samples using the default settings of the Opus codec in WebRTC: a bit rate of 32 kbps and a sampling rate of 48 kHz in FB audio.

As second codec we selected the G.711 a-law codec [5], standardised by ITU-T, and a longstanding representative of sampling-based waveform coding in NB speech. It is often taken as a reference for speech quality in this bandwidth and has been in use since the 1970s.

4.2. Prosodic parameters

As it was our aim to focus on prosody, we chose three specific attributes of speech that can be expressed almost exclusively by prosody in German. Thus, it was possible to use the same wording combined with varying prosodic attributes to express different meanings, e. g. “Es regnet.” *(It’s raining.)* as opposed to “Es regnet?” *(It’s raining?)*. The three selected attributes were *focus*, *type of sentence* and *situation*, representing different categories of the experiment.

Focus contained phrases, where a different word was emphasised in each variation, whereas the category *type of*

sentence consisted of utterances that could be either question or statement, depending on intonation.

In *situation*, there were sentences which could be divided into different phrases which changed the situation expressed in these sentences. One example for this category is

"Max dachte, Lisa kommt aus Hamburg." *(Max thought, "Lisa is from Hamburg.")* or

"Max, dachte Lisa, kommt aus Hamburg." *("Max", thought Lisa "is from Hamburg.")*.

4.3. Speech data

We collected speech samples for each category and recorded four speakers (2 males aged 14 and 18, two females, 23 and 46) who uttered the variations and repetitions. In the end, we were able to choose from 220 utterances.

For the tests, we selected three sentences in two variations each for the category *type of sentence*. Four phrases in two or three variants were chosen for *focus* and another four sentences with two variations per phrase were used in the category *situation*.

4.4. Test design

The experiments incorporated both, listening test and quality assessment using the instrumental POLQA method [2]. While the POLQA testing required nothing else than representative samples, the listening test needed to be efficiently designed for human participants. Consequently, we used the Web-based platform Percy [6], an adequate form of distribution for this experiment.

Our listeners were asked to first evaluate the overall speech quality of the current sample, then to match it to one out of several given variants and finally to assess their own difficulty of choosing one variant. Within the categories, original and coded samples were presented in random order.

For the participants, the only difference between the categories was their choice of variants to match. In *type of sentence*, they had to choose between *question* and *statement*, in *focus* there were as many choices as there were words in each utterance, for example:

o Alle o Jungen o spielen o Fußball.

In *situation*, we drew two images per sentence, visualising both possible variants of one phrase with short descriptions below them, from which the listeners had to choose the one they heard, e. g.: "Das Schiff verließ den Hafen nicht ohne Benachrichtigung des Kapitäns."*(The ship didn't leave the harbour without notification of the captain.)*. In this example, either the ship did *not* leave and the captain wasn't notified (as shown in Figure 1) or the ship *did* leave, but not until the captain knew about it, which can be seen in Figure 2.

4.5. Listeners

The online listening test was distributed among native German listeners. In total, we had 20 participants, although only 14 of them actually completed the test – including eight male and six female listeners between 21 and 76 years (mean age of 34) from various regions of Germany.

The test environment was widely quiet (home or office) and involved different audio equipments. The average testing time resulted to ca. 30 min.

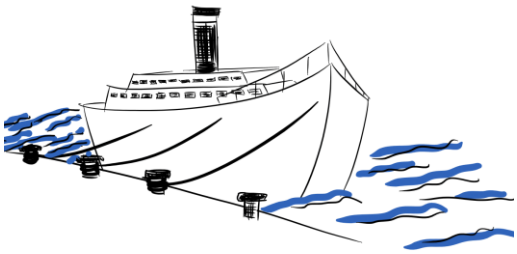


Figure 1: *Das Schiff verließ den Hafen NICHT.*

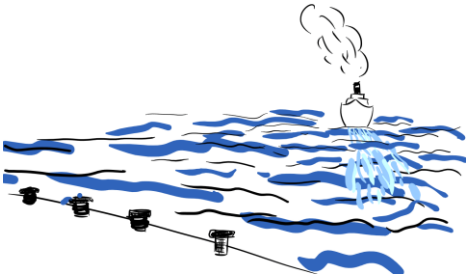


Figure 2: *Das Schiff VERLIESS den Hafen.*

5. Results

In the POLQA test using SWB mode, the Opus-coded samples achieved a score of 4.58 whereas for G.711 as a NB codec, this mode was not suitable.

The results of the listening test are shown in the Figures 3 and 4, first for the overall assessment and the second for the success rate of matching speech samples to given variants.

5.1. Overall speech quality

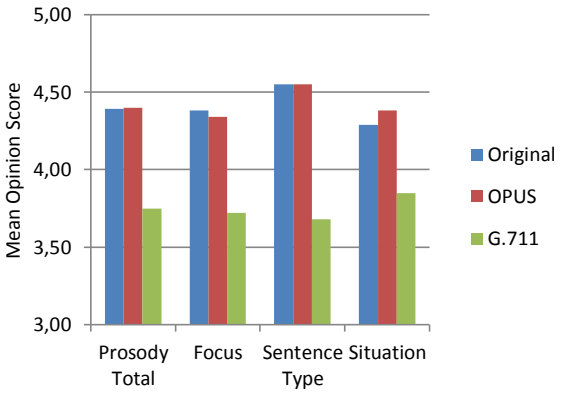


Figure 3: *Speech quality per category and codec*

Figure 3 shows the results of the MOS test when participants were asked to assess the speech quality of what they heard on the scale from 1 to 5. As expected, taking into consideration the different bandwidths of the speech signals, the results for the NB signals of G.711 are noticeably lower than those for the Opus coded FB samples and the original reference.

In the overall speech quality on the MOS scale, G.711 achieved a score of 3.75 whereas the Opus coded and original samples rated at 4.40 respectively 4.39. Interestingly,

averaged over all prosodic phrases, the Opus coding results seem marginally better than the original samples without any degradation.

Furthermore, the quality scores by prosodic category are shown. The categories, as described in the experimental setup are *focus*, *sentence type* and *situation*. In general, the speech samples in *sentence type* received the highest scores of 4.55 for both original and Opus-coded speech, whereas this category scored the lowest at 3.68 when coded with the G.711 codec.

For Opus codec, the other categories show no significant differences with 4.34 in *focus* and 4.38 in *situation*. G.711 performed different, where the *focus* samples were rated only slightly better than those from *sentence type* with a MOS of 3.72. The category *situation* on the other hand, received a rating of 3.85 which is significantly higher.

The original samples were rated higher than or equal to the Opus codec in the two categories *focus*, receiving 4.38, and *sentence type* with 4.55. The lowest rating for original samples is found in the *situation* category with an MOS of 4.29 only.

5.2. Category matching task

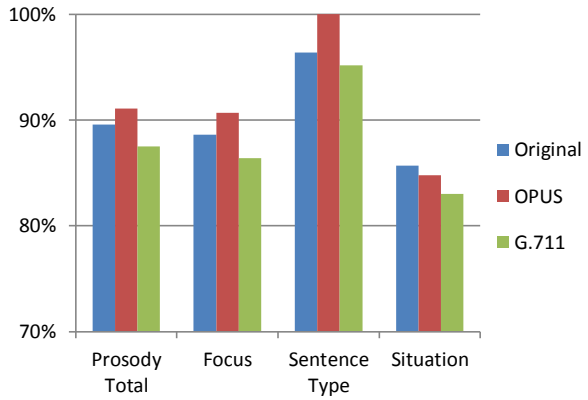


Figure 4: *Success rates within the categories*

In Figure 4, the success rates of the intelligibility test are presented. The *success rate* represents the percentage of samples that were matched to the correct variant depending on the category. These were the emphasised word in *focus*, either question or answer for *sentence type* and one image versus the other in case of *situation*.

In all coding methods, *sentence type* yielded the highest proportion of correct matches while differentiating between two *situations* resulted in the highest number of mistakes. Synchronously with the MOS results, G.711 samples were most frequently mismatched in comparison to the other two, but the relative gap between different coding methods is smaller than in the speech quality assessment. Overall, Opus scored 91.1 %, followed by the original samples at 89.6 % and G.711 succeeding in 87.5 % of the samples.

Overall, the success rates for prosody-only variations were rather high. Again, Opus outperformed the original samples in two of the three categories, namely *focus* and *sentence type* by 2.1 % and 3.6 % respectively and even in the *situation* category there was only a small difference of 0.9 % between both.

5.3. Matching effort

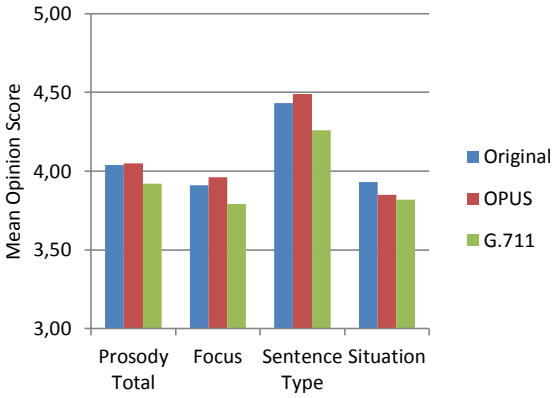


Figure 5: *Matching difficulty per category and codec*

Figure 5 displays the average difficulty of the listeners to match their samples to one of the given answers, using scale from 1 to 5, wherein 1 represents high effort/difficulty while 5 means that the listeners found the matching task easy. For a better comparison (and as in Figure 3 and 4), the results are listed by category and coding method.

The subjective *difficulty* of matching the speech sample to one variant showed a higher variability, although one constant was the category *sentence type* being again assessed with the highest scores of 4.49 for Opus, 4.43 for the original samples and 4.26 in case of G.711 coding (this time meaning that people typically found it easy to decide whether they had just heard a question or a statement). Additionally, this diagram shows once again that Opus and the originals received almost identical assessments of overall 4.05 and 4.04 respectively. G.711 scored worst at 3.92 but in this case with a smaller difference than in MOS or success rate evaluation. Except for Opus, the difficulty of matching for *focus* and *situation* was evaluated almost equal within each coding method: For Opus the *focus* category achieved a difficulty of 3.91 whereas in *situation* it was 3.93 and G.711 was assessed with 3.79 in *focus* and 3.82 in *situation*. Opus on the other hand, scored 3.96 in the category *focus* and 3.85 in *situation*.

6. Discussion and Conclusions

Comparing the MOS results for speech quality to the success rate, it is evident that the success rate scores are relatively higher than the scores in speech quality, which is especially significant in case of the G.711 codec. Therefore it would be one conclusion that prosodic variations are generally well recognised even when the overall speech quality is degraded significantly.

Out of the three prosodic categories, *type of sentence* proved to be the most easily one and also gained the highest scores independent of coding algorithm and setting. In general, there were only small assessment differences between the original samples and its coded equivalents, and in most cases the assessment of Opus-coded speech was even slightly higher compared to the assessment of original samples.

Overall, the assessments were comparably high – approx. 4.3 and above on the five-point MOS scale.

As a preliminary conclusion, the Opus algorithm seems to represent the prosodic features of speech as well as (or even slightly better?) than original speech data do, but further tests need to prove the significance of this finding.

Beyond, prosodic differences in questions and declarative sentences are recognized easier by listeners compared to other prosodic attributes of speech, regardless of external interferences. Our further research will focus on further prosodic and paralinguistic features, alternative coding algorithms and a larger set of validation data.

7. Acknowledgements

We would like to thank Christoph Draxler from LMU Munich for the opportunity to perform our experiment on the Percy Web platform and to our volunteer listeners. Further thank goes to the SwissQual AG, Zuchwil (a Rhode & Schwarz company in Switzerland) for supplying the POLQA testing software – in particular to Jens Berger.

8. References

[1] JOKISCH, O.; MARUSCHKE, M.; MESZAROS, M; IAROSHENKO, V.: Audio and speech quality survey of the opus codec in web real-time communication, Elektronische Sprachsignalverarbeitung. Tagungsband der 27. Konferenz (O. Jokisch, ed.), vol. 81 of Studientexte zur Sprachkommunikation, Leipzig, Germany, pp. 254–262, TUDpress, 2016.

[2] ITU-T: Methods for objective and subjective assessment of speech quality (POLQA): Perceptual Objective Listening Quality Assessment, REC P.863, International Telecommunication Union (Telecommunication Standardization Sector), Sept. 2014.

[3] DOBBRINER, J.: Beeinflussung prosodischer Sprachmerkmale durch Sprach- und Audiocodecs. Bachelorarbeit, Hochschule für Telekommunikation Leipzig, Mai 2016.

[4] VALIN, J; VOS, K.; TERRIBERRY, T.: Definition of the opus audio codec, RFC 6716 (Proposed Standard), Internet Engineering Task Force, Sep. 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6716.txt.

[5] ITU-T: Pulse code modulation (PCM) of voice frequencies, REC G.711, International Telecommunication Union (Telecommunication Standardization Sector), November 1988. Available: https://www.itu.int/rec/T-REC-G.711

[6] DRAXLER, C.: Percy – An HTML5 Framework for Media Rich Web Experiments on Mobile Devices. Proc. 12th Interspeech Conference, pp. 3339 – 3340, Florence, August 2011.

Wahrnehmungsexperimente mit Hilfe eines Computerspiels

Daniel Duran, Natalie Lewandowski, Antje Schweitzer

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

daniel.duran@ims.uni-stuttgart.de

Abstract

Wir präsentieren ein neu entwickeltes Computerspiel zur Durchführung von Sprachwahrnehmungsexperimenten. Klassische computergestützte Experimente in der Sprachwahrnehmungsforschung stellen keine natürliche Versuchsumgebung dar. Sie sind hochgradig überwacht, lenken die Aufmerksamkeit der Versuchsteilnehmer explizit auf bestimmte zu untersuchende Aspekte und sie unterscheiden sich deutlich von natürlichen Gesprächssituationen. Eine Lösung für dieses Problem ist der Einsatz eines Computerspiels, bei dem Aufmerksamkeit auf phonetische Details sich aus der Spielsituation ergibt. Computerspiele finden immer weitere Verbreitung in der Psychologie oder sprachwissenschaftlichen Verhaltensexperimenten. Unsere neue Experimentumgebung implementiert ein klassisches Kategorisierungsexperiment in Form eines Computerspiels. Mit einer modernen Spiel-Engine wurde ein Egoshooter entwickelt, in welchem die Spieler so schnell wie möglich bestimmte Objekte auf dem Bildschirm anklicken müssen. Die Spieler müssen in einer dreidimensionalen Umgebung auf Stimuli reagieren, die durch Spielfiguren dargestellt werden. Die zwei Antwortkategorien werden zunächst durch akustische sowie visuelle Merkmale dargestellt. Die Unterscheidung ist dann zunehmend nur noch anhand der akustischen Stimuli möglich. Die Spieler werden durch diese Spielumgebung motiviert das zu Grunde liegende Kategorisierungsexperiment möglichst gut zu lösen ohne dabei in einer völlig unnatürlichen Situation explizit darauf hingewiesen zu werden. Wir diskutieren praktische sowie theoretische Aspekte unseres Spiels und präsentieren erste Erfahrungen damit aus einer Perzeptionsstudie mit manipulierten phonetischen Details in natürlichen Sprachstimuli.

Schlüsselbegriffe: Sprachwahrnehmung, Experiment, Computerspiele

1. Einleitung

Computerspielbasierte Ansätze kommen unter psychologischen Experimenten zunehmend zum Einsatz [1, 2, 3]. So kommt etwa Washburn [1] zu dem Schluss, dass Computerspiele, zumindest unter bestimmten Bedingungen, die Realitätsnähe psychologischer Forschung erhöhen können. Foreman [2] stellt fest, dass psychologische Forschung von der Entwicklung virtueller Umgebungen profitiert hat und Kimball et al. [3] zeigen den Nutzen von Computerspielen in der Lernforschung.

1.1. Hintergrund

In ihrer Dissertation zur phonetischen Konvergenz (dem Phänomen, bei dem sich zwei Sprecher während eines Dialogs in ihrer Aussprache aneinander annähern) fand Lewandowski [4], dass phonetisch talentierte Teilnehmer in einer fremdsprachlichen Gesprächssituation mehr konvergieren als weniger talentierte. Diese Feststellung wurde mit der Hypothese erklärt, dass *Aufmerksamkeit* für phonetische Details eine Voraussetzung für deren erfolgreiche Speicherung im Gedächtnis sowie ihre spätere Wiederverwendung in der Sprachproduktion darstellt. Die Fähigkeit feinen phonetischen Details seine Aufmerksamkeit zu schenken wiederum wird als Substrat phonetischen Talents angesehen, welches sich einem bewussten Zugang entzieht und welches ein Grundelement des Konvergenzmechanismus darstellt (neben anderen individuellen Persönlichkeitsmerkmalen). Diese Hypothese wird gestützt von einer post-hoc Analyse der Konvergenzresultate in [4] anhand von Daten eines Tests der mentalen Flexibilität, welche eine positive Korrelation zwischen den beiden Dimensionen zeigt. In diesem Test müssen Probanden ihre Aufmerksamkeit schnell an ein sich veränderndes Szenario anpassen. Je schneller die Teilnehmer in diesem Test waren, umso mehr konvergierten sie in den Dialogen [5]. Segalowitz [6] schlägt vor, dass zwei Prozesse zur Redeflüssigkeit sowie zur Sprachperzeption beitragen: *access fluidity* (AF) und *attention control* (AC). AC wird als die Fähigkeit definiert, die Aufmerksamkeit auf unterschiedliche semantische Ebenen zu fokussieren (d.h. lokale vs. globale Bedeutungszusammenhänge). Während Segalowitz sich auf die Verschiebung zwischen lokalen und globalen Bedeutungszugriffen konzentriert, schlagen wir vor, dass AC auch beim Umschalten zwischen verschiedenen Dimensionen des Sprachsignals beteiligt sein kann, z.B. zwischen detaillierter akustischer Form und der Bedeutung. AC wird üblicherweise in einem *alternating runs paradigm* [7] getestet, wobei Probanden eine Reihe von Entscheidungen in zwei alternierenden unterschiedlichen Aufgaben fällen müssen. Demgegenüber beschreibt AF die Schnelligkeit und Automatisierung bei der Verknüpfung von Worten mit deren Bedeutung. AF wird üblicherweise mittels Reaktionszeiten in lexikalischen oder semantischen Entscheidungs- oder Verständnistests gemessen [6]. In einer aktuell laufenden Studie zur phonetischen Konvergenz in Dialogen [8] wenden wir verschiedene psychologische Tests an, um die individuelle Aufmerksamkeit der Teilnehmer für phonetische Details zu erfassen. Die typischen Aufgaben in derartigen Tests sind hochgradig überwacht, und eher unnatürlich, da sie die Aufmerksamkeit der Versuchsteilnehmer explizit auf bestimmte zu untersuchende Aspekte lenken. Eine Lösung für dieses Problem stellt der Einsatz eines Computerspiels dar, in welchem AC und AF Bestandteile des Spiels selbst darstellen. Unsere Annahme ist, dass ein Computerspiel natürlichere Daten liefert, da Aufmerksamkeit eine Notwendigkeit der Spieleumgebung ist und bestimmte Aktionen als Reaktion auf Ereignisse im Spiel verlangt werden anstelle expliziter bzw. bewusster Entscheidungen.

1.2. Forschungsstand

Wade & Holt [9] untersuchen beiläufiges perzeptuelles Lernen von nicht-sprachlichen Geräuschen mit komplexer spektraler Struktur. Sie heben hervor, dass sich der übliche Trainingsan-

satz der “Kategorisierung-mit-Feedback” in Studien zur nichtsprachlichen auditiven Kategorisierung nachweisbar von jenen Prozessen unterscheidet, durch welche Menschen natürlichen Sprachlauten ausgesetzt sind, und dass er möglicherweise so fundamental anders ist, dass ein informativer Vergleich ausgeschlossen ist. Sie schlagen daher eine Methode vor, welche “einige essentielle Aspekte des phonetischen Erwerbs erfasst”: ein Computerspiel. Lim & Holt [10] verwendeten ein derartiges Computerspiel um erwachsene japanische Probanden in ihrer Studie zu trainieren zwischen den englischen /r/ und /l/-Kategorien zu unterscheiden. Ihre Ergebnisse zeigen, dass die Teilnehmer die nicht-muttersprachlichen phonetischen Kategorien ohne explizites Training lernten. Das Ergebnis nach 2,5 h computerspielbasierten Trainings war vergleichbar mit 2–4 Wochen Training mittels herkömmlicher Unterrichtsmethoden mit explizitem Feedback.

Im Folgenden beschreiben wir unsere neue computerspielbasierte Umgebung für phonetische Sprachwahrnehmungsexperimente. Zusätzlich stellen wir eine Anwendung unseres Computerspiels in einer experimentellen Studie zur Aufmerksamkeit bei natürlicher Sprachwahrnehmung vor.

2. Die Ψ X 732 Computerspielumgebung

Das Computerspiel Ψ X 732 wurde mit der Unity Spiel-Engine implementiert [11]. Unity bietet eine Spiel-Engine für qualitativ hochwertige 3-D Spiele auf dem aktuellen Stand der Technik, welche für Versuchsteilnehmer, die Erfahrung mit modernen Computerspielen haben, ansprechend sind. Das erste Konzept und eine Machbarkeitsstudie wurden von Lange & Pfeiffer durchgeführt und mit einem Perzeptionsexperiment getestet [12]. Ψ X 732 bezieht Ideen und Konzepte dieser früheren Arbeit ein, ist aber eine vollständige Neuimplementierung. Die gesamte Spielelogik (wie z.B. die Verarbeitung der Benutzereingaben, das Verhalten der virtuellen Agenten, die Experimentsteuerung oder das Logging) ist in C# implementiert. Versuchsparameter sind nicht hart codiert sondern können durch eine einfache Textdatei konfiguriert werden, welche zur Laufzeit vom Spiel geladen wird. Die Liste der konfigurierbaren Parameter enthält unter anderem Zeitlimits, Versuchsspezifikationen sowie die im Spiel angezeigten Texte. Die Audiodateien (im wav-Format) sind ebenfalls nicht in den Quellcode des Spiels integriert, sondern werden auch erst zur Laufzeit von der Festplatte in das Spiel geladen. Durch dieses Konzept ist Ψ X 732 sehr flexibel und bietet eine sprachunabhängige Versuchsumgebung für verschiedene Perzeptionsexperimente.

2.1. Spielaufbau

Das Spiel gehört zum Genre der Egoshooter. Die Spieler treffen in einer virtuellen natürlich aussehenden Umgebung auf *Agenten* (im Sinne autonomer Akteure künstlicher Intelligenz) und müssen auf diese reagieren. Die Agenten gehören jeweils zu zwei Kategorien (im Spiel als “außerirdische Invasoren” bzw. “menschliche Zivilisten” bezeichnet). Nähern sich die Spieler einem Agenten wird dieser aktiviert und beginnt, die Spieler zu jagen. Der akustische Stimulus wird abgespielt und gleichzeitig erscheint eine visuelle Anzeige neben dem Agenten mit einer farblichen Kennzeichnung sowie einem beschreibenden Text. Die Farben entsprechen den Strahlen, die von den “Waffen” bzw. Werkzeugen erzeugt werden, die der jeweiligen Agentenkategorie zugeordnet sind (im Spiel wird das Wort “Waffe” vermieden). Ein Fadenkreuz im Zentrum des Bildschirms ermöglicht präzises Zielen. Die Spieler sind mit zwei Werkzeugen

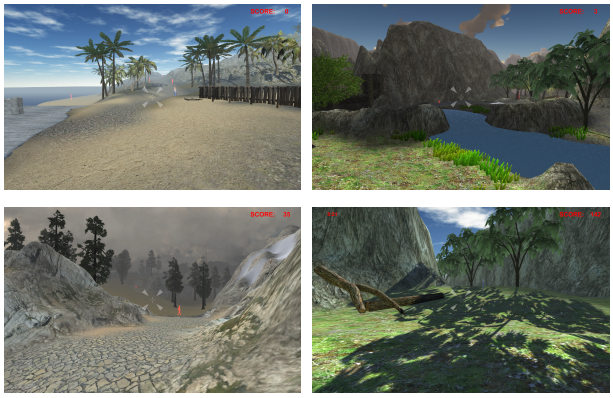


Abbildung 1: Screenshots der vier Experimentlevels.

ausgerüstet: eines, das die getroffenen Agenten in einem blauen Eisblock einfriert (der *Freezer*) und eines, das die getroffenen Agenten mit einem Bündel grüner Lichtstrahlen auf ein Raumschiff in Sicherheit beamt (der *Beamer*). Diese Werkzeuge sind jeweils mit der rechten oder linken Maustaste verknüpft. Die Strahlen von den Werkzeugen erscheinen für eine kurze Zeit als leuchtende blaue oder grüne Linien vom rechten oder linken Bildschirmrand (entsprechend der Maustaste) hin zum getroffenen Objekt. Eine Lichtkugel erscheint kurz an der getroffenen Stelle, sofern kein Agent getroffenen wurde. Dieses Feedback hilft den Spielern bei der Orientierung und beim Zielen. Die Zuordnungen von Maustaste, Ausrüstung, Farbe und Agentenkategorie bleibt während des gesamten Spiels konstant (*blau* = *Freezer* = *Alien*, *grün* = *Beamer* = *Mensch*).

Ψ X 732 besteht aus einem Einführungslevel und vier Experimentleveln. Ein Willkommensbildschirm zeigt zu Beginn des Spiels eine rotierende Raumstation von außen mit der Erde im Hintergrund. Hier werden den Spielern die Steuerung des Spiels und die Grundlagen der Geschichte als Text angezeigt. Im Einführungslevel innerhalb der Raumstation erfahren die Spieler mehr über die Hintergrundgeschichte und werden in die Szenerie eingeführt. Ein wesentlicher Bestandteil des Einführungslevels ist ein *Trainingsprogramm*, bei dem die Spieler zunächst durch einen verwinkelten Korridor navigieren müssen. Danach wird das Zielen und Treffen mit der Maus an Testdummies auf einer Art Schießstand geübt. Es hat sich gezeigt, dass Unterschiede zwischen erfahrenen Spielern und Personen ohne Erfahrung mit 3-D Computerspielen schon nach kurzer Trainingszeit in der virtuellen Umgebung ausgeglichen werden können [13]. Das Einführungslevel mit dem Training dient darüber hinaus dazu, die individuelle *Baseline* für jeden Spieler in einer relativ entspannten Umgebung zu ermitteln. Die verbleibenden vier Levels des Spiels stellen die eigentlichen Experimente dar. Sie spielen alle in einer offenen Landschaft auf der Erde (siehe Abbildung 1). In diesen Levels treffen die Spieler entsprechend der Spielkonfiguration auf Agenten. Im offenen Gelände können die Ziele teilweise schon aus großer Distanz erkannt werden. Zunächst erscheinen die Agenten als rot leuchtende, halbdurchsichtige Gestalten. Nähern sich die Spieler ihnen, werden sie durch eine weibliche, menschliche Figur ersetzt. Dies stellt einen bewussten Bruch mit üblichen Spielekonventionen dar: Agenten sind in unserem Spiel leicht zu finden, so dass die Spieler sich schon auf sie vorbereiten können.

Ein Versuchs-*Trial* wird durch folgende Ereignisse im Spiel definiert: (1) Er beginnt mit der Wiedergabe des akustischen Sti-

Tabelle 1: Teilnehmergruppen mit Experimentreihenfolgen

Gruppe	Exp.1	Exp.2	Exp.3	Exp.4	w	m
1	f0	FRIC	VOT	F2	4	3
2	F2	VOT	FRIC	f0	3	3
3	FRIC	F2	f0	VOT	3	4
4	VOT	f0	F2	FRIC	3	4

mulus. In diesem Moment wird der dazugehörige Agent aktiv. (2) Er endet, wenn die Spieler den Agenten treffen oder ohne zu reagieren am Agenten vorbeigehen. Ein weiterer Bruch mit gängigen Spielekonventionen ist dabei der Umstand, dass immer nur jeweils ein Agent gleichzeitig aktiv wird und den Spieler jagt. Beginnt ein Trial, bleiben alle anderen Agenten inaktiv, auch wenn die Spieler ihnen nahe kommen. Außerdem werden Agenten nur aktiv, wenn sie sich in einem engen Fenster direkt vor dem Spieler befinden. Dadurch können sich die Spieler auf die Agenten vorbereiten und bereits auf sie zielen. Dies ermöglicht schnellere Reaktionen und vermeidet Ambiguitäten bei der Kategorisierung. Es reduziert darüber hinaus auch Drehungen um die eigene Achse, was unerfahrenen Spielern unangenehm sein kann (siehe Diskussion).

3. Experiment

Wir präsentieren hier die Anwendung von Ψ X 732 im Rahmen einer laufenden Studie über die Perzeption akustischer Details in der Sprachwahrnehmung [8].

3.1. Daten und Teilnehmer

Als Stimuli wurden kurze Äußerungen einer deutschen Sprecherin in einem schallisolierten Raum aufgezeichnet. Manipulierte Versionen wurden mit Praat [14] erstellt. Vier phonetische Merkmale wurden verändert: der Umfang der Stimmgrundfrequenz (f0), die Höhe des zweiten Formanten in Vokalsegmenten (F2), die länge der Stimmansatzzeit in Plosiven (VOT), das Frikativspektrum durch Entfernen niedriger Frequenzbereiche (FRIC). In einem Experimentlevel wird jeweils nur eine Art der Manipulation im Vergleich zu unveränderten Stimuli getestet. Insgesamt 27 erwachsene deutsche Muttersprachler (13 weiblich) absolvierten die Experimente mit dem Spiel. Alle Teilnehmer gaben keine bekannten Hörschäden an. Um Reihenfolge-Effekten vorzubeugen, wurden die Teilnehmer in vier Gruppen eingeteilt (Tabelle 1).

3.2. Ablauf

Die Teilnehmer wurden in einem ruhigen, fensterlosen Raum an einen Computer gesetzt (MS Windows Betriebssystem, Bildschirmauflösung 1680 × 1050 Pixel, mit gewöhnlicher Tastatur und Maus als Eingabegeräte). Die Teilnehmer trugen hochwertige Sennheiser Kopfhörer. Zu Beginn der Sitzung wurden sie aufgefordert, die Lautstärke auf ein angenehmes Niveau einzustellen. Es wurde jedoch nicht darauf hingewiesen, dass sie auf die Sprachstimuli achten müssten, um die Agenten im Spiel zu unterscheiden. Die Trials sind pro Level in drei Blöcke aufgeteilt. *1:* Zusätzlich zum akustischen Stimulus wird eine visuelle Hilfe angezeigt, die den aktiven Agenten als Ziel markiert und dessen Kategorie durch Text und Farben kennzeichnet. *2:* Die visuelle Kennzeichnung verblasst, so dass die Farbe graduell mit jedem neuen Trial verschwindet. *3:* Die Kategoriezugehörigkeit der Agenten wird visuell nicht mehr gekennzeichnet (lediglich die Markierung als aktuelles Ziel bleibt erhalten). Die Zuordnung der Agenten ist für die Spieler nur noch über den akustischen Stimulus möglich. Die ersten beiden Blöcke stellen Trainings-Trials dar wobei die Spieler die Kategorisierung anhand der Stimuli erlernen können. Der dritte Block stellt das eigentliche Kategorisierungsexperiment dar. Die Anzahl der Stimuli in den ersten beiden Blöcken ist in der Konfiguration vorgegeben. Im dritten Block werden so viele Stimuli dargeboten, bis entweder das Zeitlimit für das Level erreicht oder die Liste der Stimuli vom Spieler abgearbeitet wurde.

3.3. Ergebnisse

Für die hier besprochenen Ergebnisse wird jeweils nur der erste Mausklick der Spieler als Reaktion auf einen Experimentstimulus gewertet. Die jeweilige Zielrichtung zu diesem Zeitpunkt ist zwar in den Protokollen enthalten, wird hier jedoch nicht berücksichtigt.

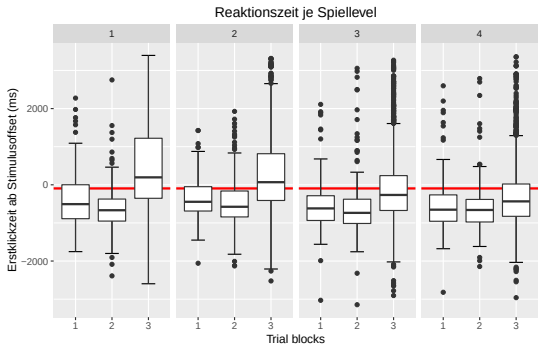


Abbildung 2: Zeit bis zum ersten Klick nach Ende des Stimulus je Spiellevel für alle Spieler.

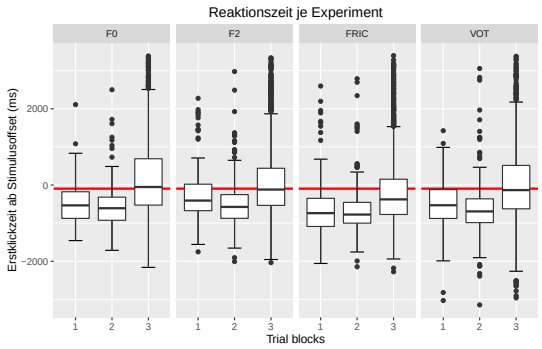


Abbildung 3: Zeit bis zum ersten Klick nach Ende des Stimulus je Experiment für alle Spieler.

Abbildung 2 zeigt die Reaktionszeiten für alle Spieler relativ zum Ende des akustischen Stimulus nach Spielleveln sortiert. Negative Werte entsprechen dabei Mausklicks, die erfolgten während der akustische Stimulus noch abgespielt wurde. Die überwiegend negativen Werte in den ersten beiden Blöcken zeigen, dass die Spieler sich hier im wesentlichen noch auf die visuelle Information verließen und nicht das Ende des akustischen Stimulus abwarteten. Im dritten Block ist ein Sinken der Reaktionszeiten über die vier Spiellevel hinweg zu erkennen. Dies lässt auf einen Gewöhnungseffekt im Umgang mit dem Spiel schließen. Abbildung 3 zeigt die Reaktionszeiten für alle

Spieler relativ zum Ende des akustischen Stimulus nach Experimenten sortiert. Das Experiment mit dem manipulierten Frikativspektrum (FRIC) zeigt die schnellsten Reaktionen, insgesamt sind die Werte aber alle sehr ähnlich.

Die Genauigkeit des ersten Mausklicks für alle Spieler gemessen an der Auswahl der rechten oder linken Maustaste (Kategoriezugehörigkeit), sortiert nach Spielleveln ist in Abbildung 4 zu sehen, und in Abbildung 5 sortiert nach Experimenten. Abbildungen 4 und 5 zeigen, dass die Spieler im dritten Block im Mittel mit ihrer Kategorisierungsgenauigkeit auf Zufallsniveau liegen (was in etwa bei 0,5 liegt). Die Aufgabe für die Spieler war also nicht trivial. Die breite Streuung zeigt aber, dass es hier deutliche individuelle Unterschiede gibt. In Abbildungen 2 bis 5 entspricht die rote horizontale Linie jeweils dem Mittelwert über alle Spalten.

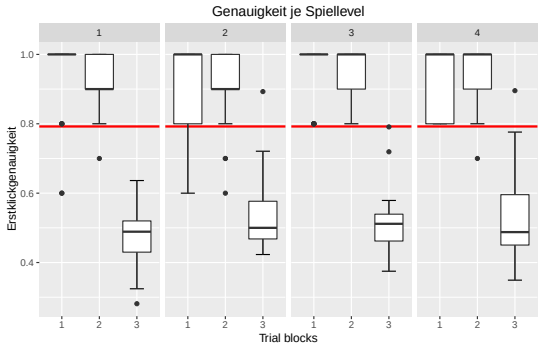


Abbildung 4: Kategorisierungsgenauigkeit des ersten Klicks für alle Spieler je Spiellevel.

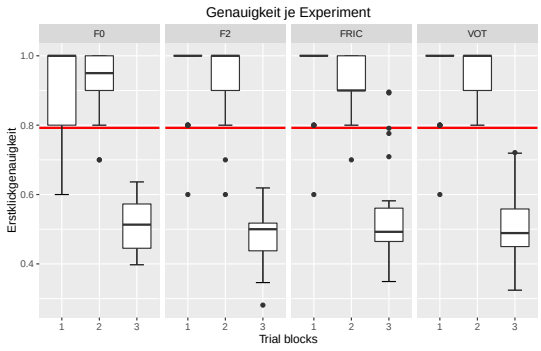


Abbildung 5: Kategorisierungsgenauigkeit des ersten Klicks für alle Spieler je Experiment.

3.4. Diskussion und Ausblick

Bei der Interpretation der oben gezeigten Ergebnisse ist zu beachten, dass sie von den eingesetzten Stimuli abhängen. In wie weit sie sich mit Ergebnissen klassischer Perzeptionsexperimente vergleichen lassen und in wie weit sie möglicherweise Artefakte der virtuellen Spielumgebung enthalten, bleibt zu untersuchen. Die breite Streuung der Ergebnisse zeigt, dass die hier präsentierte Anwendung von Ψ X 732 der Aufgabe angemessen war: individuelle Unterschiede in der akustisch-phonetischen Sprachwahrnehmung können anhand eines Kategorisierungsexperiments ermittelt werden, welches in Form eines Computerspiels dargeboten wird.

Die vorläufigen Ergebnisse zeigen, dass Ψ X 732 eine geeignete Experimentierumgebung für Perzeptionstests ist. Ein potentielles Problem von Computerspielen in phonetischen Experimenten ist das Phänomen der *Cybersickness*. Frey et al. [13] berichten über Symptome wie Unwohlsein oder Kopfschmerzen bei neun von 85 Probanden. In unserer Studie musste nur ein Teilnehmer im dritten Level das Spiel abbrechen. Frey et al. stellen fest, dass das Risiko für Cybersickness durch geeignetes Leveldesign reduziert werden kann und dass virtuelle 3-D Umgebungen generell für Experimente eingesetzt werden können, auch wenn Probanden über keine Erfahrung damit verfügen.

4. Dank

Diese Arbeit wurde im Rahmen des SFB 732 (A4) von der Deutschen Forschungsgemeinschaft (DFG) finanziert.

5. Bibliographie

[1] D. A. Washburn, "The games psychologists play (and the data they provide)," *Behavior Research Methods, Instruments, & Computers*, vol. 35, no. 2, pp. 185–193, May 2003.

[2] N. Foreman, "Virtual Reality in Psychology," *Themes in Science and Technology Education*, vol. 2, no. 1-2, pp. 225–252, 2009.

[3] G. Kimball, R. Cano, J. Feng, L. Feng, E. Hampson, E. Li, M. G. Christel, L. L. Holt, S.-j. Lim, R. Liu, and M. Lehet, "Supporting research into sound and speech learning through a configurable computer game," in *IEEE International Games Innovation Conference (IGIC)*, 2013, pp. 110–113.

[4] N. Lewandowski, "Talent in nonnative phonetic convergence," Doctoral dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2012. [Online]. Available: http://dx.doi.org/10.18419/opus-2858

[5] ——, "Phonetic convergence and individual differences in non-native dialogs," in *New Sounds*, Montréal, Canada, 2013.

[6] N. Segalowitz, "Access fluidity, attention control, and the acquisition of fluency in a second language," *TESOL Quarterly*, vol. 41, no. 1, pp. 181–186, 2007.

[7] R. D. Rogers and S. Monsell, "Costs of a predictable switch between simple cognitive tasks." *Journal of Experimental Psychology: General*, vol. 124, no. 2, pp. 207–231, 1995.

[8] A. Schweitzer, N. Lewandowski, and D. Duran, "Attention, please! Expanding the GECO database," in *Proceedings of ICPhS 18*, Glasgow, UK, 2015, paper number 620. [Online]. Available: http://www.icphs2015.info/pdfs/Papers/ICPHS0620.pdf

[9] T. Wade and L. L. Holt, "Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2618–2633, 2005.

[10] S.-j. Lim and L. L. Holt, "Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization," *Cognitive Science*, vol. 35, no. 7, pp. 1390–1405, 2011.

[11] Unity Technologies, "Unity," Computer program, 2016, version 5. [Online]. Available: http://unity3d.com/

[12] L. Lange, B. Pfeiffer, and D. Duran, "ABIMS – auditory bewildered interaction measurement system," in *Proceedings of Interspeech*, Dresden, 2015, pp. 1074–1075. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_1074.html

[13] A. Frey, J. Hartig, A. Ketzel, A. Zinkernagel, and H. Moosbrugger, "The use of virtual environments based on a modification of the computer game Quake III Arena® in psychological experimenting," *Computers in Human Behavior*, vol. 23, no. 4, pp. 2026–2039, 2007.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2016, version 6. [Online]. Available: http://www.praat.org/

Recording a High-Quality German Speech Database for the Study of Speaker Personality and Likability

Laura Fernández Gallardo

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
laura.fernandezgallardo@tu-berlin.de

Abstract

The ongoing process for recording a personality and likability database in German is motivated and described. Overall, high-quality and consistency among recordings is pursued, in order to avoid possible biases when rating speaker characteristics and low performance when automatically detecting them. Prescribed and spontaneous human-human dialogs are recorded using three different microphones in an acoustically-isolated room with 48 kHz sampling frequency. Natural and neutral speech is recorded, controlling the absence of background noises. So far, 101 German speakers without accent have participated and two listening tests have been conducted with part of these data. Our goal is to extend the number of recorded participants to at least 200, conduct more listening tests, and share the speech files, metadata, and associated labels, features, and analyses with the scientific community.

Index Terms: database recording, voice likability, speaker personality

1. Motivation for a New Database

Speech-based human-machine communications are no longer constrained to the detection of the users' message (automatic speech recognition) and to speech synthesis but also the automatic characterization of the users has been gaining attention over the last decades. Of particular interest is the detection of speakers' social characteristics such as personality and voice likability [1, 2]. Successful automatic personality traits and likability recognition—correlating well with human judgments—can enable the systematic prediction of social human behavior and the improvement of speech synthesis for human-machine interactions.

Generally, client-server architectures are being incorporated in speech-based services, which imply that coded speech signals are transmitted to a server to process the information from the signals. However, the influence of different telephone transmissions on the automatic detection performance of voice personality and likability has been overlooked. Commonly, the data are collected through telephone communications or present a sampling frequency of 8 kHz (see Section 2), filtering out the possibly relevant high-frequency speech cues. The aim of our ongoing research is to determine whether and to which extent transmission channel degradations affect the human and the automatic detection of speaker personality and likability.

To investigate the influence of transmission channels, our approach is to systematically apply channel degradations to speech recorded in clean conditions. Therefore, two main requirements for the speech data of our study are 1) clean speech, that is, recorded through high-quality microphone in quiet conditions, and 2) sampling frequency of at least 16 kHz, needed for wideband (WB, 50–7,000 Hz) transmissions, from which possible benefits are to be contrasted to conventional narrowband (NB, 300–3,400 Hz) transmissions. A sampling rate of at least 32 kHz would be needed for super-wideband (SWB, 50–14,000 Hz) conditions to be studied. Since no speech database has been found to meet these requirements (see review in Section 2), a new high-quality database of German speech is being recorded, and is presented in this paper. The aim is to record the voices of 200–300 speakers (approx. half of them female) over high-quality microphones and with 48 kHz sampling frequency. This database will serve not only for our study of speaker personality and likability but also for any speech-related research requiring high-quality clean recordings in German.

2. Brief Review of Existing Databases

Unfortunately, the publicly available speech databases employed for personality and likability research seem not to be sufficient for our intended investigation. Objections to these datasets for our work are that their segments are already transmitted or have a sampling frequency lower than 16 kHz (SSPNet Speaker Personality Corpus (SPC) [3], Personable and Intelligent virtual Agents (PersIA) [4], Electronically Activated Recorder (EAR) Corpus [5], and Speaker Likability Database (SLD) [2], VoiceClass [6]), that the language is not German (SPC, PersIA, EAR, AMI, Maptraits), or that the dataset contains a very reduced number of speakers (EmoDB [7]). PhonDat [8] presents read speech, which may difficult the manifestation of personality traits. No previous study on speaker personality has been found that employed read speech. Datasets like Smartkom or other emotional databases (such as VAM, SEMAINE, etc.) contain very little neutral content, which also makes them unsuitable for our research.

Other corpora from the speech recognition and speaker verification areas present one or more drawbacks as well. These are: different microphones and recording environments (Voxforge), NB telephone quality (Yoho, NIST SRE data, Switchboard, Mixer, Phonebook, Call-home, VeriDat, SpeechDat), too little material (OLLO), English databases (TIMIT, XM2VTS, CSR, AusTalk, BANCA, etc.), and non-native speakers of the recording language (MOBIO).

Differently, the T-Labs Personality Database (TPDB) was recorded at a sufficient sampling rate (44.1 kHz) and contains German conversational speech. It was shown to be useful for the detection of personality characteristics by humans [9]. The dataset, however, only contains 29 speakers recorded over a table microphone and 35 speakers recorded over a headset and is hence too small for training and testing regression and classification systems. An approach similar to that used for the recording of TPDB has been adopted in our recording procedure to elicit spontaneous speech. Both databases could be used to-

gether for research on speech since the recording conditions are comparable.

It should be noted that the review presented in this section does not intend to be exhaustive.

A closer look is taken at the SPC [3] and SLD [2] databases, proposed in the Interspeech Speaker Trait Challenge [10] organized in 2012 (IS12) for the personality and for the likability sub-challenges, respectively. A total of 640 speech clips of 10 s or shorter were selected for the SPC database from 322 speakers (263 males and 59 females) of French. The spontaneous speech enabled the speakers' personality to be labeled by 11 raters using the Big-Five inventory (BFI)-10 questionnaire [11]. However, the speech data, extracted from transmitted radio signals and sampled at 8 kHz, present channel distortions and low quality. Besides, the radio speaking style is not conversational, in contrast to a desirable scenario involving human interactions.

The SLD database is a subset of the larger aGender database [6], recorded over telephone lines presenting different degradations, and sampled at 8 kHz. It contains 800 speakers divided into 6 groups according to their age and gender, and one sentence per speaker. The sentences, with an average length of 3.05 s, have been used to obtain likability ratings in [2]. Besides the low and variable quality of the recordings, an important drawback of this database for the study of voice likability is that the content of the utterances varies across speakers. In [2], the authors state: *"We're aware of the fact that the meaning of the words might affect the perceived likability and it would have been better to have the same text spoken by all test speakers, (...)"*.

The design of our data collection procedure aimed at embracing the revealed advantages from these databases while avoiding their drawbacks. The maximum unweighted accuracy (UA) reached by the participants of the IS12 challenge using SPC and SLD was only UA=69.0% for personality and UA=65.8% for likability, both in binary classification tasks [12]. This rather modest performance could partly be attributed to some of the mentioned disadvantages.

3. Database Material

3.1. Speech

Different dialogs (human-human) were chosen as recording material, in order to promote a communications scenario, the focus of the future investigations with these data. The contents of the utterances recorded are all related to information requests over the telephone. In order to collect the same segments from all speakers (and thus avoid biases in the later perceptions of likability [2]), it was proposed that the speakers read dialog turns maintaining the given wording (prescribed speech). On the other hand, because personality traits are more easily manifested in non-read speech, the speakers are also asked to hold free conversations given its structure and information that speakers should provide and collect (spontaneous speech).

A female German speaker of 26 years old assists this project conducting all recording sessions. She was born and raised in northern Hesse, Germany, and claims not to have a noticeable German accent deviating from the standard High German dialect.

A recording session is divided into two parts. In the first part, the speakers are asked to read turns from four different dialogs. The dialogs simulate telephone calls held with the recording assistant, who always plays the role of a contact person or agent. The recorded speech, from the client's side, includes

Tabelle 1: *Recording material for our database.*

Prescribed speech	Dialog 1: Health insurance (Table 2)
	Dialog 2: Mobile phone rate plan
	Dialog 3: Car rental—inquiry
	Dialog 4: Real estate agency
Spontaneous speech	Dialog 5: Car rental—booking
	Dialog 6: Pizza
	Dialog 7: Book from the library
	Dialog 8: Doctor's appointment

Tabelle 2: *Dialog 1 (prescribed speech). A: agent turns (recording assistant), B: client turns (speaker).*

A: *Guten Tag, DKS Versicherungen. Mein Name ist Heinmüller, was kann ich für Sie tun?*
B: Guten Tag, mein Name ist Schmidt. Ich hab eine Frage bezüglich meiner Krankenversicherung.
A: *Geben Sie mir bitte Ihre Vertragsnummer durch.*
B: Meine Vertragsnummer ist die 4035.
A: *Einen Moment bitte. Alles klar, wie lautet Ihre Frage?*
B: Ich hab von meinem Hausarzt ein Rezept über 10 Physiotherapie-Anwendungen bekommen. Deshalb wollte ich mal fragen, ob Sie die Kosten in vollem Umfang übernehmen.
A: *Ich verstehe. Ich leite Sie eben zu Ihrem persönlichen Sachbearbeiter weiter.*
B: Alles klar, Dankeschön!

different inquiries about some information: to a health insurance company (Table 2), a mobile telecommunications company, a car rental company, and a real estate agency, for each dialog, respectively. The speakers are asked to read the exact given text as naturally as possible, yet without emotions or exaggerated friendliness. To the extent possible, it was avoided that the utterances sound like read speech.

While the first recording part comprises prescribed texts, spontaneous speech is elicited in the second part. Four spontaneous telephone dialogs are held between the speaker and the recording assistant: renting a car, ordering a pizza, ordering a book from the library, and making an appointment at the doctor's. These dialogs follow the scenarios known as Short Conversation Tests, found in the International Telecommunication Union (ITU)-T Rec. P.805. Again, the participants play the client's role and the recording assistant the agent's role. A summary of the recorded speech contents is given in Table 1.

The speakers are instructed to always talk naturally and neutrally, without hesitations, and with the correct pronunciation (i.e. without accent and not tongue-tied). Special attention is paid to avoiding any background noise, e.g. papers rustling, and any disturbance in the speaker's voice, e.g. croaky or harsh voice, clearing one's throat, snickering, etc. Unsatisfactory dialog turns (specially for prescribed speech) are always repeated until an acceptable turn is uttered.

When 1 s silences are inserted between dialog turns, the prescribed dialogs have an approximate mean duration of 18 s, and the spontaneous dialogs of 45 s.

3.2. Metadata

Socio-demographic data about the speakers are systematically collected when the recording session is completed and stored as metadata. This information comprises: age, gender, place of birth, chronological places of residence and duration of stay,

place of birth of the mother, place of birth of the father, highest education level, educational background, main occupation, past occupations (if any), years of work experience (if any), and self-assessed personality traits (the speakers are asked to fill in the BFI-10 questionnaire [11]).

These data, collected from every speaker, may be useful for future investigations of effects on speech idiosyncrasies. All data are pseudo-anonymized, i.e. the mapping between the speakers' true names and their speech files and metadata is securely stored. This is only done for the solely purpose of being able to delete a speaker's participation if he/she wishes at any time.

Before the recording session starts, the speakers are asked to carefully read information about the project and to sign a consent form. They are informed that:

- their voices will be recorded and that socio-demographic data will be collected in the session
- the data are stored and pseudo-anonymized for scientific analyses which can be published
- the data will be treated confidentially and be part of a database that may be publicly released under license agreement for exclusive use in scientific studies
- the participation is voluntary and it is possible to withdraw from the study at any time without penalty

Any question they may have is answered and the participants only proceed to the recording session if they agree with the consent form. They keep a copy of the signed form and of the project's information sheet with our contact details.

4. Speakers' Recruitment

Most of the speakers who participate in the recording process are recruited via the 'Probandenportal' of the Technische Universität (TU) Berlin, mailing lists and paper announcements. The requisites for the participation are:

- Native German speaker without dialect or accent
- 20–35 years old
- No hearing or speech disorders
- No sickness or cold at the time of the recording

Special effort has been made to only recruit speakers of the standard High German dialect, since pronunciation has been shown to have an important effect on the impressions of voice likability. It has been reported that listeners seem to prefer to hear voices with their own accent [13] or with no accent [14]. Some persons who register to participate and attend the agreed time slot but from whom a marked accent can be perceived by the conductor of the recordings are, regrettably, not accepted for the participation.

The age range was restricted to young adults in order to avoid inconsistencies in likability or personality ratings caused by age differences. Using the SLD data [2], it was found that senior speakers (55–80 years old) were rated as less likable than adult speakers (25–54 years old).

The duration of the recording session is approximately 30 minutes and the speakers are compensated with 6 €.

5. Recording Setup

The recording sessions are conducted in the acoustically-isolated room *Nautilus*, at the TEL-building of the TU Berlin. The room's dimensions are 2.75 m x 2.53 m x 2.10 m, and RT60 = 0.08 s at 2 kHz. The speaker sits alone in this room, where three microphones are placed. These microphones are:

Abbildung 1: Room *Nautilus* with the recording setup. In the position of a speaker is Laura Fernández Gallardo.

- Stand-up microphone: AKG C 414B-XLS (frequency range 20–20,000 Hz)
- Table microphone: Audio-Technica U851R (frequency range 30–20,000 Hz)
- Headset microphone: Sennheiser HMD 46 (frequency range 100–12,000 Hz)

Figure 1 shows the recording setup. The approximate distances from each microphone to the speaker's mouth are 35 cm, 50 cm, and 3 cm for the stand-up, table, and headset microphone, respectively.

The three signals corresponding to each microphone are amplified and then digitalized employing the RME Fireface UCX Audio Interface, conducted to the outside of the room via a USB port, and recorded using the software Cubase 4 with 48 kHz sampling frequency and 32-bit quantization. The recording assistant sits outside of the *Nautilus* room, from which she can listen to the speaker, give the pertinent instructions, and hold the dialogs. She employs headsets also conducted to the interior of the room, to the speaker's headset.

6. Listening Tests conducted with the Recorded Data

So far, two listening tests have been conducted using part of the recorded speakers and the signals acquired through the stand-up microphone.

6.1. Likability and Personality-based Social Relations

The first listening test was part of our study presented in [15]. A group of 30 speakers of our database were asked to participate in a listening test where the voices from the group were played, following a round-robin design. Hence, the participants mutually rated each another in terms of voice likability and personality. These ratings were collected by means of continuous scales with labels at their ends. Speech signals transmitted through a narrowband channel (codec G.711 at 64 kbit/s), through a wideband channel (codec G.722 at 64 kbit/s), and clean speech (sampled at 44.1 kHz) were used as stimuli. The Social Relations Model (SRM) [16] was then employed as a statistical approach to analyze the mutual perceptions.

In short, it was found and reported in [15] that persons which are perceived as extroverted and agreeable are also rated with a higher likability. In addition, people similar in agreeableness and neuroticism tend to rate each other's voice likability more positively. WB voices, with respect to NB, were si-

gnificantly higher rated in terms of likability on average, led to lower variance among perceivers' rating tendencies, and allowed listeners to better distinguish between non-likable and likable speech. It could also be shown, using a reduced set of acoustic features, that a better model to predict the likability ratings could be built using WB speech, compared to that using NB speech.

6.2. Paired-Comparison Listening Test

A second listening test [17] was conducted with partly the same speech stimuli and the same listeners as the experiment in [15]. Instead of using a continuous scale as in [17], a paired-comparison listening test was proposed for obtaining listeners' likability ratings. The intention was to detect and assess possible advantages of this method with respect to the direct scaling test, which typically contributes to undesirable low agreement between judges. In the paired-comparison task, the rater was presented with voice stimuli in pairs and was asked each time to determine which of each voice is preferred and to which extent, i.e. how much likable is the preferred utterance over the other.

It could be asserted that paired-comparison constitutes a reliable method for voice likability assessment while facilitating simple comparative judgments. The Bradley-Terry-Luce (BTL) probabilistic choice model [18, 19] could be successfully applied and ratio scale preference measures were derived. These preference measures correlated with $R^2 = 0.90$ with those obtained by the direct scaling test. While the direct scaling test leads to a somewhat lesser agreement between raters, it may still be generally preferred over the paired-comparison approach, given the considerable test length of the later.

7. Future Directions

A total of 101 speakers (47 males, 54 females) have already been recorded as of June 27th, 2016. Their mean age is 26.3 years with standard deviation 3.76. Out of the 101 participants, 68 are students, 20 are employed, and 13 gave other answers: 'Auszubildender', 'Arbeitssuchende', etc.

Our intention is to reach the recording of at least 200 speakers, which may suffice for training and testing automatic classification or regression systems in the frame of our project. It will be controlled that about half of them are female. The next step after completing the recordings is to label all data in clean conditions via listening tests regarding personality and likability. Then, part of the data (the speakers presenting extreme likability values and personality traits), will be transmitted through communication channels and the effects of the distortions on the subjective ratings will be analyzed. Furthermore, the influence of the channel degradations on the performance of automatic systems will be investigated.

We are very open to various forms of collaborations. Once our data recording is accomplished, they will be freely released to the scientific community under a license agreement. Also the results derived from these data such as raw scores, speech features, or other analyses can be shared with interested researchers.

8. Acknowledgements

The database recording and its associated studies are supported by the German Research Foundation (DFG, Grant FE 1603/1-1 to Laura Fernández Gallardo).

9. Bibliographie

[1] A. Vinciarelli and G. Mohammadi, "A Survey of Personality Computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[2] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "'Would you Buy a Car From Me?' – On the Likability of Telephone Voices," in *Interspeech*, 2011, pp. 1557–1560.

[3] G. Mohammadi and A. Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[4] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of Personality Traits from Human Spoken Conversations," in *Interspeech*, 2011, pp. 1549–1552.

[5] M. Mehl, S. Gosling, and J. Pennebaker, "Personality in Its Natural Habitat: Manifestations and Implicit Folk Theories of Personality in Daily Life," *Journal of Personality and Social Psychology*, vol. 90, no. 5, pp. 862–877, 2006.

[6] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A Database of Age and Gender Annotated Telephone Speech," in *Language Resources Evaluation Conference (LREC)*, 2010, pp. 1562–1565.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Interspeech*, 2005, pp. 1517–1520.

[8] F. Schiel and A. Baumann, "Phondat 1, corpus v. 3.4." Bavarian Archive for Speech Signals (BAS), Tech. Rep., 2006.

[9] T. Polzehl, "Personality in Speech - Assessment and Automatic Classification," Ph.D. dissertation, Technische Universität Berlin, Germany, 2014.

[10] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Interspeech*, 2012, pp. 254–257.

[11] B. Rammstedt and O. P. John, "Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[12] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A Survey on Perceived Speaker Traits: Personality, Likability, Pathology, and the First Challenge," *Computer Speech & Language*, vol. 29, no. 1, pp. 100–131, 2015.

[13] N. Dahlbäck, Q. Wang, C. Nass, and J. Alwin, "Similarity is More Important than Expertise: Accent Effects in Speech Interfaces," in *Conference on Human Factors in Computing Systems*, 2007, pp. 1553–1556.

[14] B. Weiss and F. Burkhardt, "Is 'Not Bad' Good Enough? Aspects of Unknown Voices' Likability," in *Interspeech*, 2012, pp. 510–513.

[15] L. Fernández Gallardo and B. Weiss, "Speech Likability and Personality-based Social Relations: A Round-Robin Analysis over Communication Channels," in *Interspeech*, 2016.

[16] D. A. Kenny, *Interpersonal Perception: A Social Relations Analysis*. New York, U. S.: Guilford Press, 1994.

[17] L. Fernández Gallardo, "A Paired-Comparison Listening Test for Collecting Voice Likability Scores," in *Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication*, 2016.

[18] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[19] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. New York, USA: Wiley, 1959.

The impact of animacy and rhythm on the word order of conjuncts in German

Isabelle Franz [1], *Gerrit Kentner* [1], *Frank Domahs*[2]

[1] Goethe - Universität Frankfurt
[2] Philipps - Universität Marburg

i.franz@emuni-frankfurt.de, kentner@lingua.uni-frankfurt.de, domahs@uni-marburg.de

Abstract

In this study we investigated the impact of two constraints on the linear order of constituents in children's speech production. Two types of constraints have been found to influence serialization in English speaking participants: a rhythmic (*LAPSE) and a semantic (ANIM) one. We tested 18 German children aged three to six years. Participants were instructed to produce coordinated bare noun phrases in response to picture stimuli (e.g., *Delphin und Planet)*. Disyllabic target words were controlled with respect to word stress and animacy.
Overall, children preferably produced animate items before inanimate ones, confirming findings of Prat-Sala, Shillcock and Sorace (2000) [1]. Furthermore, the order of the conjuncts was affected by the rhythmic constraint, such that disrhythmic constructions (resulting in a sequence of unstressed syllables, a so called stress lapse) were avoided. The latter results were significant when the factor animacy didn't vary and can be taken as evidence for the *prosodic licensing hypothesis* (Demuth 2007) [2]. In sum, our findings suggest a stronger influence of animacy compared to rhythmic well-formedness on conjunct ordering for German speaking children, as it was shown for English speaking adults by McDonald and colleagues (1993) [3].
Key words: child language, prosody, lapse, animacy, picture naming, word order

1. Introduction

Prosodic as well as semantic constraints have an effect on children's and adults' speech production. For language acquisition there are many reports indicating a preference for an alternating rhythm in speech, especially for the avoidance of lapses (two or more unstressed syllables in a sequence). Gerken (1996) [4] showed that toddlers are more likely to produce grammatically ill formed sentences when this leads to prosodic well formedness. Furthermore, there are many findings demonstrating a preference for animate referents to be produced before inanimate ones. An influence of this animacy constraint (ANIM) was shown (among others) by Prat Sala and colleagues (2000) [1], and Drenhaus and Féry (2008) [5].

McDonald and colleagues (1993) [3] revealed in their study with English adult participants that *LAPSE has an effect on the linear order of conjuncts as long as animacy doesn't vary as a factor. Hence, in English speaking adults' speech production, the animacy constraint has more power than the rhythmic constraint. If and how this interaction can be transferred to child language is still an open question. On the one hand, findings from speech perception suggest special importance of prosodic constraints in child language (Schröder & Höhle 2011 [5], Gutman et al. 2014 [6]). Furthermore, results from speech production support the *prosodic licensing hypothesis* (Demuth 2007 [2]). Studies show that toddlers omit segments or even functional words when omission leads to prosodically optimal structures. Such omissions even take place when they yield ungrammatical sentences or non-existing monosyllabic words (Domahs et al., 2016 [7]; Miles et al. 2015 [8]; Gerken 1996 [4]). Accordingly data from speech acquisition indicate that prosodic constraints including *LAPSE are effective in children's speech production. On the other hand, there is accumulating evidence showing that the animacy constraint ANIM affects child language (Demuth 2005 [9]; Prat Sala et al. 2000 [1]; Drenhaus & Féry 2008 [5]). However, so far, there are no studies comparing the influence of ANIM with *LAPSE in child language. Derived from findings on the importance of prosody in child language, one may assume a different interaction of *LAPSE and ANIM in child language as compared to adults – potentially reversing the ranking into *LAPSE > ANIM.

In our study, we examined this interaction between *LAPSE and ANIM in German preschool children's speech production.

2. Methods

We tested 18 children aged three to six years with normal language abilities (as confirmed by the TROG-D norms, Fox 2006 [9]). In a picture naming task, the children were instructed to produce coordinated noun phrases (e.g.: 'dolphin and planet'*)* without determiners or any prespecified order of the conjuncts. As target-items, we used 20 bisyllabic nouns with one corresponding black and white drawing each.
Target items varied in the factors stress pattern (trochaic, iambic) and animacy (animate, inanimate). Stimuli were diagonally arranged picture pairs matched for visual salience (which was controlled in a pretest). 30 picture pairs were presented in a square, separated by an invisible diagonal line running from top left to bottom right. Pictures did not touch the diagonal so that they were located bottom left and top right within the square. Further, picture pairs were presented in reversed spatial order (Fig. 1, left and middle panel).

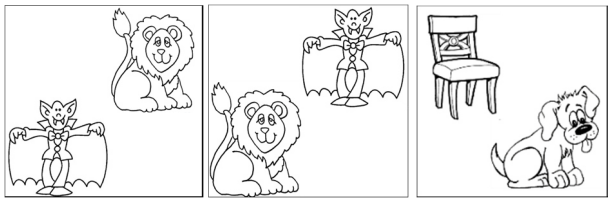


Figure 1: The left and middle panel show an example of an item pair in the two spatial orders and the right panel shows a filler pair.

We added 16 filler pairs arranged in the opposite direction to the item pairs (Fig. 1, right panel). Thus, filler item pictures were placed top left or bottom right inside the square with the invisible diagonal line running from bottom left to top right. Every participant was asked to name 46 picture pairs.

We analyzed the sequences the children chose, yielding either violations of *LAPSE (*Rátte und Planét, 'rat and planet'*), or ANIM *(Planét und Rátte, 'planet and rat'*), or both (*Hóse und Delfín, 'trousers and dolphin'*) or none (*Delfín und Hóse, 'dolphin and trousers'*) to examine the constraints' influence on sequencing the nouns within a phrase. Participants were familiarized with the target-items and particular stimulus-pictures in advance.

3. Results

On average, each of the 18 participants produced 21.3 valid conjoined noun phrases (range: 11 to 28) on the basis of 30 stimuli pictures. This corresponds to 385 valid phrases obtained in the experiment. In 61% of cases, children named the top right conjunct before the bottom left one, indicating a general preference for this order. However, no child used one of the orders exclusively.

A generalized linear mixed-effects regression model (glmer, Bates, Mächler, Bolker & Walker 2015 [10]) was employed in R statistical software (R Core Team 2015) to evaluate the effects of ANIM, *LAPSE, their interaction, as well as word frequency on word order. Overall, children preferably produced animate items before inanimate ones, showing a significant influence of ANIM on word order ($z = 4.654$, $p < 0.001$) as can be seen in Figure 2.

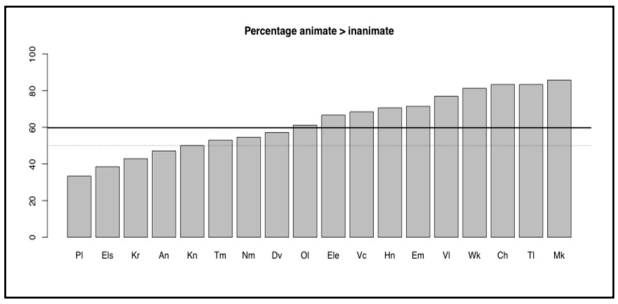


Figure 2: Overall mean percentage for animate > inanimate conjunct orders (solid line) with each bar representing individual percentages per child (chance line dotted)

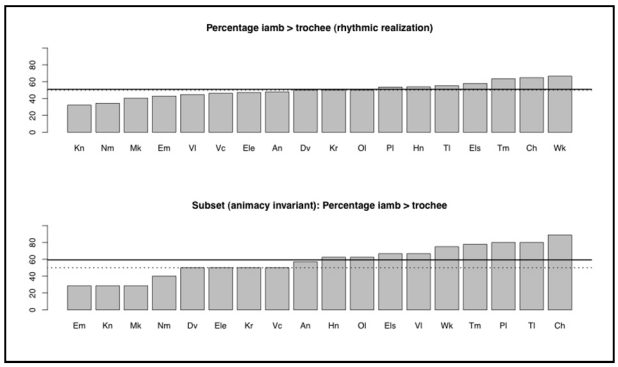


Figure 3: Top panel: Overall mean percentage (solid line) for rhythmic realizations (iamb > trochee). Bottom panel: Mean percentage for rhythmic realizations in subset with items not varying in animacy. Bars represent individual percentages per child.

Noun frequency had no significant influence on naming order. The prosodic constraint also showed some impact on the linear order so that *LAPSE constructions were avoided (Fig. 3). These results were only significant when animacy didn't vary as a factor ($z = 2.423$, $p = 0.0154$).

4. Discussion

Children preferably produced animate items before inanimate ones, showing a significant influence of ANIM on word order. These results are consistent with findings reported by Prat Sala and colleagues (2000) [1] for English speaking children. Furthermore, participants tended to name iambs before trochees, and by doing so, avoided disrhythmic structures, in line with the prosodic licensing hypothesis (Demuth 2007 [2]). However, the rhythmic effect only holds for the subset of data in which the two conjuncts did not vary w.r.t. animacy. That is, while conjunct order is demonstrably affected by ANIM and *LAPSE, the first constraint clearly outweighs the second one. This result conforms with the findings by McDonald et al. (1993) [3] for English speaking adults.

Given that the iambic stress pattern is less frequent than the trochaic one in German, iambic words may result in increased processing effort in speech production. As in our design participants had to name iambs before trochees to produce a rhythmically optimal phrase (*Klavíer und Rátte*), increased processing difficulty with the iambic structure could explain the rather week effect of *LAPSE (as in *Rátte und Klavíer*) on serialization. Indeed, Schiller et al. (2004) [11] found that English adults need more time for naming iambs than they do for naming otherwise comparable trochees. As a consequence, there could be counteracting effects of prosody in our design with shorter naming latencies for trochees (trochee > iamb) and a preference for iambs for rhythmic structures (iamb > trochee).

Working as linguistic constraints while naming two pictures in a conjoined noun phrase, ANIM and *LAPSE are effective on two different levels of speech planning – semantics and phonology. Taking classical speech production models into account (Levelt 1989 [12]), ANIM already becomes effective during semantic encoding whereas *LAPSE doesn't work until the later stadium of phonological encoding is reached. With animate items being more visually salient (Carniglia et al. 2012 [13]) this timing issue becomes even more interesting for comparing *LAPSE and ANIM as constraints in speech planning. It seems plausible, that for ANIM there are visual and linguistic (semantic) mechanisms which are becoming effective at an earlier stage in picture naming than the purely phonological ones for *LAPSE. In this way, ANIM might become effective before *LAPSE gets the chance to influence word order. These aspects could explain the sizes of effects in our study, i.e. the fact that the rhythmic constraint is only becoming significantly important for word order, when animacy is balanced out. However, we would like to hightlight the fact that *LAPSE *does* influence word order. With phonological encoding being assumed to happen after syntactic encoding in typical speech production models, such an impact of a prosodic constraint on word order is strongly suggesting an interaction of syntax and phonology.

5. Bibliography

[1] Prat-Sala, M. & Branigan, H. P. "Discourse Constraints on Syntactic Processing in Language Production: A Cross-linguistic Study in English and Spanish" *Journal of Memory and Language,* 42, pp. 168-182, 2000.

[2] Demuth, K. "Acquisition at the Prosody-Morphology interface" *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition North America* (GALANA), pp.84-91, 2007.

[3] McDonald, J. L.,Bock, K. & Kelly, M. H. "Word and World order: Semantic, Phonological and Metrical Determinants of Serial Position" *Cognitive Psychology*, 25, pp. 188-230, 1993.

[4] Gerken, L.A. "Prosodic structure in young children's language production" *Language,* 72, pp. 683-712, 1996.

[5] Drenhaus, R. & Féry, C. "Animacy and child grammar: An OT account" *Lingua*, 118, pp. 222-244, 2008.

[6] Gutman, A., Dautriche, I., Crabbé, B. & Christophe, A. „Bootstrapping the Syntactic Bootstrapper: Probabalistic Labeling of Prosodic Phrases", *Language Acquisition*, 22, 3, pp. 285-309, 2015.

[7] Domahs, F., Blessing, K., Kauschke, C. & Domahs, U. "Bono Bo and Fla Mingo: Reflections of speech prosody in German second graders' writing to dictation" *Frontiers in Psychology,* 7, 2016.

[8] Miles, K., Yuen, I., Cox, F. & Demuth, K. "The prosodic licensing of coda consonants in early speech: interactions with vowel length" *Journal of Child Language* 42, pp. 682– 694, 2015.

[9] Demuth, K., Machobane, M., Moloi, F. & Odato, C. "Learning Animacy Hierarchy Effects in Sesotho Double Object Applicatives" *Language* 81, pp. 421-427, 2005.

[10] Bates, D., Mächler, M., Bolker, B. & Walker, S. „Fitting Linear Mixed-Effects Models using {lme4}" *Journal of Statistical Software* 67, 1, pp. 1-48, 2015.

[11] Schiller, N.O., Fikkert, P. & Levelt, C.C. "Stress priming in picture naming: An SOA study" *Brain and Language* 90, pp. 231-240, 2004.

[12] Levelt, W. *Speaking: From Intention to articulation*, Cambridge, MA: MIT Press, 1989.

[13] Carniglia, E., Carputi, M., Manfredi, V., Zambarbieri, D. & Pessa, E. "The influence of emotional picture thematic content on exploratory eye movements" *Journal of Eye Movement Research* 5(4), pp. 1-9, 2012.

On the relationship between pointing gestures and speech production in German counting out rhymes: Evidence from motion capture data and speech acoustics

Susanne Fuchs [1], *Uwe D. Reichel*[2]

[1] Zentrum für Allgemeine Sprachwissenschaft, Berlin, Germany
[2] Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
fuchs@zas.gwz-berlin.de, uwe.reichel@nytud.mta.hu

Abstract

We investigated the interplay between pointing gestures and speech by means of motion capture and acoustics. Counting out rhymes served as a testbed, since they involve clear index finger turning points. The distance between the participant and an interlocutor (a teddy) was varied between close and far. Additionally, speaking with a normal speech rate in comparison to a fast rate was examined. Results of 1352 pointing gestures provide evidence that: a) the number of syllables realized per stroke are in general relatively stable across condition, but differences occur among subjects, b) turning points occur frequently in vowels, but also in consonants when syllables have a complex phonological structure, c) fast speech rate not only affects speech, but also leads to a shortening in pointing gesture duration, d) rhythmicity of the strokes is reduced with high speech rate and e) the impact of stroke rate on the acoustic energy contour is larger in normal than in fast speech. Distance showed no strong effects. We believe that counting out rhymes show great potential for further research in which further insight could be gained into the rhythmic and prosodic characteristics of a language as well as the coordination between pointing gestures and speech.

Index terms: pointing gestures, counting out rhymes, motion capture, syllable structure, rhythm, DCT

1. Introduction

This study investigates the relationship between manual pointing gestures of the index finger and speech acoustics in counting out rhymes. Counting out rhymes are an interesting area of investigation, because they naturally include pointing gestures with clear turning points and they can be recorded without constraining the motion of a speaker and therefore have a high external validity in comparison to very controlled, unnatural situations. Moreover, they are considered to be part of the oral poetry tradition found in many languages and cultures [7]. Kelly and Rubin [8] stated that rhythmic structure of poetry can provide evidence regarding the speech rhythm of a language. Finally, counting out rhymes are perfect games to investigate the acquisition of prosodic rules, since the rhymes are frequently used in childhood. The intention of the game is to select one person pseudo-randomly out of two (or a whole group). Since in counting out rhymes the number of syllables and words for each phrase often changes (see Table 1), it is to some extent unpredictable where the whole rhyme will end and which person will be "out". The person who counts speaks the rhyme and moves his/her index finger back and forth between the other and him/herself with very clear turning points. It is therefore an ideal testbed to investigate the relationship between speech production and pointing gestures. In particular it allows investigating the following questions:

(1) Does the number of turning points resemble the number of syllables? Is this behavior speaker specific or rather stable among speakers and conditions?

(2) Where do turning points occur within the speech flow (only in the vowel or also elsewhere)?

(3) To what extent is the relationship between speech and pointing gestures affected by time pressure and distance between two players?

These questions are motivated by the following proposals from the literature. Rochet Cappelan and colleagues [9] proposed an optimal 2:1 frequency relation between speech (jaw) and pointing gestures. They tested this relationship in an experiment in which adult participants had to point to a target while naming it. The target consisted of either 1, 2, 3 or 4 /pa/ syllables. Based on temporal measures of the pointing gesture and the jaw motions, the authors confirmed their proposal. They write: "… two syllables might be the maximum number of syllables that could be realized on one finger pointing motion without affecting the duration of the pointing period." (p. 5). We wish to extend their work to counting out rhymes and are particularly interested in the stability of this relation across speakers and tasks (questions 1 and 3). We would always expect speech production to be much faster than pointing gestures. Although the two belong to the same body of a person, the two motor control systems have very different mass (heavier for the arm), dynamic behavior (soft tissue dynamics for the tongue, joints for the arm and fingers), and space within they can move (much larger for pointing gestures).

Moreover, we are interested in how speakers coordinate their pointing gestures with the speech flow [6,10]. Krivokapic et al. (2016) [6] hypothesized that the maximum displacement of the index finger motion would be coordinated with the vocalic gesture or the tonal target of the stressed vowel, but not the consonantal gesture. They investigated the coordination between pointing gestures and oral gestures in bisyllabic words (CVCV), with either stress on the first or second syllable. Their findings provide evidence for a stable coordination of tonal targets (measured as f0 peaks) and the maximum displacement of the pointing gesture. We wish to follow up on this, but also to examine natural speech material which also contains voiceless portions as well as syllables with complex onsets and/or codas. It is unclear whether or not a similar coordination between pointing targets and the vowel occurs in closed or complex syllables (question 2).

2. Methodology

2.1. Experimental setup

Participants stood in front of a chair with a teddy bear and were instructed to play the counting out rhyme game with the bear as a fictitious person. We proposed various counting out rhymes, but recorded only the ones the participant knew (see Table 1). These counting out rhymes were recorded in four different conditions in successive order: a) close distance & normal speech rate, b) far distance & normal speech rate, c) far distance & fast speech rate, and d) close distance & fast speech rate.

In the close distance condition, the teddy was placed about 1m in front of the participant while in the far distance it was placed ca. 2m away. The latter two conditions also involved a faster speech rate. To evoke fast rate, subjects were instructed to imagine that they want to finish quickly with the counting out rhyme and start the following game.

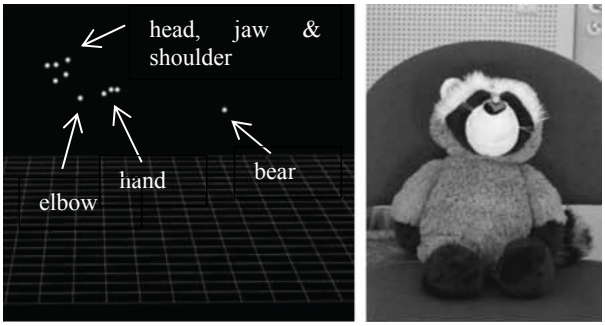


Figure 1: *Left: Display of a subject with markers located on different body parts pointing to a bear. Right: Teddy bear with a marker at the nose.*

Pointing gestures were measured by means of a motion capture system (OptiTrack, *Motive* Version 1.9.0) with 12 cameras (Prime 13). Motion data were recorded with a sampling rate of 200 Hz and a precision of 0.4 mm after calibration. One camera of the twelve was used as a video camera (200 Hz) to protocol the whole recording session. Acoustics were simultaneously recorded by means of a Sennheiser microphone. The sampling rate was 44.1 kHz.

Three markers were placed on a frontlet (one anterior, one posterior and one at the right lateral side), one marker was glued on the chin, one at the right shoulder joint (for right handers, and the left for left handed people), one at the wrist, one at the finger joint of the index finger, and one at the tip of the index finger. In the analyses we describe here, we focused only on x, y, and z motions of the index finger motion.

2.2. Participants and speech material

So far we have recorded five females with no known history of self-reported speech, language or hearing disorders. All were between 35 and 50 years old and worked in academia.

2.3. Preprocessing and annotation

Motive output files were saved in c3d format and subsequently converted into Matlab using the *Biomechanical Toolkit* [1]. The marker of the index finger was manually selected and it was checked for artefacts due to hidden movements during the relevant pointing gestures. To label the data and select the respective turning points, the 3D matrix was converted into a motion rate vector and saved as a wav-file. All data were then annotated in PRAAT (version 5.3.53 [2]). On the basis of the motion rate we labeled the velocity minima which correspond to the index finger endpoints (turning points). The interval between two turning points will hereafter be called a stroke. In a further step we added the text which was spoken within a stroke. If a turning point occurred in the middle of a word or sound, we added a full stop to the text marking the continuation of that sound to the next stroke. An example is displayed in Figure 2.

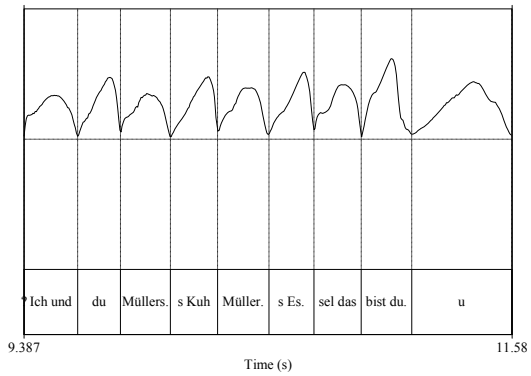


Figure 2: *Example of the motion rate vector. Vertical lines mark the labelled velocity minima. Lower track: the text of the rhyme (rhyme 1) was added for each stroke.*

Syllable nuclei were detected automatically. For this purpose the energy (root mean squared deviation: RMSD) in an analysis window was compared to the energy in a longer reference window with the same time midpoint moved along the bandpass filtered signal. If the RMSD in the analysis window was above a threshold relative to the RMSD in the reference window, a syllable nucleus was set. Further details of this procedure are provided in [3].

2.4. Location of turning point within the speech flow

A common proposal in the literature is that turning points of pointing gestures occur during vowels. Since vowels are produced with a high intensity in the acoustic envelope, we suppose that turning points would coincide with these values. For this purpose, we automatically extracted the intensity value of each turning point, subtracted it by the intensity minimum in the interval, multiplied it by 100 and divided it by the intensity range (as the difference between the max-min). The higher the output value (in percent), the closer the turning point to the intensity maximum within the stroke.

2.5. Quantification of rhythm

2.5.1. Rhythmicity: Pairwise modulo variability index

To quantify the rhythmicity of syllables and strokes we adopted the Pairwise Variability Index (PVI, [4]) a well-established measure of speech rhythm. This index measures the mean deviation of duration of neighbouring segments in an utterance. High PVI values indicate low rhythmicity and low PVI values high rhythmicity. For counting out rhymes however, this measure did not capture rhythmic variation based on varying assignment of beats per syllable or stroke. To give an example from rhyme 1, "Ich und du" consists of four

beats, “du” receiving two. Even though this results in a perfectly rhythmic four-fourth beat, PVI values would be large due to the large duration differences between the last two syllables. In order to take such a regular temporal variation into account, we extended the PVI to the Pairwise Modulo Variability index (PMI) which is normalized for speech rate (nPMI). Differences between the two indexes are illustrated in Figure 3. For a sequence X = $x_1 \ldots x_n$ (here: stroke intervals) it is calculated as follows:

$$\mathrm{nPMI}(X) = \frac{\sum_{i=1}^{n-1} [\frac{\mathrm{flip}(\max(x_i, x_{i+1}) \% \min(x_i, x_{i+1}))}{\min(x_i, x_{i+1})}]}{n-1}$$

’%’ is the modulo operator. The difference is not directly measured between the two duration values but between the larger value and the closest whole-number multiple of the smaller value. To allow this closest multiple to be larger than the greater value we had to extend the standard modulo calculation by a ’flip’ operation that replaces modulo values

$$v > \frac{\min(x_i, x_{i+1})}{2} \quad \text{by} \quad \min(x_i, x_{i+1}) - v$$

To give a concrete numerical example for this operation: the comparison of 1 with 0.4 (closest multiple 0.8) would yield a similar result as 1 with 0.6 (closest multiple 1.2). In both cases 0.2 is left.

The divisor $\min(x_i, x_{i+1})$ normalizes for speech rate, so that the nPMI is the same for pairs with the same relative durational difference, e.g. nPMI(1, 0.6) = nPMI (2, 1.2) = 0.33. As the PVI, also the nPMI gives the mean value of all pairwise differences.

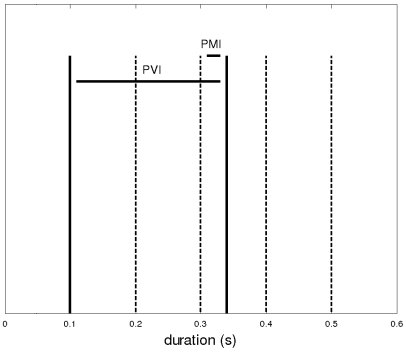


Figure 3: *Duration difference measurements PVI and PMI. PMI measures the difference between the larger value and the closest whole-number multiple of the smaller value.*

2.5.2. Influence of the strokes on the energy contour

To quantify the influence of strokes on the acoustic properties of the speech signal, we performed a digital cosine transform (DCT) on the energy contour [5]. The energy contour was taken, because it is a continuous signal and not interrupted, as e.g. fundamental frequency by voiceless events.

Then we calculated stroke influence *s* as the relative weight of the coefficients around the stroke rate *r* (± 1 Hz) within all coefficients below 10 Hz:

$$s = \frac{\sum_{c: r-1 \le f(c) \le r+1\mathrm{Hz}} |c|}{\sum_{c: f(c) \le 10\mathrm{Hz}} |c|}$$

Greater *s* values correspond to a greater impact of stroke rate on the energy contour.

3. Results

3.1. Relationship between number of syllables & strokes

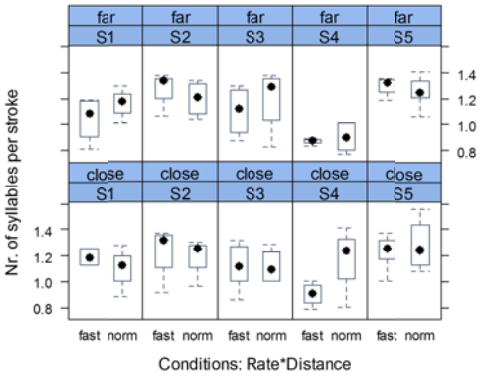


Figure 4: *Boxplots with number of syllables per stroke split by Speaker (S1-S5), Rate (fast, norm) and Distance (far, close).*

Figure 4 displays the results for how many strokes are on average spoken per pointing gesture. Except for S4, speakers are relatively stable across conditions, which speaks for a close link between the number of realized syllables and pointing gestures. The results are also in agreement with [9] that two syllables might be an upper bound for a pointing gesture.

3.2. Position of turning points within the speech flow

In our dataset 1352 pointing gestures were analyzed. In about 50% of the cases the turning points occurred between 80-100% within the intensity range, i.e. close to the intensity maximum each pointing gesture interval. We assume that these high intensity values correspond to vowels and confirm earlier proposals [6]. However, it is also evident that the other half of the data may not show turning gestures that correspond so clearly to the syllable nucleus, particularly in the fast rate. In comparison to previous studies, our speech material also involved coda consonants and complex syllable onsets. First visual inspections reveal that, particularly in closed syllables, turning points occur in coda consonants.

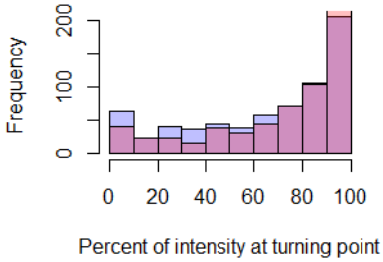


Figure 5: *Histogram for the percentage of the intensity at the turning point with respect to the overall intensity range within the stroke interval. Red: normal, Blue: fast.*

3.3. Effect of time pressure and distance

While distance shows no effect on the duration of the pointing gesture, speech rate does. Figure 6 (top) shows a shortening of the pointing gesture with faster speech. The shortening could be required if the number of syllables spoken in a stroke (Figure 4) should remain rather stable.

Furthermore, increased rate leads to a lower rhythmicity (higher nPMI values) in strokes (Figure 6 middle), and in normal speech rate to a wider distance. These effects also have consequences for speech (Figure 6 bottom). They result in a reduced impact of stroke rate on the acoustic energy contour.

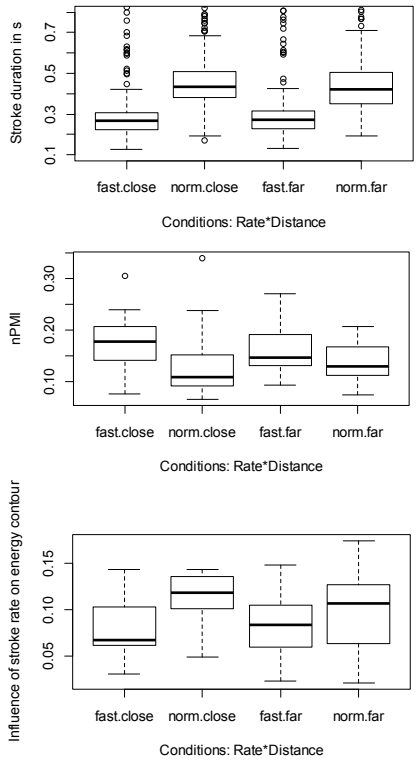


Figure 6: *top: Boxplots for stroke duration, middle: normalized pairwise modulo variability index, bottom: effect of stroke rate on energy contour split by rate and distance.*

4. Conclusion

The results of this study reveal a relatively stable, speaker-specific realization of the number of syllables per pointing gesture across conditions. When subjects speak faster, they also shorten their pointing gestures to maintain this ratio. Turning points are often produced within the vowel of a syllable, but can also occur elsewhere. Faster speech rate increases the likelihood of turning points being produced in consonants, probably because vowels are phonemes which are heavily affected by faster rate: they are reduced. Apart from the relative stability, fast speech rate reduces the rhythmicity of the pointing gestures and their impact on the energy contour of the speech signal. Such findings may be comparable to two motor systems with different properties, which adapt to each other, but can also reorganize with increased time pressure.

5. Acknowledgements

This work was supported by a grant from the BMBF to SF and by the Alexander von Humboldt society to UDR. Thanks to Jolanda Fuchs for providing her teddy bear Tom.

6. References

[1] A. Barré and S. Armand, "Biomechanical ToolKit: Open-source framework to visualize and process biomechanical data," *Computer Methods and Programs in Biomed.*, 114, 80–87, 2014.

[2] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer," Version 5.3.53, http://www.praat.org/

[3] U. Reichel, "Linking bottom-up intonation stylization to discourse structure," *Computer, Speech, and Language*, 28, pp. 1340–1365, 2014.

[4] E. Grabe and E. Low, "Durational variability in speech and the rhythm class hypothesis," in *Papers in Laboratory Phonology 7*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2002, pp. 515–546.

[5] C. Heinrich and F. Schiel, "The influence of alcoholic intoxication on the short-time energy function of speech," J. Acoust. Soc. Am., vol. 135, no. 5, pp. 2942–2951, 2014.

[6] J. Krivokapic, M. Tiede, M.E. Tyrone and D. Goldenberg, "Speech and manual gesture coordination in a pointing task", In *Proc. of Speech Prosody*, Boston, paper nr. 392, 2016.

[7] P.N.A. Hanna, P. Lindner, and A. Dufter, "The meter of nursery rhymes: universal versus language-specific patterns," In *Sounds and systems: studies in structure and change*, Berlin/New York: Mouton de Gruyter, 2002, pp. 241–267.

[8] M.H. Kelly, and D. C. Rubin. "Natural rhythmic patterns in English verse: Evidence from child counting-out rhymes." *Journal of Memory and Language,* 27.6, 718–740, 1988.

[9] A. Rochet-Capellan, J.L. Schwartz, R. Laboissiere, and A. Galvan "Two CV syllables for one pointing gesture as an optimal ratio for jaw-arm coordination in a deictic task: A preliminary study." *2nd EuroCogSci07*, 2007.

[10] A. Rochet-Capellan, R. Laboissiere, A. Galvan, and J.L. Schwartz, "The speech focus position effect on jaw–finger coordination in a pointing task." *JSLHR* 51.6, 1507–1521, 2008.

7. Appendix

Table 1. *5 counting out rhymes (1st column) in German orthography. Dots: syllable boundaries within words; Line breaks: prosodic boundaries; Number of syllables = 2nd column, Number of words =3rd column, the sums and ratio between syllables and words are displayed below the text.*

Orthographic representation of counting out rhymes		Nr. of syllables	Nr. of words
(1.) Ich und du		3	3
Mül.lers Kuh		3	2
Mül.lers E.sel		4	2
der bist du		3	3
Ratio =1.3	Sum	13	10
(2.) E.ne me.ne Mis.te		6	3
es rap.pelt in der Kis.te		7	5
e.ne me.ne Meck		5	3
und du bist weg		4	4
Ratio =1.47	Sum	22	15
(3.) Ei.ne klei.ne Dick.ma.dam		7	3
fuhr mal mit der Ei.sen.bahn		7	5
Ei.sen.bahn die krach.te		6	3
Dick.ma.dam die lach.te		6	3
eins, zwei, drei		3	3
und du bist frei		4	4
Ratio = 1.57	Sum	33	21
(4.) Ei.ne klei.ne Mic.ky.maus		7	3
Zog sich mal die Ho.sen aus		7	6
Zog sie wie.der an		5	4
Und du bist dran		4	4
Ratio = 1.35	Sum	23	17
(5.) E.ne, me.ne Mo.pel		6	3
Wer frisst Po.pel		4	3
Sau.er, süß und saf.tig		6	4
Für ne Mark und acht.zig		6	5
Für ne Mark und zehn		5	5
Und du musst gehen		4	4
Ratio = 1.29	Sum	31	24

Die akustischen und artikulatorischen Korrelate des /r/ im Norddeutschen. Eine Ultraschallstudie.

Riccarda Funk, Christina Otto

Institut für germanistische Sprachwissenschaft, Friedrich-Schiller-Universität Jena
riccarda.funk@uni-jena.de, otto.christina@uni-jena.de

Abstract

Die hier vorgestellte Studie untersucht, wie der /r/-Laut in der standardnahen norddeutschen Varietät im Silbenanlaut und im Silbenauslaut nach Langvokal artikulatorisch und akustisch realisiert wird. Im Silbenonset wurden Minimalpaare mit /r/ oder /h/ und in der Silbencoda Wörter mit /r/ nach Langvokal in Trägersätze eingebettet. Diese wurden artikulatorisch durch Ultrasound Tongue Imaging mittels *AAA* [1] und akustisch (F1, F2) von 10 Sprechern aus dem Norddeutschen Raum untersucht.
Die Analyse zeigt die Bildung einer dorso-uvularen Enge bei /r/ im Silbenonset im Vergleich zu /h/. Bei /r/ nach Langvokal kommt es zu einer Diphthongbildung, deren Ausprägung in Abhängigkeit zum Vokal auftritt. Je näher sich der Ausgangsvokal an [ɐ] befindet, umso schwächer ausgeprägt ist die Zungenbewegung.

Schlüsselbegriffe: Ultraschall, Norddeutsch, /r/, /r/-Vokalisierung

1. Einleitung

Varietäten des Deutschen weisen eine Vielzahl vokalischer und konsonantischer Korrelate von /r/ auf. Dabei gibt es zum einen eine systematische Alternation im Silbenonset und der Silbencoda, zum anderen einen Zusammenhang zwischen /r/-Laut, Dialekt und Sprechstil, weshalb die artikulierten /r/-Varianten auch innerhalb desselben Sprechers stark variieren können.
In standardnahen Varietäten wie dem Norddeutschen treten **konsonantische** Realisationen des /r/ im Wort- und Silbenonset vor Vokal und nach silbeninitialen Plosiven und Frikativen auf. Diese werden als dorso-uvulare Frikative artikuliert. Im Onset und nach stimmhaften silbeninitialen Plosiven werden sie stimmhaft [ʁ] oder entstimmt [ʁ̥] realisiert, nach stimmlosen silbeninitialen Plosiven oder Frikativen hingegen stimmlos [χ]. In postvokalischer Stellung nach Kurzvokal wird in der kanonischen Standardtranskription ein stimmhafter dorso-uvularer Frikativ [ʁ] transkribiert. In rheinländischen Dialekten wird nach Kurzvokal ein stimmloser uvularer Frikativ [χ] gebildet, in süddeutschen Dialekten ist im Onset die Verwendung eines alveolaren Vibranten [r] üblich, der zum alveolaren Flap [ɾ] reduziert werden kann. In sächsischen Dialekten kann das /r/ nach Kurzvokal durch Pharyngalisierungen ausgedrückt werden [2] [3].
Vokalische /r/ Realisationen treten in standardnahen Varietäten in unbetonten Silben mit /ər/, in den Präfixen <er-, her-, ver-, zer-> und in postvokalischer Stellung nach Langvokal auf. In unbetonten Silben mit /ər/ und in den genannten Präfixen werden zentralisierte Vokale produziert, deren Qualität von ihrer Lautumgebung abhängt und die einen großen Bereich im Vokalraum eines Sprechers abdecken. In postvokalischer Stellung nach Langvokal wird das /r/ zu einem zentralisierten Vokal reduziert und bildet mit dem Langvokal einen Diphthong. Die Vokalqualität des [ɐ] und die Dynamik des Diphthongs unterscheiden sich, je nachdem, welcher Langvokal vorausgeht. Je tiefer der Langvokal liegt, umso geringer ist die Bewegung der Zunge Richtung [ɐ], sodass ein Monophthong als /r/ Variante entstehen kann [2] [3].
Zu dieser Problemstellung gibt es bisher vor allem akustische Untersuchungen, beispielsweise von Simpson [3] und Barry [4], artikulatorische Analysen jedoch kaum. Schiller und Mooshammer [5] analysieren Reduktionsprozesse des /r/ im Deutschen mittels Elektromagnetischer Artikulographie und erkennen unterschiedliche Konfigurationen des Zungendorsums in Abhängigkeit von Vokal und Position des /r/-Lautes im Wort. Otto und Simpson [6] untersuchen silbenauslautendes /r/ nach Kurzvokal im Ostmitteldeutschen mit Hilfe von Ultraschall und belegen, dass das postvokalische /r/ hier zur Pharyngalisierung des Vokals führt und Einfluss auf die Zungenkonfiguration und Spektralstruktur des gesamten Vokals und angrenzende Konsonanten hat.
In dieser Studie sollen folgende Hypothesen untersucht werden:

1. Steht das /r/ in postvokalischer Stellung nach Langvokal, findet eine Vokalisierung des /r/ mit Diphthongbildung statt, bei der sich die Zungenkonfiguration im Ultraschall über das Segment /Vːr/ ändert. Die Stärke der Veränderung unterscheidet sich in Abhängigkeit zur Zungenhöhe- und Lage des vorhergehenden Vokals [2]. Bei /iːr/ und /uːr/ fällt sie sehr stark aus, bei /aːr/ hingegen schwach, dazwischen liegen die anderen Vokale. Die Diphthongbildung zeigt sich akustisch in der vokalabhängigen Veränderung der Formantwerte über das Segment hinweg.
2. Steht das /r/ im Onset, ist im Ulltraschall eine dorsale Enge sichtbar [5]. Bei /r/ sind Dorsum und Radix zurückgezogen, bei /h/ hingegen nicht.

2. Methode

2.1. Sprecher und Sprachmaterial

Für die Ultraschallaufnahmen wurden fünf männliche und fünf weibliche Studenten im Alter zwischen 19 und 26 Jahren ausgewählt, die aus Niedersachen, Schleswig-Holstein und von der Insel Rügen stammen und zurzeit in Jena wohnen. Keiner dieser Sprecher war bisher in logopädischer Behandlung oder leidet an Sprach-, Sprech- oder Hörproblemen.

Die Probanden lasen in fünf Durchgängen 45 randomisierte Sätze, die je ein Zielwort im selben syntaktischen Kontext enthalten.

Es wurden in der deutschen Sprache vorkommende Zielwörter und keine Pseudowörter verwendet, um die Natürlichkeit der Äußerungen zu gewährleisten, auch wenn der phonetische Kontext im Sprachmaterial dadurch nicht vollständig kontrolliert ist.
Ausgewertet wurden sieben Wörter mit /r/ nach Langvokal (Bier, Tür, Uhr, Meer, Stör, Moor, Paar) und sieben Minimalpaare mit /r/ und /h/ im Onset (Ritt/Hit, Rübe/Hübe, Rufe/Hufe, Recht/Hecht, hören/röhren, Rosen/Hosen, Rasen/Hasen).

2.2. Aufnahmen

Verwendet wurde das Ultraschallgerät DP2200 mit der Sonde 35C20EA bei einer Frequenz von 6,0 MHz und einer Tiefeneinstellung von 10,8 cm, die in einer internen Bildrate des Gerätes von 66 bps resultiert. Nach De-interlacing der einzelnen Bildrahmen beträgt die Videoframerate ungefähr 60 bps [7]. Die Akustik wurde mit einem Mikrophon AKG C1000S aufgezeichnet und zusammen mit dem Videosignal mit der *Articulate Assistant Advanced* Software [1] verarbeitet.

Die Sonde wurde mit einem *Head and Transducer Support System* [8] zwischen Kinn und Kehlkopf befestigt. Vor der Aufnahme wurde mithilfe einer aus medizinischem Kunststoff hergestellten Bissplatte die Sonde so ausgerichtet, dass die Bissebene möglichst horizontal und damit über Sprecher hinweg vergleichbar ausgerichtet ist. Von dieser Bissebene wurde am Anfang und Ende der Aufnahmen je ein Bild gemacht, um beide zu vergleichen und rotationsbedingte Fehler auszuschließen. Zusätzlich wurden durch Pressen der Zunge gegen den Gaumen, trockenes Schlucken und Wassertrinken Konturen des Gaumens aufgenommen, die in der Analyse als oberer Referenzpunkt dienen.

2.3. Datenauswertung

Die Segmentierung der Sätze erfolgte mittels *praat* [9] nach einheitlichen Regeln [10]. Bei /r/ nach Langvokal wurden die Segmentgrenzen bei Anwesenheit des 2. Formanten gesetzt. Bei /r/ und /h/ im Onset wurden die Grenzen bei sinkender Intensität und kleiner Amplituden im Oszillogramm gesetzt.

In *AAA* wurde für jeden Sprecher ein identisches Template angelegt, das zur Orientierung der Zungenposition dient. Über das Videobild wurde ein Gitter mit 42 radialen Achsen gelegt und eine Kurve anhand der Punkte erstellt, an denen die sichtbare Zungenkontur (weiße Linie im Ultraschallbild) diese Achsen schneidet. Dieser Prozess wurde semi-automatisch für alle Aufnahmen mit /r/ in postvokalischer Stellung nach Langvokal und im Onset mit dem in *AAA* implementierten Edge-Detection Algorithmus durchgeführt und manuell korrigiert. Mithilfe der Gaumenaufnahme wurde für jeden Sprecher eine Gaumenkontur als Referenz zur Zunge erstellt. Die Zungenkurven wurden in allen Fällen unterhalb der weißen Kontur gezeichnet, die Gaumenkurven oberhalb.

Die Zungenkurven bei /r/ nach Langvokal wurden über die Wiederholungen gemittelt und zwei Zeitpunkte (nach 25% und 75%) miteinander verglichen. Bei /r/ und /h/ im Onset wurde derselbe Zeitpunkt (50%) gewählt, die Kurven wurden ebenfalls gemittelt und miteinander verglichen.

Neben der bildlichen Darstellung wurde das quadratische Mittel des Kurvenabstands (root mean square, RMS) im Programm berechnet, um die Differenz beider Zungenkurven zu ermitteln. Der Vorteil ist hierbei, dass die RMS-Werte über alle Sprecher gemittelt werden können, was aufgrund anatomischer Unterschiede in der Graphik nicht möglich ist.

Für die akustische Auswertung der /r/ Varianten in postvokalischer Stellung nach Langvokal wurden die ersten beiden Formanten F1 und F2 jedes Sprechers in *praat* ermittelt. In Anlehnung an die artikulatorische Auswertung wurden die Formanten nach 25% und 75% des Lautsegments /V:r/ gemessen und ein Mittelwert über die Wiederholungen berechnet.

3. Ergebnisse

3.1. /r/ nach Langvokal

Der Vergleich der Zungenkurven beider Zeitpunkte im Segment /V:r/ zeigt bei allen Vokalen eine Veränderung der Position, da eine Diphthongbildung stattfindet. Richtung und Stärke der Zungenbewegung sind je nach Vokal verschieden.

Zur Veranschaulichung werden die Daten eines Sprechers mit den Eckvokalen /i:/, /u:/ und /a:/ vorgestellt.

Im Segment /i:r/ findet eine starke Zungenbewegung statt (vgl. Abb. 1). Die gesamte Zunge bewegt sich nach unten, Dorsum und Radix zudem nach hinten. Die RMS-Distanz beträgt hierbei 0,642 cm.

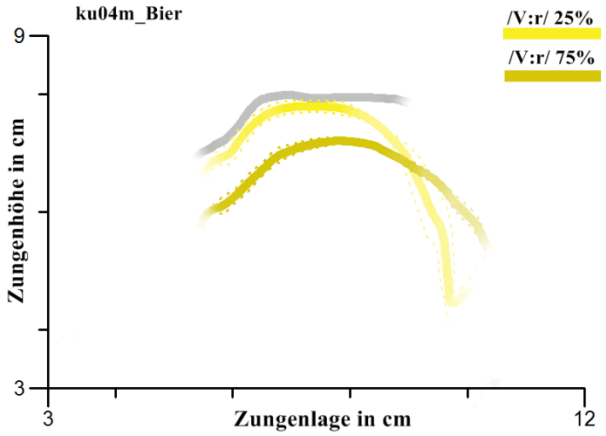


Abbildung 1: Segment /i:r/ des Sprechers ku04m, über alle Wiederholungen gemittelt. Die gepunktete Kurve zeigt die Standardabweichung, die graue Markierung die Gaumenreferenz.

Im Segment /u:r/ (vgl. Abb. 2) findet ebenfalls eine starke Zungenbewegung statt. Die gesamte Zunge bewegt sich nach unten, Apex und Radix schwächer als Lamina und Dorsum. Der RMS-Wert beträgt hier 0,668 cm, ähnlich wie bei /i:r/.

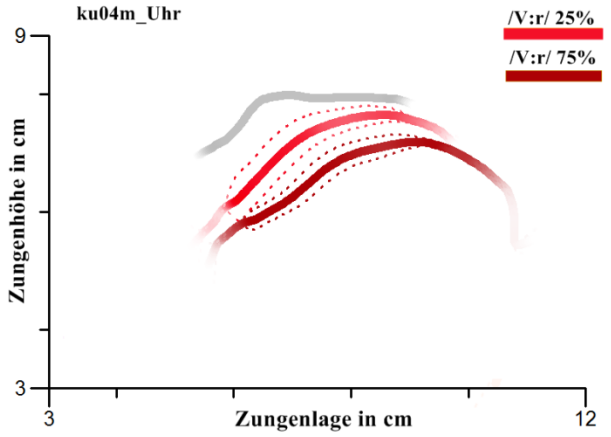


Abbildung 2: Segment /u:r/ des Sprechers ku04m, über alle Wiederholungen gemittelt. Die gepunktete Kurve zeigt die Standardabweichung, die graue Markierung die Gaumenreferenz.

Im Segment /a:r/ (vgl. Abb. 3) findet im Gegensatz zu /i:r/ und /u:r/ fast keine Zungenbewegung statt. Der RMS-Wert beträgt lediglich 0,101 cm.

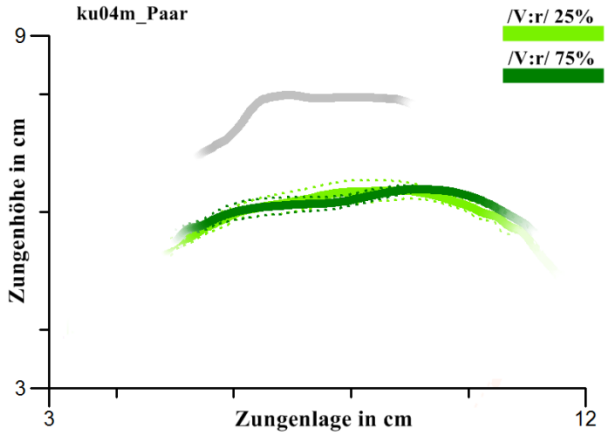


Abbildung 3: Segment /a:r/ des Sprechers ku04m, über alle Wiederholungen gemittelt. Die gepunktete Kurve zeigt die Standardabweichung, die graue Markierung die Gaumenreferenz.

In dieser Auswertung ist ein Zusammenhang zwischen Richtung und Stärke der Zungenbewegung in Abhängigkeit des Vokals im Segment /V:r/ nachweisbar. Allerdings fallen im Vergleich aller Probanden sprecherspezifische Unterschiede auf. Nicht bei allen Sprechern sind die RMS-Werte so differenziert wie bei ku04m, manche Sprecher zeigen nur geringe RMS-Unterschiede zwischen den Vokalen auf. Das zeigt sich in einer unterschiedlich großen relativen Standardabweichung, wenn die RMS-Distanzen der Sprecher über alle Vokale gemittelt werden. Am größten ist die relative Standardabweichung mit 58% bei ku04m, am kleinsten mit 29,4 % bei ku08m. Im Mittel sind die Zungenbewegungen bei /u:r/ und /i:r/ jedoch deutlich größer als bei /a:r/ (vgl. Abb. 4).
In allen untersuchten Segmenten /V:r/ außer /o:r/ treten zudem geschlechterspezifische Unterschiede insofern auf, dass die RMS-Werte der männlichen Sprecher im Mittel größer ausfallen als die RMS-Werte der weiblichen Sprecher (vgl. Abb. 4). Grund dafür können anatomisch bedingte Unterschiede sein. Da der Artikulationsraum bei Männern größer ist als bei Frauen [11], fallen die Bewegungsunterschiede größer aus.

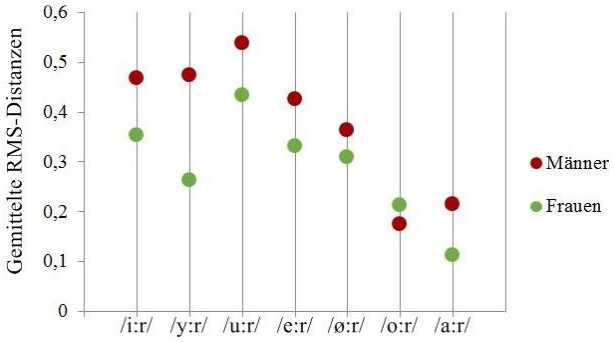


Abbildung 4: Gemittelte RMS-Werte der Vokale aller Männer und Frauen.

In der akustischen Analyse der beiden Zeitpunkte des Segments /V:r/ zeigt sich die /r/-Vokalisierung in einer Bewegung von F1 und F2. Abbildung 5 zeigt die Formanttransitionen eines Sprechers getrennt nach Vokal. Alle Formanten bewegen sich in Richtung eines zentralisierten Vokals. Bei ku04m tritt die größte Formantbewegung bei /u:r/ durch den starken Anstieg von F2 auf, die Zunge bewegt sich demnach stark nach vorne. F1 steigt durch die Zungenbewegung nach unten leicht an. Bei /i:r/ ist ebenfalls eine Bewegung der Zunge nach unten erkennbar, da F1 leicht ansteigt (vgl. Abb.1). Die Bewegung nach hinten führt zu einem Absinken von F2, die Bewegung fällt jedoch schwächer aus als bei /u:r/. Bei /a:r/ ist vor allem ein Absinken von F1 durch die Zungenbewegung nach oben erkennbar, F2 verändert sich kaum, was auf eine geringe Veränderung der vertikalen Zungenlage hindeutet (vgl. Abb. 3).

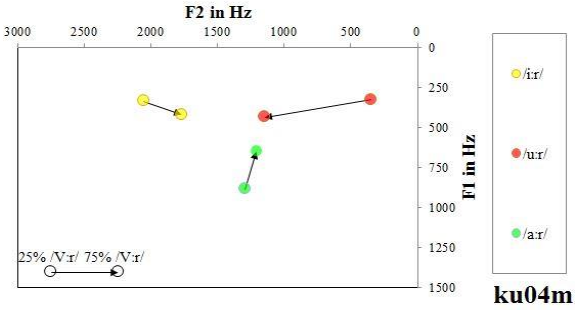


Abbildung 5: Formantbewegung des Sprechers ku04m.

3.2. /r/ im Onset

Beim /r/-Laut im Onset ist im Vergleich zu /h/ im Lautmittelpunkt bei allen Sprechern ein Unterschied in den artikulatorischen Daten erkennbar. Steht /r/ im Onset, sind Dorsum und Radix im Vergleich zu /h/ zurückgezogen. Darin lässt sich die dorsale Enge bei der Frikativbildung erkennen.
Auch hierbei treten Differenzen der RMS-Werte zwischen den Sprechern auf, die möglicherweise anatomisch zu erklären sind. Bei einem ohnehin geringen Abstand von Dorsum und Uvula muss die Zunge für eine Geräuschbildung nicht weit

nach hinten verlagert werden. Die Zusammenhänge zwischen Größe des RMS-Wertes und Sprechergeschlecht sprechen ebenfalls hierfür, da der RMS-Mittelwert der männlichen Probanden kleiner ist (0,448 cm, SD 14,7%) als der Mittelwert der weiblichen Probanden (0,523 cm, SD 10,5%).
Zur Veranschaulichung werden die Ultraschalldaten von /r/ und /h/ des Sprechers ku04m dargestellt (vgl. Abb. 6).

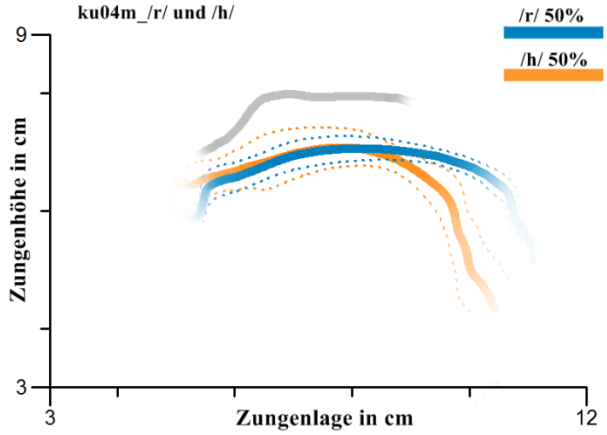


Abbildung 6: Vergleich von /r/ und /h/ bei Sprecher ku04m.

Das /r/ kann im Onset sowohl stimmhaft als auch stimmlos realisiert werden. Im Sonagramm wird ein Intensitätsabfall (gelbe Linie) sichtbar. Die Amplituden sind schwächer ausgeprägt als in der Lautumgebung. Häufig ist eine Geräuschbildung erkennbar. Je stärker die dorsale Enge gebildet wird, umso stärker treten diese Merkmale auf. Außerdem fällt auf, dass sich die Stimmhaftigkeit bei /r/ während der Lautdauer verändert. Viele /r/ Laute beginnen zunächst stimmhaft, werden dann stimmlos und kurz vor Vokalbeginn noch einmal stimmhaft (vgl. Abb. 7).

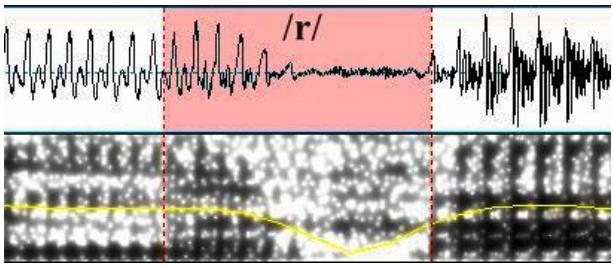


Abbildung 7: Oszillogramm und Sonagramm von /r/ des Sprechers ku04m im Satz <Ich habe den Rasen gemocht>.

4. Zusammenfassung

Die Auswertung der Daten hat bestätigt, dass bei /r/ in postvokalischer Stellung nach Langvokal eine Vokalisierung des /r/ mit Diphthongbildung stattfindet, bei der sich die Zungenkonfiguration im Ultraschall zu verschiedenen Zeitpunkten innerhalb des Segments /Vːr/ ändert. Die Stärke der Bewegungsänderung unterscheidet sich in Abhängigkeit zur Zungenhöhe- und Lage des vorhergehenden Vokals. Bei Segmenten mit hohen Vokalen wie /iːr/ und /uːr/ ist sie stärker als in Segmenten mit tieferen Vokalen wie /aːr/. Die /r/-Vokalisierung zeigt sich in der akustischen Analyse durch vokalabhängige Formantwertbewegungen in Richtung eines zentralisierten Vokals.
Steht das /r/ im Onset, ist im Ultraschall eine dorsale Enge sichtbar, die sich im Vergleich der Zungenkonfigurationen im Lautmittelpunkt von /r/ gegenüber von /h/ zeigt. Bei /r/ sind Dorsum und Radix zurückgezogen, bei /h/ hingegen nicht. Im Oszillogramm zeigt sich die dorsale Enge durch kleine Amplituden, im Sonagramm durch eine Abschwächung der Formanten und durch Friktion.
Bei der artikulatorischen und akustischen Analyse fällt auf, dass es in der Zungenbewegung starke sprecherspezifische Unterschiede gibt. Möglicherweise sind diese anatomisch zu erklären, was im Rahmen dieser Studie jedoch nicht untersucht werden kann.
Bei allen Ergebnissen ist zu beachten, dass die Stichprobe der Sprecher sehr gering ist. Durch die artikulatorische und akustische Analyse von zehn Sprechern können keine allgemein gültigen Aussagen über die Zungenbewegungen getroffen werden. Es können lediglich Gemeinsamkeiten, Unterschiede und sichtbare Tendenzen in den Bewegungen und zwischen den Sprechern aufgezeigt werden.
Die Wahl der untersuchten Zeitpunkte kann problematisch sein. Bei manchen Sprechern spiegeln diese Zeitpunkte die Zungenbewegung gut wieder, bei anderen wäre ein Vergleich anderer Zeitpunkte möglicherweise besser gewesen.

5. Bibliographie

[1] Articulate Instruments Ltd (2014). Articulate Assistant Advanced User Guide: Version 2.16, Articulate Instruments Ltd., Edinburgh, UK.
[2] Kohler, K. (1977). Einführung in die Phonetik des Deutschen. Berlin, Erich Schmidt Verlag.
[3] Simpson, A. P. (1998). Phonetik und Phonologie des deutschen r. In: Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung. AIPUK Nr. 33, Hrsg. von K. J. Kohler, Kiel. S. 105-147.
[4] Barry, W.J. (1997). Another R-tickle. Journal of the International Phonetic Association, Vol. 27 No. 1-2, S. 35.
[5] Schiller, N., Mooshammer, C. (1995). The character of /r/-Sounds. Articulatory Evidence for different reduction processes with special reference to German. Paper presented at ICPhS.
[6] Otto, C., Simpson, A.P. (2015). Pharyngealization of East Thuringian postvocalic /r/. Articulation, acoustics and temporal extent. Paper presented at ICPhS, Glasgow UK, available at: http://www.icphs2015.info/pdfs/Papers/ICPHS0712.pdf.
[7] Wrench, A.A. und Scobbie, J.M. (2006), "Spatio-temporal inaccuracies of video-based ultrasound images of the tongue", paper presented at International Seminar on Speech Production, Ubatuba, Brazil.
[8] Scobbie, J.M., Wrench, A.A. und van der Linden, Marietta (2008), "Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement.", in Sock, R. and Fuchs, Susanne. and Laprie, Yves (Eds.), *Proceedings of the 8th ISSP*, INRIA, Strasbourg, S. 373–376.
[9] Boersma, P., Weenink, D. (2013). Praat. Doing phonetics by computer. Version 5.3.45. Download: http://www.praat.org.
[10] Thomas, E. R. (2011). Sociophonetics. An Introduction. Basingstoke, Hampshire, Palgrave Macmillan.
[11] Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. Journal of the Acoustical Society of America 85. S. 1699–1707.

A Shadowing Experiment with Natural and Synthetic Stimuli

Iona Gessinger[1], Eran Raveh[1], Johannah O'Mahony[1]
Ingmar Steiner[1,2], Bernd Möbius[1]

[1]Saarland University
[2]DFKI GmbH
lastname@coli.uni-saarland.de

Keywords: phonetic convergence, human-computer interaction, shadowing task

1. Introduction

Inter-speaker accommodation is a phenomenon observed in human communication. Phonetic convergence is one way for a speaker to accommodate to an interlocutor. It is defined as an increase in segmental and suprasegmental similarities between two speakers [1]. Phonetic convergence has been found for human-to-human interaction in both spontaneous, conversational speech [1, 2] and non-conversational speech occurring in experimental settings such as the shadowing task [3, 4]. Previous studies on convergence in human-to-human interaction looked at suprasegmental features such as f0 range [5] and speaking rate [6], as well as segmental features such as spectral properties of vowels [3] and voice onset time [7].

Thus far, phonetic convergence has received little to no attention in the field of human-computer interaction (HCI). In the experiment introduced in this paper, we take a first step in investigating whether human speakers also converge to synthesized speech by conducting a shadowing experiment using both natural and computer-generated stimuli, concentrating on selected segmental features (cf. 2.1).

Based on previous findings in human-human interaction, we expect to observe phonetic convergence on the segmental level for the natural stimuli. Since the quality of synthesized speech is improving and HCI is becoming ever more used for various tasks in everyday life, humans are likely to interact in a similar way with computers as they do with humans. Therefore, we expect to observe convergence for the synthetic stimuli as well. However, the degree of convergence might still be influenced by the perceived naturalness of the synthetic stimuli.

2. Experiment

The following experiment consists of two conditions. In the first condition a group of participants is presented with short sentences recorded by natural speakers. In the second condition a different group of participants is presented with a synthesized version of the same sentences. The amount of convergence in the natural speech condition will serve as a baseline for the synthetic speech condition. Only the natural condition will be discussed in this paper.

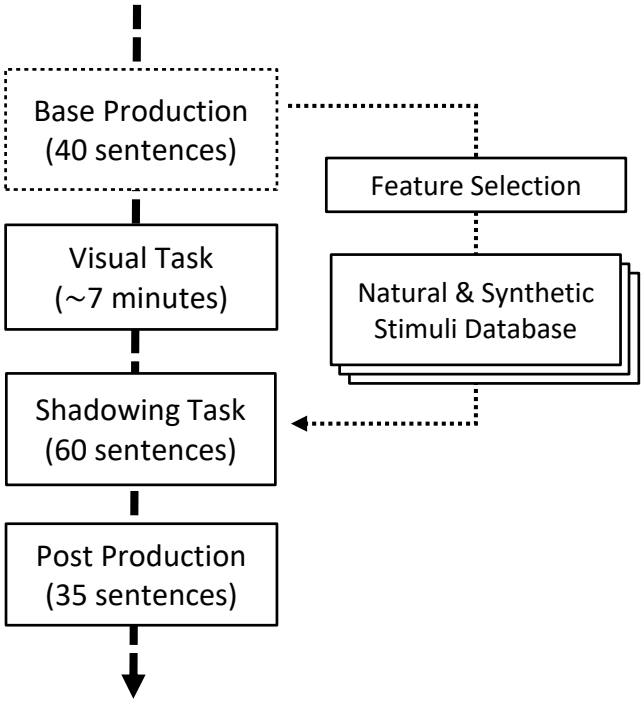


Figure 1: *Workflow of the experiment, showing its four phases. The stimuli presented in the shadowing task are selected based on feature realization in the baseline production.*

2.1. Target phenomena

To investigate convergence at the segmental level, three target features were selected that show variation across native speakers of German: [ɛː] vs. [eː] as a realization of the vowel *-ä-* in stressed position, as in *Gerät*, [ɪç] vs. [ɪk] as a realization of the final syllable *-ig*, as in *König*, and elision or epenthesis of [ə] in the final syllable *-en*, as in *reden*.

The former two features vary mainly regionally, with a preference for [eː] and [ɪç] in Northern Germany and a preference for [ɛː] and [ɪk] in Southern Germany [8, p. 64ff.]. All four forms are part of the phonetic inventory of standard German (*Pfirsich* [ɪç] / *Plastik* [ɪk] / *Säle* [ɛː] / *Seele* [eː]) and are hence expected to be known to all speakers. Despite the regional distribution of the two features, they are not strong dialectal markers [9, p. 560 for [ɪç]/[ɪk]] and often persist in otherwise standard German

Table 1: *Examples of target and filler sentences with corresponding target features.*

target sentence	target feature
Die Best**ä**tigung ist für Tanja.	[ɛː] vs. [eː]
Ich bin süchti**g** nach Schokolade.	[ɪç] vs. [ɪk]
Wir begleit**en** dich zur Taufe.	[ən] vs. [n̩]
filler sentence	
Der Kaffee war ja schon kalt.	—

productions of native speakers. Nevertheless, [ɛː] and [ɪç] are the official standard German forms of the respective features in the given contexts.

The elision of [ə] in the final syllable *-en* after plosives and fricatives is a highly expected phenomenon in standard German speech. In this position, [ə] is only produced when speaking particularly slowly and clearly [8, p. 39].

2.2. Stimuli

Thirty short German sentences (15 targets and 15 fillers) serve as stimuli for the shadowing task. Each target sentence contains one target feature only (i.e. five sentences per feature). The filler sentences do not contain any of the target features (cf. Table 1).

10 additional filler sentences are included in the baseline production, five of which are shown at the beginning of the recording to familiarize the participants with the task. The other five additional fillers contain tokens such as *Pfirsich* and *Plastik* to verify that participants were able to produce [ɪç] and [ɪk] respectively in a context other than *-ig*.

The first set of stimuli was recorded by two native speakers of German (1 female, 25 years old and 1 male, 23 years old). All 30 sentences were presented on a computer screen in random order. The speakers were instructed to speak naturally. Then the 15 target sentences were presented again, grouped by target feature. Possible feature variations as presented above were pointed out, and speakers were instructed to produce both variations. The best tokens regarding target feature production and overall clarity were selected for presentation in the shadowing task.

The second set of stimuli was generated using the text to speech system MaryTTS [10] with HMM synthesis.[1] One female and one male synthetic voice were used to match the gender of the natural speakers. In order to control for potential differences in information structure between the natural and the synthetic stimuli, prosodic characteristics of the synthetic stimuli were manipulated to match the natural stimuli.

2.3. Participants

21 native speakers of German (17 females, 19-33 years old, mean = 25.8, and 4 males, 23-34 years old, mean = 29.5) with no speech, language, or hearing impairments were recruited as participants for the first condition of the experiment. Another group of participants will be recruited for the second condition.

[1] http://mary.dfki.de/

2.4. Procedure

The experimental procedure consists of four tasks: baseline production, visual task, shadowing task and post production (see Figure 1). For the baseline production, 40 short sentences (15 targets, 15 fillers, and 10 additional fillers) were presented to the participants on a computer screen in random order. There were no instructions with respect to speaking style. Productions were recorded under the same conditions as model speaker productions. The realizations of the target features were noted by the experimenters during the baseline production. In order to weaken the mental representation of their first production, the participants were asked to perform a visual task after the baseline production.

In the shadowing task, the participants were presented with the productions of the two model speakers (15 targets and 15 fillers per model speaker; grouped by model speaker; semi-randomized for balanced distribution of targets over the two sets; alternating order of model speaker presentation). The target sentences played back to the participants always contained the opposite target feature realization of that observed in the participants' baseline productions (for instance, a participant who predominantly produced [ɪk], [ɛː] and elided [ə] in the baseline condition was exposed to [ɪç], [eː] and [ən] in the shadowing condition).

Words such as "repeat" and "imitate" were avoided in the instructions, so that converging behavior was not encouraged by the choice of words. Immediately after the shadowing task, participants were again presented with the written form of the stimuli to record the post production.

The second group of participants will undergo the same experimental procedure, but with synthetic instead of natural stimuli in the shadowing task.

3. Results

For a preliminary analysis of the target feature [ɛː] vs. [eː] the participants were divided into two groups based on their baseline production. The first two formants of the vowels were measured at mid-point and plotted along with the productions of the models the respective group heard in the shadowing task (cf. Figure 2).

For the group of participants which prefer [eː], the productions in the shadowing condition tend to move toward the model speakers' productions compared to the baseline condition. The same, slightly smaller effect can be observed for the post condition.

For the group of participants which prefer [ɛː], the productions in the shadowing condition also tend to move toward the model speakers' productions, but to a lesser extent than in the first group. The post condition, however, does not differ remarkably from the baseline condition for the second group.

These tendencies also become apparent when calculating the mean Euclidean distance between each of the participants' productions and the mean production of the models, using the formula

$$d(p,m) = \sqrt{(p_{F1} - m_{F1})^2 + (p_{F2} - m_{F2})^2} \qquad (1)$$

where p and m are points in the two-dimensional space

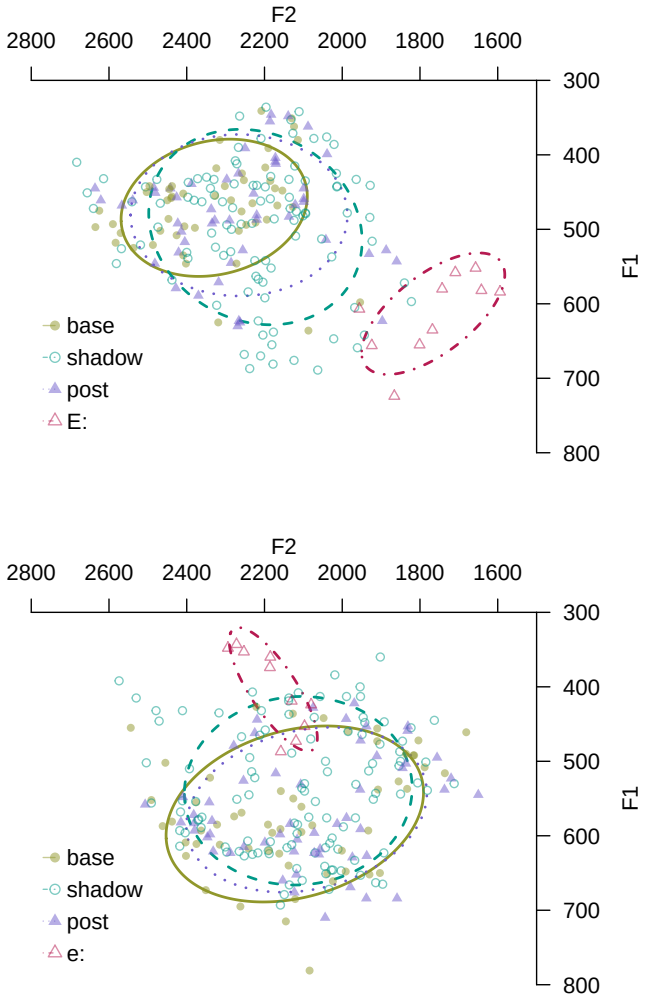


Figure 2: *Visualization[2] of target feature [ɛː] vs. [eː] produced by participants with baseline preference [eː] (upper figure; n = 11) and [ɛː] (lower figure; n = 10) in the production tasks **base**, **shadow** and **post**, as well as the human models (n = 2) they heard in the shadowing task, producing [ɛː] (upper figure) and [eː] (lower figure) (**red**). The ellipses visualize the confidence level of the estimated true mean (here: ±1 standard deviation).*

representing the production of the first two formants of vowels (see also Figure 2), p_{F1} and m_{F1} are the values of the respective first formants of the productions in Hertz, and p_{F2} and $mF2$ are the respective values of the second formants of the productions in Hertz. As this measure of distance is derived from frequency in Hertz, its unit will be called Hertz distance (HzD) in the following.

For the group of participants which prefer [ɛː] the Euclidean distance is 587 HzD (sd = 149) for the baseline condition, 482 HzD (sd = 175) for the shadowing condition, and 524 HzD (sd = 176) for the post condition. For the group of participants which prefer [eː] the Euclidean distance is 276 HzD (sd = 90) for the baseline condition, 241 HzD (sd = 92) for the shadowing condition, and 266 HzD (sd = 96) for the post condition.

The target feature [ɪç] vs. [ɪk] was categorically evaluated. Each segment was categorized either as a fricative (accounting for [ɪç]) or as a plosive (accounting for [ɪk]). The fricative category also includes the variations [ʃ] or [j̊], which were produced in a small number of cases. Figure 3 shows the distribution of changes between productions in baseline condition and shadowing condition. It comprises productions of speakers with plosive preference producing a fricative ([k] → [ç] convergence), productions of speakers with fricative preference producing a plosive ([ç] → [k] convergence), and productions that were the same category as in the baseline production of the speaker (no convergence).

In total, convergence was observed in 39% of the productions (20% [k] → [ç] and 19% [ç] → [k]). Participants' productions showed the following three patterns: consistent production in the baseline condition and consistent opposite production in the shadowing condition (i.e. complete convergence), same production in baseline and shadowing conditions (i.e. no convergence) or non-consistent productions in one or both of the conditions (i.e. partial convergence).

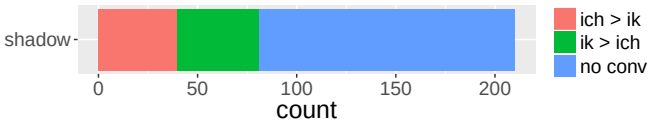


Figure 3: *Visualization of target feature [ɪç] vs. [ɪk] in the shadowing condition, showing cases where speakers changed their production compared to the baseline condition from fricative to plosive, from plosive to fricative, or didn't change their production.*

Regarding the target feature [ən] vs. [n̩], no participant showed a natural preference for schwa epenthesis in the baseline production. Hence, all participants were presented with the unreduced productions of the model speakers. For the preliminary analysis, each final syllable *-en* was acoustically and visually checked for epenthesis of schwa. A segment was only counted as schwa if it was longer than 30 ms.

Under this condition, the following number of schwa epentheses was observed: 2 out of 105 trials in the baseline condition (1.9%), 22 out of 210 trials in the shadowing condition (10.5%) and 4 out of 105 in the post condition (3.8%).

As in the case of [ɪç] vs. [ɪk], different individual production patterns were observed across the participants, ranging from no convergence to complete convergence.

4. Conclusion

We presented first results from an ongoing shadowing experiment with natural and synthetic stimuli. In this experiment, phonetic convergence on the segmental level is examined in the context of short sentences. Preliminary analysis of the natural condition shows convergence for all three target phenomena. The degree of convergence varied across the participants.

[2]Plots were generated using *phonR*, http://drammock.github.io/phonR/

5. Acknowledgments

We would like to thank Dr. Les Sikos for helpful suggestions concerning the experimental design and Dr. Sébastien Le Maguer for technical assistance in generating the synthetic stimuli.

6. References

[1] J. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[2] N. Lewandowski, "Talent in nonnative phonetic convergence," Ph.D. dissertation, Universität Stuttgart, 2012.

[3] K. Shockley, L. Sabadini, and C. Fowler, "Imitation in shadowing words," *Perception & Psychophysics*, vol. 66, no. 3, pp. 422–429, 2004.

[4] M. Babel, G. McGuire, S. Walters, and A. Nicholls, "Novelty and social preference in phonetic accommodation," *Laboratory Phonology*, vol. 5, no. 1, pp. 123–150, 2014.

[5] C. Smith, "Prosodic accommodation by french speakers to a non-native interlocutor," in *Proceedings of the XVIth International Congress of Phonetic Sciences*, 2007, pp. 313–348.

[6] J. Pardo, I. Jay, and R. Krauss, "Conversational role influences speech imitation," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2254–2264, 2010.

[7] A. Walker and K. Campbell-Kibler, "Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task," *Frontiers in Psychology*, vol. 6, no. 546, 2015.

[8] S. Kleiner and R. Knöbel, Eds., *Der Duden in zwölf Bänden*, 7th ed. Berlin: Dudenverlag, 2015, vol. 6: Duden - Das Aussprachewörterbuch.

[9] H. Mitterer and J. Müsseler, "Regional accent variation in the shadowing task: evidence for a loose perception-action coupling in speech," *Attention, Perception & Psychophysics*, vol. 75, no. 3, pp. 557–575, 2013.

[10] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: a tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, pp. 365–377, 2003.

Wenn Stotterer nicht stottern. Quantifizierung dynamischer Ultraschalldaten

Cornelia J. Heyde [1], *James M. Scobbie* [1]

Clinical Audiology, Speech and Language (CASL) Research Centre, Queen Margaret University, Edinburgh, UK

cheyde@qmu.ac.uk, jscobbie@qmu.ac.uk

Abstract

Stottern wird traditionell akustisch definiert, wobei man sich oftmals ausschließlich auf akustisch-perzeptive Unterbrechungen im ansonsten flüssigen Redefluss stützt. Die detaillierte artikulatorische Analyse der flüssigen Sprache von Stotterern soll Auskunft darüber geben, ob Elemente des Stotterns eventuell selbst dann nachweisbar sind, wenn sie akustisch nicht wahrnehmbar sind. Eine weitere Frage, die wir mit der dynamischen Analyse der Ultraschalldaten beantworten wollen, ist, wo genau sich Stottern manifestiert. Mit Hinblick auf die Fault-Line Hypothese von Wingate [1] untersuchen wir insbesondere die Bewegung der Zunge in die Verschlußstellung (Anglitt) und vergleichen diese mit der Bewegung im Übergang von der Verschlußstellung zum darauffolgenden Vokal (Abglitt). Die Ergebnisse unserer Untersuchung deuten an, dass sich Stotterer selbst in der scheinbar flüssigen Sprache von Kontrollsprechern unterscheiden. Artikulatorische Unterschiede zwischen den beiden Sprechergruppen wurden im Übergang von Konsonant zu Vokal beobachtet, was die Annahme Wingate's bestärkt, dass beim Stottern das Problem nicht auf einem bestimmten Wort oder Laut, sondern im Übergang von einem zum nächsten Segment, liegt.

Schlüsselbegriffe: Artikulation, , Motorkontrolle, Stottern, Ultraschall, Kinematik

1. Einleitung

Stottern ist eine Unterbrechung im Redefluss und wird häufig in drei Hauptsymptome unterteilt: Wiederholungen (/k-k-kafe/), Verlängerungen (/ʃ:::u:/) und Blockaden (/k---ɔp͡f/) [2]. Hier stellt sich die Frage, ob das Problem in der Realisierung des initialen Konsonanten selbst oder, wie von Wingate [1] vorgeschlagen, im Übergang von Konsonant zu darauf folgendem Vokal besteht. Wingate vertritt die Meinung, dass Unterbrechungen im Redefluss auftreten, wenn der Sprecher den auf den Konsonanten folgenden Vokal nicht problemlos integrieren kann. Die drei Hauptsymptome des Stotterns stellen jeweils akustische Ereignisse dar, die in Form von Unterbrechungen im ansonsten nicht-auffälligen Sprachfluss wahrgenommen werden. Hier stellt sich die Frage, ob Stottern wirklich, wie weithin angenommen [3], die Artikulation lediglich kurzzeitig lokal beeinflusst, oder ob es sich um eine motorische Störung handelt, die sich generell auf die Artikulation der Betroffenen auswirkt und somit die akustisch-perzeptiven Stotterereignisse eventuell lediglich die „Spitze des Eisberges" darstellen.

Diese beiden Fragen nach der Reichweite (i.e., lokal vs. global) und der Lokalisierung von Stotterereignissen (auf einem Segment oder im Übergang von einem zum nächsten Segment) wurden anhand von akustisch unauffälligen Ultraschalldaten untersucht. So wurde die flüssige Sprache von Stotterern mit der von Kontrollsprechern auf kinematische Unterschiede hin untersucht. Zwei Bewegungsmuster wurden hierfür quantifiziert [4] – zum einen die Bewegung in die Verschlußstellung (Anglitt) und zum anderen die Bewegung im Übergang von der Verschlußstellung zum darauffolgenden Vokal (Abglitt). Für beide Bewegungsverläufe wurden Parameter der Dauer und der erreichten Maximalgeschwindigkeit erhoben. Anschließend wurden die Ergebnisse der stotternden Erwachsenen mit denen der Kontrollsprecher (KS) verglichen. Wingate's „Fault-Line" Hypothese ließ Unterschiede zwischen beiden Sprechergruppen im Übergang von Konsonant zu Vokal (Abglitt) vermuten, wohingegen im Anglitt keine Unterschiede erwartet wurden. Die Quantifizierung der kinematischen Ultraschalldaten erlaubt es auch feinere Unterschiede zwischen den Sprechergruppen statistisch zu analysieren und auszuwerten.

2. Methode

2.1. Teilnehmer

Daten von neun stotternden Erwachsenen und neun Kontrollsprechern wurden erhoben. Die Stotterer waren mindestens 18 Jahre alt und berichteten alle von persistierendem idiopathischem Stottern, welches vor dem achten Lebensjahr begonnen hat. Alle Teilnehmer wurden zudem offiziell diagnostiziert. Zwei standardisierte Tests („*Stuttering Severity Instrument*" und „*Overall Assessment of Speaker's Experience of Stuttering*") gaben Auskunft über die Schwere (von mild bis stark ausgeprägt) des Stotterns zum Zeitpunkt der Datenerhebung.

Die Kontrollsprecher wurden bestmöglich mit den stotternden Erwachsenen bezüglich des Alters (20 bis 60 Jahre; Durchschnittsalter: SE: 34,4; KS: 33,6, Standardabweichung: SE: 14,2; KS: 12,2), des Geschlechts (SE und KS: 6 männliche und 3 weibliche Sprecher), der Bildung (höchster Bildungsabschluss) und der Händigkeit (Rechtshänder mit der Ausnahme eines Teilnehmers), sowie Muttersprache (Britisches Englisch) abgestimmt. Keiner der Teilnehmer berichtete über neurologische, Hör-, Seh- oder andere Störungen, die die Ergebnisse der Studie hätten beeinträchtigen können.

2.2. Datenerhebung

Audiosignale und kinematische Ultraschalldaten wurden parallel erhoben. Die Daten bestanden aus einsilbigen CV Silben mit velarem konsonantischem Silbenkopf /k/ und variierendem vokalischen Silbenkern. Der Silbenkern bestand aus den beiden Kardinalvokalen /a/ und /i/, sowie dem Zentralvokal Schwa. Jeder Silbenproduktion ging ein Schwa-Laut voraus, der es uns ermöglichte, den Anglitt (Artikulation hin zum konsonantischen Verschlusslaut /k/) für die kinematische Analyse

zugänglich zu machen. Die Daten wurden in neun Listen randomisiert und die gleichen Listen wurden von beiden Sprechergruppen produziert. Pro Sprecher wurden 12 Wiederholungen jeder der drei Silbenkombination (/kɑ/, /ki/, /kə/) aufgenommen. In einer Perzeptionsstudie wurden die Daten, die eindeutig für unflüssig (gestottert) befunden wurden, von der weiteren Analyse ausgeschlossen. Insgesamt wurden 637 Silbenproduktionen von allen 18 Sprechern analysiert.

2.3. Datenanalyse

Die Ultraschalldaten wurden mithilfe der AAA Software [5] mit 120 Bildern pro Sekunde aufgenommen und bearbeitet. Splines (mathematische Kurven) wurden der Zungenkontur mittels Edge-Tracking nachempfunden und in die Ultraschallbilder eingezeichnet (*Abbildung 1a)*. Die Splines einer CV Silbe wurden überlagert um die Koordinierung der Zungenoberfläche im zweidimensionalen Raum darzustellen (*Abbildung 1b*) und ein radialer Vektor wurde in der Region fixiert, in der die größte radiale Verschiebung (ausgehend von der Ultraschallsonde unterhalb des Kinns) der Zungenoberfläche beobachtet wurde (*Abbildung 1c*). Von der Verschiebung der Zungenoberfläche entlang des Vektors wurden zwei Verläufe abgeleitet (*Abbildung 2*): Die positive Verschiebung (mit größer werdendem Abstand zur Ultraschallsonde) wird in der Folge als „Anglitt“ bezeichnet. Sie stellt die Bewegung der Zungenoberfläche hin zum konsonantischen Verschluss dar. Die negative Verschiebung (mit kleiner werdendem Abstand zur Ultraschallsonde) wird in der Folge als „Abglitt“ bezeichnet. Sie stellt die Bewegung der Zungenoberfläche vom konsonantischen Verschluss hin zur vokalen Öffnung dar. Der Anfangspunkt des Anglitts, wie auch der Endpunkt des Abglitts wurden (in Anlehnung an die Methode vieler EMA Studien) mit einem Grenzwert von 20% der Maximalgeschwindigkeit bestimmt.

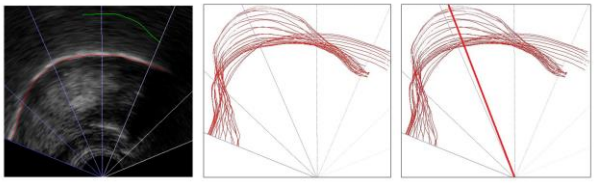


Abbildung 1: a) Ultraschallbild (Zungenspitze rechts; Zungenwurzel links) mit Spline (rote Kurve); b) überlagerte Splines relativ zu Anglitt und Abglitt; c) Fixierung des radialen Meßvektors (rote Linie) in der velaren Region

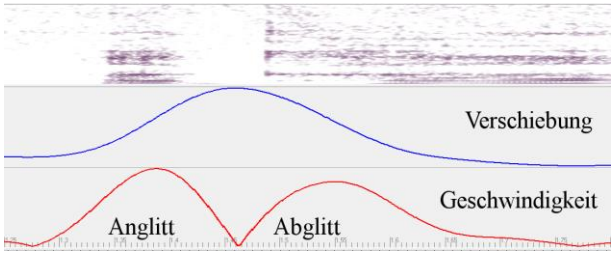


Abbildung 2: Artikulatorische Messungen von Dauer und Geschwindigkeit der Zungenoberfläche in der positiven (Anglitt) und negativen Verschiebung (Abglitt) entlang des Meßvektors für den Stimulus /ə ka/

Durchschnittswerte für die Dauer von Anglitt und Abglitt, sowie die Maximalgeschwindigkeit, die in Anglitt und Abglitt erreicht wurden, wurden für beide Sprechergruppen erhoben und verglichen.

3. Ergebnisse

Linear Paneldatenmodelle (mixed-effects models) wurden für die Analyse der Daten verwendet. Die Dauer (*Abbildung 3*) wie auch die Maximalgeschwindigkeit (*Abbildung 4*) wurden jeweils für An- und Abglitt auf signifikante Gruppeneffekte hin untersucht.

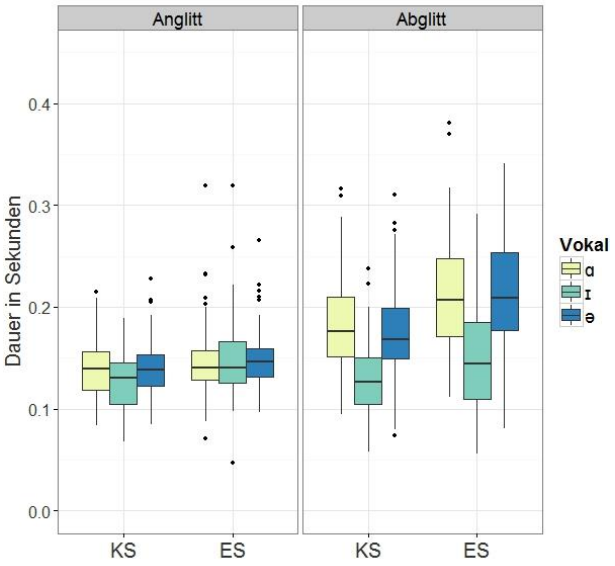


Abbildung 3: Durchschnittsdauer von Anglitt (Bewegung in die Verschlußstellung) und Abglitt (Bewegung im Übergang von der Verschlußstellung zum darauffolgenden Vokal) in Abhängigkeit von Sprechergruppe (Erwachsene Stotterer/ES im Vergleich zu den Kontrollsprechern/KS) und Vokal (/a/, /i/, /ə/)

Der beste Bestimmungskoeffizient (Model-Fit) für die Analyse der Dauer des Anglitts (*Abbildung 3* – linke Spalte) beinhaltete Sprechergruppe (ES, KS) als Fixed-Effekt und Vokal des Silbenkerns (/kɑ/, /kə/, /ki/) in Abhängigkeit von Individuum als Random-Effekt. Dem Model zufolge gab es keinen signifikanten Unterschied zwischen den beiden Sprechergruppen in der Dauer des Anglitts (β = -0.008, SE = 0.007, t = 1.130). Für die Dauer des Abglitts (*Abbildung 3* – rechte Spalte) hat das Hinzufügen von Vokal als Fixed-Effekt den Model-Fit signifikant verbessert ($\chi^2(2) = 0.980$, p = 0.613). Der beste Model-Fit enthielt somit Sprechergruppe (SE, KS), wie auch Vokal (/ɑ/, /i/, /ə/) als Fixed-Effekt wobei die Steigung für den Vokal je Sprecher variieren konnte (Random-Effekt). Der Abglitt von stotternden Erwachsenen dauerte demnach signifikant länger als der der Kontrollsprecher (β = 0.030, SE = 0.009, t = 3.143). Der Abglitt hin zum Kardinalvokal /i/ war im Durchschnitt signifikant kürzer als der zum Vokal /a/ (β = -0.056, SE = 0.009, t = -6.423), wohingegen sich der Abglitt zum Schwa nicht von dem hin zum /a/ unterschied (β = -0.005, SE = 0.007, t = -0.678). Das Hinzufügen der Interaktion von Sprechergruppe und Vokal hat den Model-Fit nicht verbessert.

Die Maximalgeschwindigkeit von Anglitt und Abglitt verhielt sich ähnlich wie die Dauer: Während im Anglitt (*Abbildung 4* – linke Spalte) kein Unterschied zwischen den Sprechergruppen beobachtet wurde (β = -4.482, SE = 11.684, t = -0.384 mit Sprechergruppe als Fixed-Effekt während die Steigung für den Vokal je Sprecher variieren konnte), wurde im Abglitt (*Abbildung 4* - rechte Spalte) ein Effekt von Sprechergruppe (β = -24.576, SE = 9.923, t = -2.477), sowie von Vokal des Silbenkerns beobachtet. Der beste Model-Fit beinhaltete Sprechergruppe und Vokal als Fixed-Effekt mit Vokal in Abhängigkeit von Individuum als Random-Effekt.

Mit Bezug auf die Maximalgeschwindigkeit, die während des Abglitts erreicht wurde, wurde eine signifikant geringere Maximalgeschwindigkeit hin zum Vokal /i/ im Vergleich zum Vokal /a/ erreicht (β = -38.167, SE = 9.704, t = -3.933). Die erreichte Maximalgeschwindigkeit im Abglitt zum Schwa war ebenso bedeutend geringer als die im Abglitt zum Vokal /a/ (β = -9.738, SE = 4.419, t = -2.204). Die Interaktion von Sprechergruppe und Vokal hat den Model-Fit nicht verbessert ($\chi^2(2)=0.269$, p = 0.874).

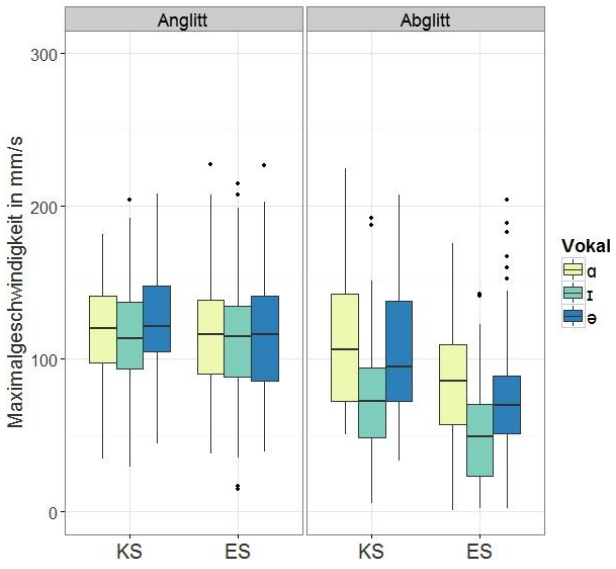


Abbildung 4: Maximalgeschwindigkeit von Anglitt und Abglitt in Abhängigkeit von Sprechergruppe (Erwachsene Stotterer/ES im Vergleich zu den Kontrollsprechern/KS) und Vokal (/a/, /i/, /ə/)

4. Diskussion

Unsere Ergebnisse zeigen, dass selbst die akustisch-perzeptuell flüssige Sprache von Stotterern sich in der Kinematik von der Sprache der Kontrollsprecher unterscheidet, was darauf hindeuten könnte, dass die akustisch-perzeptiven Stottereignisse lediglich die Spitze des Eisbergs darstellen. Die Bestimmung von Stottereignissen sollte daher nicht nur in der akustischen Charakterisierung bestehen, sondern kinematische Faktoren mit einbeziehen. Diese könnten eventuell unser Verständnis des Stotterns erweitern, indem sie zusätzliche Symptome (zu den drei vorherrschenden Hauptsymptomen) aufzeigen.

In der Lokalisierung einer möglichen „Fault-Line“, konnten wir Anhaltspunkte finden, die die Hypothese Wingate’s unterstützen. Wingate war der Annahme, dass nicht der initiale Konsonant selbst die Unflüssigkeiten im Sprechfluss auslöst. Stattdessen schlug er vor, dass Stottereignisse daher rühren, dass der folgende Vokal nicht problemlos integriert wird. Diese Annahme wird von unseren Daten unterstützt: Während die Sprechergruppen sich im Anglitt nicht wesentlich unterschieden, konnten wir für den Abglitt (Übergang vom konsonantischen Verschluss hin zum Vokal) sowohl für die Dauer, als auch für die erreichte Maximalgeschwindigkeit einen statistischen Effekt beobachten, was die Bedeutung der artikulatorischen Untersuchung unterstreicht.

Die Methode, die wir präsentiert haben, verbindet Aspekte der traditionell statischen Ultraschall-Datenanalyse mit denen von dynamischen Fleshpoint-Verfahren wie zum Beispiel der elektromagnetischen Artikulographie. Als Ausgangspunkt bietet das Ultraschallbild umfangreiche Informationen über die Zungenform im zweidimensionalen Raum. Die Aneinanderreihung mehrerer Ultraschallbilder bei hoher Bildrate bietet zudem eine zusätzliche temporale Dimension, die Einblicke in die Koordination der Artikulation bietet. Der Meßvektor wird an der Kinematik der Zunge selbst und somit an einem den Daten internen sprecherunabhängigen Referenzpunkt orientiert [6]. Dies berücksichtigt die artikulatorische und anatomische Heterogenität des einzelnen Sprechers, was eine relativ objektive Quantifizierung der kinematischen Kennwerte und somit Vergleiche von Sprechergruppen erlaubt.

Die Tatsache, dass wir mit dieser Methode selbst feine akustisch nicht erfassbare Unterschiede in der scheinbar flüssigen Sprache beider Gruppen festhalten und messen konnten, wird als Bestätigung dieser Methode interpretiert. Zudem unterscheiden sich die Messungen beider Sprechergruppen erwartungsgemäß in Abhängigkeit der verschiedenen Silbenkerne: Während im Anglitt keine signifikanten Unterschiede messbar waren, werden im Abglitt Effekte von Koartikulation ersichtlich, wobei sich Verläufe zum hohen Vokal /i/ durch eine signifikant geringere Dauer als auch geringere Maximalgeschwindigkeit hervorhoben. Letzteres bestätigt den Anspruch der Präzision dieser Methode.

5. Dank

Die Autoren danken Prof. Alan Wrench und Steve Cowen für deren stetige Unterstützung.

6. Bibliographie

[1] Wingate, M. E. (1988). The structure of stuttering: A psycholinguistic analysis. Springer Verlag, New York, NY.

[2] Alpermann, A. (2010) Redeflussstörung: Stottern–eine psychische Störung? Sprache Stimme Gehör, 34(2), 56

[3] Kaufmann, S. (2006) Idiopathisches Stotterin – Diskussion vor dem Hintergrund eines psycholinguistischen Modells der Sprachproduktion. München, GRIN Verlag.

[4] Tasko, S. M., & Westbury, J. R. (2002). Defining and measuring speech movement events. Journal of Speech, Language, and Hearing Research, 45(1), 127-142.

[5] Articulate Instruments Ltd (2012). Articulate Assistant Advanced User Guide: Version 2.14. Edinburgh, UK: Articulate Instruments Ltd.

[6] Iskarous, K. (2005). Patterns of tongue movement. Journal of Phonetics, 33(4), 363-381.

The realisation of Albanian laterals in German as a second language: A case study

Bettina Hobel, Sylvia Moosmüller, Christian Kaseß

Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

bettina.hobel|sylvia.moosmueller|christian.kasess@oeaw.ac.at

Abstract

In this case study, it is investigated how laterals spoken by one female Albanian speaker are realised in German as a second language. Standard Albanian features two lateral phonemes, an alveolar and a velarised one, whereas Standard Austrian German (SAG) has only one phoneme but a variety of possible realisations ranging from alveolar laterals in the standard language to velarised laterals, retroflex laterals, and l-vocalisations in the dialects. The results of our Albanian speaker show that she mostly produces alveolar laterals, occasionally with a tendency towards palatalization. Only two velarised laterals occurred, both in position between consonants. Hence, it can be concluded that the Albanian speaker abandoned one of the two Albanian lateral phonemes, namely the velarised lateral, and instead, she only uses the alveolar lateral phoneme of SAG. Concerning the l-vocalisation and retroflex laterals, as found in the Styrian dialects, only a few instances occurred; in general, the realisation of l-vocalisation was restricted to unstressed positions. The phonetic context has significant influence on the F2 of the lateral, which mainly involves differences between previous and following front vowels versus consonants. F2 of laterals preceded or followed by front vowels is higher than preceded or followed by consonants.

Index Terms: laterals, German as a second language, Austrian varieties

1. Introduction

1.1. Albanian laterals

Standard Albanian features two lateral phonemes, an alveolar lateral approximant[i] /l/ and a velarised lateral /ɫ/. Two distinct graphemes correspond to them, <l> and <ll>, respectively. Palatograms depicted in [1, pp. 415, 418] show a contact in the alveolar region for the alveolar lateral, whereas the contact is produced in the dental-alveolar region for the velarised lateral.

Regarding acoustics, various studies have yielded different results, especially as concerns the alveolar lateral. As becomes evident from Table 1, all studies report a rather low F2 for the velarised lateral. However, regarding the alveolar lateral, [1, p. 416] reports a substantially higher F2 than [2], Bothorel [1969-1970, cited after 3], and Jubani-Bengu [2012, cited after 3]. A reason for these differences might be the small number of speakers analysed in Beci (three male speakers from Skodër, Gheg variety), Bothorel (one male speaker, Gheg variety)[ii], Jubani-Bengu (two male speakers, Gheg variety). [2] analysed a total of 13 male speakers from the Tosk and the Gheg variety as well as from the transition zone.

It has to be mentioned that until now, formant frequencies have been reported for male speakers only. However, in the current study, we investigate the laterals of a female speaker.

Table 1. *Formant values for alveolar and velarised laterals in Albanian.*

	Beci [1]	Moosmüller et al. [2]	Bothorel and Jubani-Bengu [cited in 3, p. 3[iii]]
Alveolar			
F1	290-395 Hz	310 Hz[iv]	354 Hz
F2	1750-1890 Hz	1501 Hz	1586 Hz
Velarised			
F1	410-420 Hz	354 Hz	355 Hz
F2	1000-1100 Hz	1069 Hz	934 Hz

1.2. Laterals in Styria

The dialectal situation in the Austrian federal state of Styria is not straightforward since this region belongs to a dialectal transition zone where two main dialect regions meet. For the current investigation, the dialects of the following two locations have to be considered due to the previous and the current residences of the Albanian speaker: the district of Leoben and the city of Graz.

Both belong to the area of Southern Middle Bavarian, a transition zone between Middle and Southern Bavarian varieties [4, supplementary map 1]. Within this zone, the influence from Middle Bavarian, as spoken in Upper and Lower Austria [4, p. 1] became stronger during the last decades, thus isoglosses have been moving southwards. I.e., Middle Bavarian features are expanding into the area of Southern Middle Bavarian. This is true for l-vocalisation, a characteristic feature of Middle Bavarian. [4, supplementary map 4] and [5] drew the isogloss of l-vocalisation in the mountainous region between the Enns valley and the Mur/Mürz valleys. Recently, however, it was shown that the area of the Mur and Mürz valleys is now a region of l-vocalisation as well [6].

Thus, in the district of Leoben, besides the alveolar lateral, which is pronounced word-initially and between vowels, two further pronunciations of the lateral are prevalent: On the one hand, the retroflex lateral as a typical characteristic of the Southern and the Southern Middle Bavarian dialect region [6; 7, p. 340; 8, p. 92], on the other hand, the vocalisation of the lateral, intruding from the northern federal states (Middle

Bavarian). Therefore, it is assumed that elderly speakers mostly pronounce the retroflex lateral in the positions where, since recently, younger speakers apply l-vocalisation.

Both the retroflex lateral and the vocalisation of the lateral are restricted to the same word positions: before consonants and in word-final positions [9, p. 1112; 10, p. 840]. Front vowels preceding a lateral are rounded. In the case of vocalisation, the following processes apply: The lateral is vocalised to a front vowel, either [i̯] or [e̯] and, finally, absorbed by the front vowel, rendering e.g., [vy:e̯] "wild" *wild*. After back vowels, the vocalised lateral is retained, e.g. ['kɔe̯n̩] "Kohle" *coal* [11, p. 342].

However, contrary to the pronounced dialectal speech behaviour characteristic for the district of Leoben, speakers of the city of Graz, especially middle and upper class speakers, gear towards the pronunciation norms of Standard Austrian German (SAG). Therefore, in the district of Leoben, most probably, l-vocalisation is prevalent, whereas in the city of Graz, the alveolar lateral of Standard Austrian German is used in these positions [6]. The mean F1 and F2 of the alveolar lateral as spoken in SAG spontaneous speech, measured from three female speakers born and raised in Vienna and having an academic background, is 319 Hz for F1 and 1753 Hz for F2 [preliminary results presented in 12].

Therefore, in the first years of her stay in Austria, our Albanian speaker, who had already studied German in Albania, was exposed to l-vocalisation, a phonological process she was not familiar with on the basis of her previous studies, and additionally to the retroflex lateral, as mainly used by elderly people. By intense contact with the people living in the district of Leoben – she worked in an inn in a village – she soon learnt the local dialect. However, when she came to the city of Graz, she was confronted with a kind of speech behaviour that resembled to what she had learnt at the university. With respect to lateral production, this means that she was confronted with less instances of dialectal l-vocalisation or retroflex laterals.

1.3. Transition of Albanian laterals to German

When acquiring a second language, a new phoneme system has to be acquired as well. Most literature refers to the acquisition of new phonemic contrasts in L2, whereas research on non-contrastive sounds or allophones in L2 is scarce.

One such study investigating how L1 English speakers realise laterals in their L2 Spanish was done by [13]: American English features two lateral allophones, a "light" and a "dark" one in auditory terms, acoustically differing in F2, their use depending on syllable position. In contrast, Spanish has one "light" lateral phoneme. Thus, L2 learners of Spanish have to ignore the distinction made in English. The results showed that the better the learners master the second language Spanish, the more they assimilate to the Spanish /l/; however, they do not achieve lateral productions identical to Spanish natives. Nevertheless, at a high level of Spanish acquisition they stop distinguishing different variants by syllable position as done in English. It has not been investigated whether the slightly different lateral production of advanced learners from those of native speakers affects perception or constitutes a foreign accent.

The pilot study presented in this paper is, to some extent, comparable to the above-mentioned study by [13], the difference being the phonological status of the laterals in English and in Albanian (allophonic vs. phonemic, respectively). However, in the same way as in [13], the target language features only one lateral phoneme, namely an alveolar lateral approximant [11, p. 342].

2. Aim of the study

This case study attempts to show how laterals spoken by one Albanian speaker are realised in German (as a second language).

As described above, Albanian has two lateral phonemes, <l> [l] und <ll> [ɫ], thus it will be examined whether in L2-pronunciation one phoneme is abandoned, or both are maintained in the phonetic output, most probably in dependence of word position or phonetic context. Additionally to the SAG alveolar lateral, the speaker, due to dialectal influences, could also produce instances of retroflex laterals or l-vocalisations.

Hypothesis 1: The speaker produces an alveolar lateral, according to Standard Austrian German.

Hypothesis 1a: No velarised laterals occur.

Hypothesis 1b: Single instances of vocalised or retroflex laterals occur, due to dialectal influences.

Hypothesis 1c: Differences in the lateral's formant frequencies exist depending on the vowel context. Between front vowels, a slight palatalisation can be observed, manifested in a higher F2, whereas between back vowels, alveolar realisations or even velarised laterals occur, showing a lower F2.

3. Method

Data for this study were collected within the project "Styrialects", at the Department of Linguistics, Karl-Franzens Universität Graz. A recording of one Albanian female speaker was conducted: it comprises a semi-spontaneous interview (approx. 50 min.), a questionnaire, and a picture naming task. The current investigation is based on an approx. ten minutes speech of the semi-spontaneous interview.

The speaker was born and raised in Elbasan (Albania), which, from a dialectological point of view, belongs to the transition zone between the Gheg and the Tosk variety. In high school, she learnt German, though the standard variety as spoken in Germany. During the first four years of living in Austria, she had intense contact with Styrian dialect speakers from a village in the district of Leoben, thus, she acquired this dialectal variety. Finally, when moving to the city of Graz about twelve years ago, she additionally had contacts with SAG speakers; thus, she took the opportunity to incorporate this variety as well. As a conclusion, the speaker has a near-native competence in SAG and in a specific dialect, her foreign accent being limited to fine phonetic details.

For this analysis, an auditive and an acoustic analysis were performed. Segmentation, annotation, and acoustic analysis were carried out with STx [14]. All voiced laterals were segmented manually (n=172); some cases had to be excluded from further analysis due to poor signal quality. Annotation comprised word and syllable position of the lateral, stress of the syllable containing the lateral, and the phonetic context. F1, F2, and F3 were extracted over time by means of LPC (window length 46 ms, overlap 95 %). Although F1-F3 were

extracted, we focussed on the analysis of F2, since F2 is the most reliable cue for the degree of velarisation [15; 16].

For statistical analysis, a linear mixed-effects model was fitted to the data of F2 with "following context", and "previous context" as fixed factors, and "word" as random factor. Linear mixed-effects modelling was carried out in *R*, version 3.1.1, [17], using the *lmer*-function of the *lme4*-toolbox [18]. Significance of the factors was determined using the Type-III ANOVA of the *lmerTest*-toolbox [19]. In case of significant effects, pairwise Tukey post-hoc tests were performed on all possible combinations with regard to the significant effect using the *lsmeans*-toolbox [20]. Additional random factors (word position, syllable position, and stress of the syllable containing the lateral) did not contribute significantly to the statistical model and were therefore not included into the model.

The Type-III ANOVA was used as the data is unbalanced with regard to the previous context and following context (back vowel/ front vowel/ consonant):

Table 2. *Contingency table regarding the previous and following context.*

		Following context		
		Back v.	Front v.	Cons.
Prev. context	Back v.	7	14	17
	Front v.	7	39	33
	Cons.	11	34	6

In the auditive analysis, we concentrated on instances of vocalisation, as typical for Middle Bavarian dialects, and retroflex laterals, as used in Southern Bavarian dialects. These cases will be analysed qualitatively as they are interesting due to possible dialect influences.

4. Results

For an overview, Table 3 presents the mean formant frequencies and standard deviations of the second formant according to the factors previous context and following context.

Table 3. *Sample mean (stand. deviation) for F2 [Hz]*.

All laterals		1827 Hz (192 Hz),		
		Following context		
		Back v.	Front v.	Cons.
Prev. con.	Back v.	1668 (154)	1749 (129)	1722 (116)
	Front v.	1827 (161)	1876 (150)	1861 (174)
	Cons.	1768 (210)	1899 (218)	1553 (304)

The linear mixed-effects model revealed that the two main effects of previous context (p=0.002) and following context (p=0.02) as well as the interaction between previous and following context (p=0.02) were statistically significant.

The post-hoc Tukey test showed that for the main effect of the previous context, the contrast between front and back vowels (p=0.005) as well as between front vowels and consonants (p=0.02) is statistically significant. In both cases, laterals after front vowels exhibit a higher F2 than after back vowels or consonants. In the main effect of the following context, only the contrast between front vowels and consonants following the lateral is statistically significant (p=0.03). Again, laterals followed by front vowels show a higher F2 than laterals followed by consonants.

As regards the interaction of the previous and following context, the post-hoc Tukey test revealed that only three pairwise contrasts were statistically significant, all of them involving the consonant – consonant – condition: This condition is statistically significantly different from the front vowel – consonant – condition (p=0.005), from the consonant – front vowel – condition (p=0.007) and from the front vowel – front vowel – condition (p=0.004). F2 for the lateral in the consonant – consonant – condition ranges from 1179 Hz to 1957 Hz with a mean of 1553 Hz. Hence, the mean of the consonant – consonant – condition is more than 300 Hz lower than the means of the other three conditions (compare Table 3).

This statistically significant interaction could be due to the fact that our data contained only six instances of a lateral between consonants, such as e.g. "Mittelschule" *secondary school*, where /ɛ/ is deleted between [t] and [l]. Despite the variation of F2 of laterals surrounded by consonants, the above-mentioned contrasts rendered statistically significant results.

Two of the six instances of laterals surrounded by consonants show F2 values below 1300 Hz, which is the upper limit of velarised laterals per definition by [2]. Except for those two laterals, no velarised laterals occur.

The contrast we were most interested in, namely the front vowel – front vowel – condition versus the back vowel – back vowel – condition (see *Hypothesis 1c*), did not reach significance in the post-hoc Tukey test (p=0.15). However, the difference between F2 in symmetrical front vs. back vowels is about 200 Hz with a lower F2 for the lateral between back vowels compared to the lateral between front vowels, as expected in *Hypothesis 1c*. Still, the mean F2 for laterals between back vowels is far from the measured F2 for a velarised lateral. The F2 mean for laterals between front vowels shows a tendency towards palatalisation with F2 exceeding 1900 Hz.

As regards the qualitative analysis, in numerous instances, the lateral gave the auditive impression of an alveolar lateral with a tendency towards palatalisation, though palatalisation was not present in all instances. We spotted palatalisation especially when the lateral was followed by a high vowel and in unstressed position, as e.g. in "weil" followed by "ich", *because I*.

However, in the qualitative analysis, we focussed mainly on instances of l-vocalisation, the characteristic feature of Middle Bavarian, and, as a comparison, on the production of retroflex laterals, typical for the southern part of Styria. In the investigated passage, only few instances were found: eight vocalisations of the lateral and two instances of retroflex laterals.

Some l-vocalisations were observed in the following words: "weil" *because* (4 times), "also" *thus* (3 times), "(ich) will" *(I) want* (2 times), and "Beispiel" *example*. We will try to give an interpretation of these four examples. First, "weil" and "also" are function words, occurring in unstressed positions, hence, in general, a vocalisation or rather absorption of the lateral is very likely. However, in our data, there are

numerous instances of function words in which the lateral is preserved. Second, both occurrences of "will" are found in stressed positions, and in both cases it is clear from the context that our speaker deliberately wants to use the dialect, thus consciously applying the process of vocalisation. Third, in the case of "Beispiel", the second syllable is unstressed and the vowel [ʏ] is already strongly nasalised due to the following word "meinem", *my*.

The two realisations of a retroflex lateral were found in the words "selber", *myself*, and "Mal", *times*, which can both be subject to retroflexion in Southern Bavarian.

5. Discussion

The speaker predominantly realised alveolar laterals in accordance with Standard Austrian German (*Hypothesis 1*). Compared with data from SAG, the mean F2 of all laterals in the data of our Albanian speaker is about 70 Hz higher (1827 Hz) than the mean F2 (1753 Hz) of alveolar laterals spoken by female speakers of SAG [preliminary results presented in 12]. This result indicates a tendency towards a palatalisation of the lateral.

This tendency towards palatalisation could be interpreted as a variant between the standard alveolar lateral and vocalisation, as our speaker is both oriented towards the standard language and the Middle Bavarian dialect, the latter holding a higher prestige than the Southern Middle Bavarian dialect. Thus, gearing towards the alveolar lateral and, at the same time, the l-vocalisation leads to an in-between form of palatalised laterals.

Compared to the formant data for Albanian, our results seem most similar to the data obtained by Beci [1]. However, we investigated a female speaker while he measured the formants of male speakers. Thus, for a detailed comparison, formant measurements of Albanian laterals spoken by female Albanian speakers would be beneficial. Additionally, to compare the obtained data with the different Austrian varieties, data of F2 in SAG as spoken in Graz as well as data of F2 from Styrian dialects should be analysed.

Except for two instances, no velarised laterals occurred; a result that corroborates our *Hypothesis 1a*.

Only two laterals were realised as retroflex laterals and only eight laterals were subject to vocalisation. This could be due to a residual influence of the Styrian dialect of Leoben, which features both dialectal variants (*Hypothesis 1b*). Both features are absent in stressed positions in Standard Austrian pronunciation, leading to social markedness, which explains the rare use of these two dialectal variants [21; 22]. As concerns our speaker's use of other dialect features as shown in Hobel et al. [23], e.g., input-switch rules – bidirectional rules switching between a dialectal and a standard variant [22, 34f.; 24, p. 348] –, she only uses dialectal features on a word level, but hardly ever on a phonological level. E.g., she realises dialectal [iː] or [nɛːd̥] as opposed to SAG [ɪç] or [nɪçt], "ich" and "nicht", *I* and *not*, respectively. However, she avoids phonological features such as e.g. dialectal [ɔ] – standard pronunciation [ɑ], as in [ɔbɐ] and [ɑbɐ] "aber" *but* [22, p. 50]. These findings explain the rare occurrences of vocalisations and retroflex laterals.

Our study revealed differences in F2 in dependence of the vowel context (*Hypothesis 1c*). Between front vowels, a tendency towards palatalisation was observed. Conversely, F2 between back vowels is lowered, but still within the range of alveolar laterals. Thus, laterals are not velarised in this context. However, this contrast did not reach significance.

To conclude, the Albanian speaker did not transfer both Albanian lateral phonemes to German, but rather produces an alveolar lateral. The velarised lateral, as found in Albanian, is abandoned in her speech production of German. The fact that F2 did not exactly correspond to F2 as obtained for SAG [12], corroborates the findings by [13] who stated that the L2 lateral production approaches the native speaker's production, but does not achieve it, not even at a very high level of proficiency; instead, an "intermediate" category is set up. This seems to apply to our data as well. However, formant values for Albanian female speaker would be necessary in order to compare these results to the results of our speaker's L2 German.

This case study can be seen as being part of a bigger survey of investigating what exactly constitutes a foreign accent when speakers speak German as a second language. Although their competence is near-native in grammatical aspects, the phonetic and probably phonological aspect lacks perfect command. Still, it is difficult to exactly trace the segmental and suprasegmental entities which differ from speakers of L1 German.

6. References

[1] B. Beci, *Të folmet veriperëndimore të Shqipërisë dhe sistemii fonetik i së folmes së Shkodrës*. Tirana: Akademia e Shkencave e RSh, 1995.

[2] S. Moosmüller, C. Schmid, and C. Kaseß, "Alveolar and Velarized Laterals in Albanian and in the Viennese Dialect," *Language and Speech*, in print.

[3] D. Müller, "Cue weighting in the perception of phonemic and allophonic laterals along the darkness continuum: Evidence from Greek and Albanian," *Albanohellenica*, vol. 6, pp. 1–14, 2015.

[4] E. Kranzmayer, *Historische Lautgeographie des gesamtbairischen Dialektraumes: Mit 27 Laut- und 4 Hilfskarten in besonderer Mappe*: Österreichische Akademie der Wissenschaften, 1956.

[5] P. Wiesinger, "Die Mundarten der Steiermark," in *Veröffentlichungen des Steiermärkischen Landesarchivs*, vol. 8, *Atlas zur Geschichte des steirischen Bauerntums*, F. Posch, M. Straka, and G. Pferschy, Eds., Graz, 1976, Blatt 12.

[6] R. Vollmann, B. Hobel, T. Seifter, and F. B. Pokorny, "The spread of /l/-vocalization in Styria," in *Phonetik in und über Österreich*, S. Moosmüller, C. Schmid, and M. Sellner, Eds., Wien: Österreichische Akademie der Wissenschaften, in print.

[7] C. J. Hutterer, "Der Stadtdialekt von Graz in Vergangenheit und Gegenwart," in *850 Jahre Graz 1128-1978. Festschrift im Auftrag der Stadt Graz*, W. Steinböck, Ed., Graz, Wien, Köln, 1978, pp. 323–354.

[8] M. Hornung and F. Roitinger, *Die österreichischen Mundarten: Eine Einführung*. Neu bearbeitet von Gerhard Zeillinger. Wien: öbv&hpt, 2000.

[9] W. Haas, "Vokalisierung in den deutschen Dialekten," in *Handbücher zur Sprach- und Kommunikationswissenschaft*, vol. 1, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung (Zweiter Halbband)*, W. Besch, U. Knoop, W. Putschke, and H. E. Wiegand, Eds., Berlin, New York: de Gruyter, 1983, pp. 1111–1116.

[10] P. Wiesinger, "Die Einteilung der deutschen Dialekte," in *Handbücher zur Sprach- und Kommunikationswissenschaft*, vol. 1, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung (Zweiter Halbband)*, W. Besch, U. Knoop, W. Putschke, and H. E. Wiegand, Eds., Berlin, New York: de Gruyter, 1983, pp. 807–900.
[11] S. Moosmüller, C. Schmid, and J. Brandstätter, "Standard Austrian German," *Journal of the International Phonetic Association*, vol. 45, no. 3, pp. 339–348, 2015.
[12] C. Schmid, "German initial laterals by bilingual L1 Bosnian migrants in Vienna," *12. Tagung Phonetik und Phonologie im deutschsprachigen Raum München*, Oct. 2016.
[13] M. Solon, "Do learners lighten up?: Phonetic and Allophonic Acquisition of Spanish /l/ by English-speaking Learners," *Studies in Second Language Acquisition*, pp. 1–32, 2016.
[14] A. Noll, J. White, P. Balazs, and W. Deutsch, *STX Intelligent Sound Processing: Programmer's Reference*, 2007.
[15] P. Carter and J. K. Local, "F2 variation in Newcastle and Leeds English liquid systems," *Journal of the International Phonetic Association*, vol. 37, pp. 183–199, 2007.
[16] D. Recasens, "A cross-language acoustic study of initial and final allophones of /l/," *Speech Communication*, vol. 54, pp. 368–383, 2012.
[17] R Core Team, *R: A language and environment for statistical computing*. Vienna, 2014.
[18] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[19] A. Kuznetsova, P. B. Brockhoff, and R. H. Bojesen Christensen, *lmerTest: Tests in Linear Mixed Effects Models: R package version 2.0-29*, 2015.
[20] L. Russel, *lsmeans: Least-Squares Means: R package version 2.20-23*, 2015.
[21] S. Moosmüller, "Soziophonologische Variation im gegenwärtigen Wiener Deutsch: Eine empirische Untersuchung," *Zeitschrift für Dialektologie und Linguistik*, vol. 56, p. 222, 1987.
[22] S. Moosmüller, *Hochsprache und Dialekt in Österreich: Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck*. Köln, Weimar: Böhlau, 1991.
[23] B. Hobel and R. Vollmann, "Phonological case study of the use of (Styrian) dialect and standard language in German as a second language," *Grazer Linguistische Studien*, vol. 84, pp. 5–20, 2015.
[24] W. U. Dressler and R. Wodak, "Sociophonological methods in the study of sociolinguistic variation in Viennese German," *Language in Society*, vol. 11, no. 3, pp. 339–370, 1982.

[i] In the following: alveolar lateral.

[ii] Bothorel measured only F2.

[iii] [3] pooled the results of Bothorel and Jubani-Bengu.

[iv] The data on F1 were not published in [2], but the measurements have been performed and are reported here.

The Acoustics of Fricative Contrasts in Two German Dialects

Stefanie Jannedy [1], *Melanie Weirich*[2]

[1] ZAS Berlin, Germany
[2] Friedrich-Schiller Universität Jena, Germany

jannedy@gmail.com, melanie.weirich@uni-jena.de

Abstract

In this study, we are investigating the acoustic characteristics of the five voiceless German fricatives [f s ç ʃ χ] elicited in non-words from 3 speakers each of two German dialects. The northern German dialect differentiates these five fricatives whereas in the middle German region and in Berlin, /ç/ and /ʃ/ have already merged or are in the process of merging [1;2]. Our previous work (submitted) has indicated that differentiating [ç] and [ʃ] acoustically in the speech of northern speakers from that of Berlin speakers works best when using DCT (discrete cosine transformation) rather than the four spectral moments. Results from our study corroborate this finding for both the northern and the middle German dialect.

Index Terms: German fricatives, spectral moments, DCT

1. Introduction

Fricatives are produced by obstructing the airstream at some place in the oral cavity, thereby creating turbulent noise. The place of the constriction affects the spectral characteristics of the noise: the more fronted a constriction is, the greater is the energy in the higher frequencies. A constriction further back in the oral cavity results in a spectrum with more energy in lower frequency regions. Thus, energy peaks of the postalveolar fricative /ʃ/ have lower frequencies than the energy peaks of the alveolar fricative /s/ [3]. Moreover, the acoustic discrimination of fricatives includes differences in intensity, spectral shape (e.g. skewness, peakedness) and also temporal parameters. Several studies have shown that fricative production is characterized by a rather high inter-speaker variability especially in sibilants cross-linguistically [4;5;6] mirroring both physiological and social sources of variation in fine phonetic detail (e.g. [7;8].

Numerous studies have investigated the various acoustic parameters potentially differentiating the fricatives of English (e.g. [3;9;10;11;12;13;14;15]). Much fewer studies have examined fricative realization in other languages such as Aleut, Scottish Gaelic or Chickasaw [6] or Polish [16;17;18]. Acoustic studies of the German fricative system are even more limited, [19] focused on the realization of German /h/ and /ç/ and compared it with the Japanese /hi/-syllable and [20] (this volume) are exploring the German fricative system.

The German fricative system is rather interesting because this system of contrasts involves the /ç/ sound which is relatively rare in the languages of the world. The distribution of /ç/ is restricted – it can only occur after high front vowels or word- or morpheme initially. Moreover, German is one of only three known languages of the world that contrasts the palatal fricative /ç/ and the postalveolar fricative /ʃ/ [21].

This contrast however has already dissolved in the middle German dialect region [22;23;24] and now the merger between /ç/ and /ʃ/ is also affecting the speech of speakers in the northeast of the middle German dialect belt up to Brandenburg and Berlin [1;2].

Previous work [17;2] has shown that the four spectral moments 1. center of gravity (COG), 2. Standard Deviation from the COG, 3. Skewness and 4. Kurtosis are less useful in differentiating /ç/ and /ʃ/ in German, and that Discrete Cosine Transformations (DCTs) [25] provide a useful method to differentiate different fricative categories in other languages with complex fricative systems such as Polish [16;17].

Thus, the purpose of the present study is to describe and evaluate the various acoustic parameters in terms of their usefulness to differentiate the contrasting five voiceless German fricatives [f s ç ʃ χ]. Fricative productions will be compared between speakers from the middle German dialect region where the contrast between /ç/ and /ʃ/ has mostly dissolved in spontaneous speech and speakers from the northwest, where the contrast between /ç/ and /ʃ/ is still fully realized.

2. Methods

For this study, we have recorded and analyzed the data of 3 female speakers from the northern part of Germany and 3 females from the middle German dialect region around Jena. All speakers were recorded with a head-mounted microphone at a recording frequency of 44 kHz, however, data was downsampled to 22kHz as fricative noise does not extend beyond 11 kHz. All speakers read the same list of 46 sentences containing a carrier phrase and 46 different target words. The reading of the list was repeated 4 times in the same order.

2.1. Recording Materials

The target words were either real German words or non-words, made up to keep the segmental context identical. For the purpose of this study, only the five non-words contrasting in the fricatives are considered for analysis. The block of these words were always read last.

1.	Ich habe	„iffa“	gesagt.	/f/
2.	Ich habe	„issa“	gesagt.	/s/
3.	Ich habe	„icha“	gesagt.	/ç/
4.	Ich habe	„ischa“	gesagt.	/ʃ/
5.	Ich habe	„acha“	gesagt.	/χ/

2.1.1. Segmentation

A text file and a sound file for each repetition was submitted to webMAUS [26] which generated a Praat TextGrid file based on a first pass of a speech recognizer, trying to align the word- and SAMPA canonical phonological transcriptions to

the audio file. The alignment of the boundaries was then hand corrected for the target words and fricatives.

2.1.2. Measurements

All acoustic measurements were done in PRAAT [27]. For the acoustic parameterization, the spectral moments (treating the spectrum as a probability density distribution) following [11] were calculated consisting of 1) the centroid frequency or Center of Gravity (COG), , 2) the Standard Deviation (SD) which is a measure of how much the frequencies in the spectrum deviate from the COG, 3) the skewness describing the energy distribution over the whole frequency range of the spectrum and expresses if the frequencies are skewed towards the higher or the lower frequencies; and 4) kurtosis which reveals the spectral peakedness of the distribution. In addition, *Discrete Cosine Transformation* (DCT) [28] was used to quantify the shape of the spectra. DCT decomposes the signal into a set of half-cycle cosine waves whereby the resulting amplitudes of these cosine waves are the DCT coefficients. We will concentrate on three DCT coefficients, which 1) are proportional to the linear slope of the spectrum (DCT1), 2) correspond to its curvature (DCT2), and 3) describe the amplitude of the higher frequencies (DCT3). Prior to analysis the data was filtered using a pass band filter (200-11025 Hz) and all acoustic measurements were automatically logged at the temporal midpoint of the fricatives.
For a better quantification of the fricative contrasts, Euclidean Distances (EDs) were calculated between all fricative pairs for each speaker separately using 1) the spectral moments and 2) the DCT1xDCT2xDCT3 space.

2.2. Speakers

For this analysis we used the data from 3 female speakers for each dialect group. The northern dialect speakers were born, raised and now live just to the south of Hamburg. They range in age between 39 and 44. The middle German speakers were born and raised in Thuringia, now living in Jena. They range in age between 23 and 34.

2.3. Statistics

All in all, 6 speakers x 4 repetitions x 5 fricatives = 120 items were analyzed. We performed a linear mixed effects analysis as implemented in the lme4 package [29] in R (version 2.14.1, R Development Core Team 2008). Likelihood ratio tests were run to test for a significant effect of the test variables by comparing the model with the factor in question to a model without that factor. Post hoc tests were carried out (using the lsmeans package in R [30]) to reveal any significant difference between the dialect areas for each fricative or fricative pair respectively.

3. Results

First, we will have a look on the average spectra of the five fricatives for both dialect areas. Then, the acoustic space of the fricatives is shown characterized by the different parameters (COG, SD, skewness, kurtosis and DCT1-3). The statistical analysis concentrates on investigating the different parameters in their usability to distinguish the German fricatives and to compare the realization of the fricative contrasts in the two dialect areas.

3.1. Fricative space

Figure 1 shows the average fricative spectra for the five fricatives for the speakers from Bux (top panel) and from Jena (lower panel). In the two dialect areas the five spectra vary in the energy distribution over the frequency range in a similar way. As expected /s/ (green) is characterized by high energy in the higher frequency regions over 6000 Hz, while /ç/ (red) and /ʃ/ (blue) have their energy peaks somewhat lower (for Bux around 3000 Hz, for Jena around 5000 Hz). For Jena, no clear peak but rather a plateau can be seen. Both /χ/ (magenta) and /f/ (orange) show flatter spectra, but this is especially the case for /f/ for the speakers from Bux.

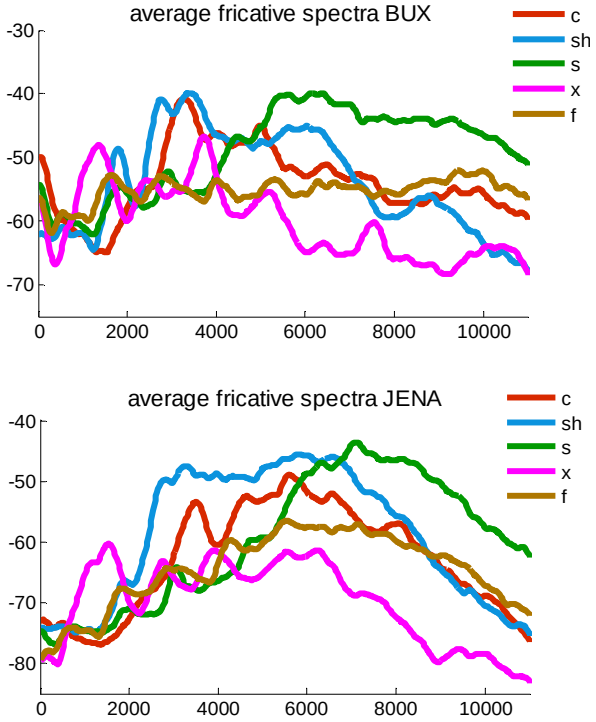


Figure 1: *Mean fricative spectra separated by dialect area, different fricatives are plotted in different colors*

Different measurements were made to see which of them are best qualified to describe and separate the spectral characteristics of the five German fricatives and which are maybe less suitable. The upper plot of Figure 2 shows the acoustic space of the fricatives spanned by the two spectral moments COG and skewness. While /s/ and /χ/ are nicely separated by COG (plotted on the x-axis), with /s/ (green) having high values and /χ/ (magenta) low values, the values for the three fricatives /ç ʃ f/ overlap. The same holds for skewness, which separates /s/ and /χ/ but fails to show dividable distributions for /ç/, /ʃ/ and /f/. The two dialect areas show similar distributions, however, overall, the values for skewness and COG are slightly shifted with higher values for COG and lower values for skewness for Jena than for Buxtehude.

The lower plot of Figure 2 shows the acoustic space of the fricatives spanned by DCT1 and DCT2: a clearer separation between the five fricatives can be seen for both dialect areas. However, /ç/ remains in the middle of the fricative cloud, thereby revealing the least clear separation to all other fricatives. Again, overall, the fricative distributions are somewhat shifted between the dialect areas.

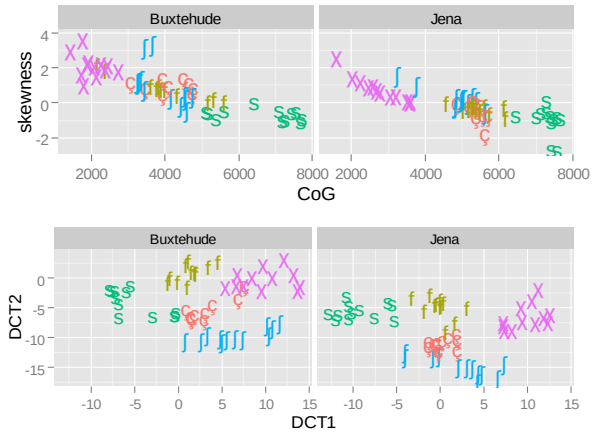


Figure 2: *Acoustic fricative space separated by dialect area using COG and skewness (upper plot) and DCT1 and DCT2 (lower plot).*

Different statistical models were run for the different parameters as dependent variable. As fixed effects, repetition was entered as a control variable, and dialect area (Buxtehude, Jena) and fricative (f, s, ç, ʃ, χ) as test variables. As random effect we entered an intercept for speaker.

Table 1: *Results of the Post Hoc Test showing significant differences between the fricatives in DCT1, DCT2 and DCT3 for the two dialect areas (*** p < 0.001, ** p < 0.01, * p < 0.05).*

	DCT1		**DCT2**		**DCT3**	
pair	Bux	Jena	Bux	Jena	Bux	Jena
/ç-f/	n.s.	n.s.	***	***	***	n.s.
/ç-s/	***	***	n.s.	***	***	***
/ç-ʃ/	*	n.s.	***	***	n.s.	n.s.
/ç-χ/	***	***	***	***	n.s.	n.s.
/f-s/	***	***	***	n.s.	***	***
/f-ʃ/	***	n.s.	***	***	***	*
/f-χ/	***	***	n.s.	n.s.	***	n.s.
/s-ʃ/	***	***	***	***	***	***
/s-χ/	***	***	***	n.s.	***	***
/ʃ-χ/	**	***	***	***	n.s.	n.s.

For all DCT coefficients we found a significant interaction of dialect area and fricative ($p < 0.001$). Results of Post Hoc Tests analyzing the differences between the fricatives for the two dialect areas are given in Table 1 (p values are adjusted using the tukey method). Overall, many of the comparisons were significant. While DCT3 shows fewer significant differences compared to DCT1 and DCT2, DCT3 clearly distinguishes /s/ from all other fricatives. Since DCT3 reflects the energy of the higher frequencies, it works best capturing the spectral characteristics of /s/. The fricative pairs most difficult to differentiate are /ç-f/, /ç-ʃ/ and /f- χ/, while the sibilants /s/ and /ʃ/ are distinguished best by the DCT parameters.

When comparing the two cities, it can be seen that Buxtehude shows more significant differences than Jena, thereby revealing a more distinct acoustic realization of the fricative contrasts.

For SD, skewness and kurtosis there was also a significant interaction of dialect area x fricative, for COG however, there was a main effect of fricative and of dialect area but no interaction thereof. The main effect of dialect area reflects a higher COG for Jena than for Buxtehude already apparent in the shifted distribution in Figure 2. In comparison to the DCT analysis, fewer fricative comparisons turned out to be significant. Moreover, for the /ç-ʃ/-contrast none of the parameters succeeded in distinguishing the spectral characteristics of the two fricatives. Parallel to the DCT analysis, more significant differences were found for Buxtehude than for Jena.

Table 2: *Results of the Post Hoc Test showing significant differences between the fricatives in COG, SD, skewness and kurtosis for the two dialect areas (*** p < 0.001, ** p < 0.01, * p < 0.05).*

	COG		**SD**		**skewness**		**kurtosis**	
pair	Bux	Jena	Bux	Jena	Bux	Jena	Bux	Jena
/ç-f/	n.s.	n.s.	***	***	n.s.	n.s.	n.s.	n.s.
/ç-s/	***	***	***	n.s.	***	n.s.	n.s.	n.s.
/ç-ʃ/	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
/ç-χ/	***	***	n.s.	**	***	***	*	n.s.
/f-s/	***	***	***	***	***	n.s.	n.s.	**
/f-ʃ/	n.s.	n.s.	***	***	n.s.	n.s.	*	n.s.
/f-χ/	***	***	***	***	***	***	***	n.s.
/s-ʃ/	***	***	***	n.s.	***	***	n.s.	**
/s-χ/	***	***	***	***	***	***	***	*
/ʃ-χ/	***	***	n.s.	n.s.	***	n.s.	**	n.s..

3.2. *ED of Fricative contrasts.*

Since the above analysis showed that the DCT coefficients are better descriptors distinguishing the acoustic characteristics of the five German fricatives we will concentrate on these parameters. Figure 3 shows the amount of the acoustic contrast between all fricative pairs expressed as the Euclidean Distance using DCT1-3 (ED_DCT123). The different dialect areas are plotted next to each other. A significant interaction of pair x dialect area was found ($\chi^2(9) = 26.213$, $p = 0.0019$). However, it only reveals that the contrasts between /s-ʃ/ ($p < .05$) and /s-χ/ ($p < .001$) were significantly higher in Jena than in Buxtehude. Overall, the order of the fricative pairs is much the same in both dialect areas: The smallest acoustic contrast is revealed for the /ç-ʃ/ pair; followed by the other /ç/-comparisons except the one with /s/.

4. Discussion

While generally, DCTs were better than the four spectral moments in separating the fricatives in each dialect, especially the curvature of the spectrum (DCT2) is a rather promising acoustic correlate for fricative differentiation, particularly when the very similar spectral shapes of /ç/ and /ʃ/ are concerned. Although the experiment was set up for speakers to visually capture the differences between the experimental tokens, and presumably all speakers made an effort to realize the acoustic contrasts reflected in the orthography, fewer fricative pairs turned out to be significantly different for the Jena speakers compared to the northern speakers. This is not surprising given the overall tendency of Thuringian speakers to merge at least /ç/ and /ʃ/ in fluent and spontaneous speech.

In light of this, it is interesting that the rare phoneme category /ç/ even evolved as a contrastive entity in a crowded fricative space as found in German. While our results capture the observation that the /ç-ʃ/ contrast is difficult to quantify and qualify even in the northern speakers, it is not a surprise that it is on the demise in the middle German Thuringian dialect. A listening experiment is planned to analyze if and how the contrast is perceived in the different areas. It remains to be seen if the /ç-ʃ/ merger also starts spreading further north. However, there may be some resistance by northern speakers which is not to sound like “southern” speakers.

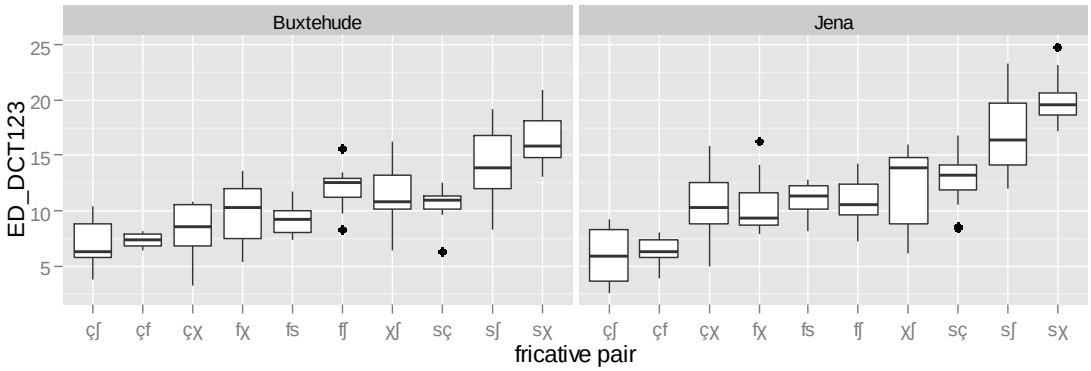


Figure 3: *Acoustic contrast between fricatives estimated as EDs in DCT1 x DCT2 x DCT3 space separated by dialect area.*

5. Acknowledgements

This work was funded by the German Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411). We would also like to thank Adrian Simpson (F.S. Univ. Jena) for allowing us to use his praat scripts for parameter extractions, Sophie Arndt for her patient work as a research assistant, and all of our informants.

6. References

[1] Jannedy, S. & Weirich, M. (2014). Perceptual divergence in an urban setting: category instability of the palatal fricative. Journal of Laboratory Phonology 5(1):91-122.

[2] Jannedy, S. & Weirich, M. (submitted) Spectral Differentiation of Intradialectal Fricative Variation. Submitted to Clopper, C. (ed). Special Issue of *JASA* on Methods in Dialect Research.

[3] Hughes, G. & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America* 28, pp. 303-310.

[4] Dart, S. N. (1998) Comparing French and English coronal consonant articulation. Journal of Phonetics 26, 71-94.

[5] Newman, R. S.; Clouse, S. A. & Burnham, J. L. (2001) The perceptual consequences of within-talker variability in fricative production. Journal of the Acoustical Society of America 109(3), 1181-1196.

[6] Gordon, M.; Barthmaier, P. & Sands, K. (2002) A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association* 32, 141-174.

[7] Weirich, M. & Fuchs, S. (2013) Palatal morphology can influence speaker-specific realizations of phonemic contrasts. *Journal of Speech, Lanuage and Hearing Research* 56, S1894–S1908.

[8] Stuart-Smith, J. (2007): Empirical evidence for gender speech production: /s/ in Glaswegian. In: Cole, J. & Hualde, J. I. (eds.): *Laboratory Phonology* 9. Berlin: Mouton. pp. 65-86.

[9] Harris, K. S. (1958): Cues for the discrimination of American English fricatices in spoken syllables. Language and Speech 1. pp. 1-7.

[10] Shadle, C. (1985) The acoustics of fricative consonants. *Technical Report* 506, Research Laboratory of Electronics, MIT Cambridge.

[11] Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R.N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America* 841, 115-123.

[12] Behrens, S. J. & Blumstein, S. E. (1988a). Acoustic characteristics of English voiceless fricatives: A descriptive analysis. *Journal of Phonetics* 18, 51-63.

[13] Behrens, S. J. & Blumstein, S. E. (1988b). On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *Journal of the Acoustical Society of America* 84, 861-867.

[14] Tomiak, G. R. (1990). *An acoustic and perceptual analysis of the spectral moments invariant with voiceless fricative obstruents.* Doctoral dissertation, SUNY Buffalo.

[15] Jongman, A.; Wayland, R. & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America* 108, 1252-1263.

[16] Bukmaier, V. & Harrington, J. (2016). The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change. *Journal of the International Phonetic Association*, pp. 1-19. doi:10.1017/S0025100316000062.

[17] Guzik, K. & Harrington, J. (2007). The quantification of place of articulation assimilation in electropalatographic data using the similarity index (SI). *Advances in Speech Language Pathology* 9 (1), 109-119.

[18] Wiktor, J. (1995). The acoustic parameters of Polish voiceless fricatives: Analysis of variance. *Phonetica* 52:252-258.

[19] Tronnier, M. & M. Dantsuji. 1993. An Acoustic Approach to Fricatives in Japanese and German. Proceedings of the 3rd EUROSPEECH'93, Berlin, vol. 1, 271-274.

[20] Lowery, M. & Kleber, F. An acoustic analysis of German fricatives. (this volume)

[21] Mielke, J. (2008). The emergence of distinctive features, pp. 1-304. Oxford: Oxford University Press.

[22] Herrgen, J. (1986). Koronalisierung und Hyperkorrektion. Das palatale Allophon des /CH/-Phonems und seine Variation im Westmitteldeutschen, pp. 1-278. Stuttgart: Franz Steiner.

[23] Dirim, I. & Auer, P. (2004). Türkisch sprechen nicht nur die Türken. Über die Unschärfebeziehung zwischen Sprache und Ethnie in Deutchland. Walter De Gruyter, pp 1-255.

[24] Hall, T. A. (2013). Alveolopalatalization in Central German as markedness reduction. Transactions of the Philological Society, pp. 143-166. doi: 10.1111/1467-968X.12002.

[25] Harrington, J. (2010). *Phonetic analysis of speech corpora*, pp. 1-424. Chichester: Wiley-Blackwell.

[26] Kisler, T. and Reichel U. D. and Schiel, F. and Draxler, Ch. and Jackl, B. and Pörner, N. (2016): BAS Speech Science Web Services - an Update of Current Developments, Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, paper id 668.

[27] Boersma, P. and D. Weenink (2012). Praat: doing phonetics by computer [Computer program]. Version 5.3.23, retrieved 7 August 2012 from http://www.praat.org/

[28] Watson, C. I. & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. Journal of the Acoustical Society of America 106, 458-468.

[29] Bates, D., Maechler, M. & B. Bolker (2011). lme4: Linear mixed-effects models using S4 classes, R package version 0.999375-42.

[30] Lenth, R. V. (2016) Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software*, 69(1), 1-33. doi:10.18637/jss.v069.i01

Fully Automated Accent Correction for Computer-Assisted Speech Rhythm Training: A Pilot Study

Markus Jochim[1], *Christoph Draxler*[1]

[1]Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich

{markusjochim|draxler}@phonetik.uni-muenchen.de

Abstract

In this project we have implemented a feedback method for CAPT (computer-assisted pronunciation training), where the learner's syllable durations are corrected in a fully automated way. The correction is based on the PSOLA algorithm [1] and uses a reference speaker's recording to determine the desired syllable durations for a given utterance. Learners get to hear a pre-defined sentence in their target language and are expected to imitate it as closely as possible. Their recording is then manipulated to match the reference speaker's syllable durations and the manipulation is presented as auditory, corrective feedback. Exploiting this feedback, learners are expected to try again and improve their pronunciation in terms of speech rhythm.

A pilot study with six learners of German with various L1 backgrounds was conducted to assess whether learners can successfully exploit the feedback to improve their pronunciation. The results suggest that learners can improve their pronunciation significantly using this method. They also suggest that the method works equally well or better than repeatedly listening to the reference speaker.

Index Terms: computer-assisted pronunciation training, CAPT, speech rhythm, software, web tool, computer-assisted language learning, CALL

1. Introduction

Language learners receive more and more assistance in their endeavor from smartphones and other devices. Computer-assisted language learning, or CALL, is a very interesting field in many respects. CALL applications can further our theoretical understanding of second language acquisition (SLA) and CALL users can benefit from advances in SLA. Designing CALL applications and materials happens at the "technology-pedagogy interface" [2] and is a challenge for software engineers and language teachers alike, and ideally they would tackle it jointly.

A complete CALL system has several important ingredients (exercises, user interface, performance rating, corrective feedback, to name just a few) and ideally covers training for speaking, listening, writing, and reading. In this paper we aim to contribute to the way pronunciation is taught in CALL systems, which is often termed CAPT (computer-assisted pronunciation training). More specifically, we will be talking about speech rhythm. We will argue that manipulating the learner's voice on-the-fly is a valuable method to provide corrective feedback.

CAPT will always involve the learner producing an utterance and the machine giving some sort of feedback. Most often this is done in a listen-and-repeat paradigm, where the learner hears a reference recording and tries to imitate it. Other options are reading tasks (often of individual words) and picture naming tasks. Completely spontaneous productions are rare and pose a special challenge for machines to evaluate.

We implemented a software that is capable of comparing two recordings – a native speaker's reference recording and a learner's attempt to imitate it – and automatically manipulating the syllable durations of the learner's recording to match the other one. We will argue that syllable duration can be used as a proxy variable for speech rhythm, which is a complex linguistic category associated with several phonetic correlates.

1.1. Why Use the Learner's Voice?

Owing to the fact that most systems are based on a listen-and-repeat paradigm, Probst et al. [3] started searching a "golden speaker", i.e. a reference speaker who is best for a certain learner to imitate. They found that learners benefit the most when imitating a reference speaker with a voice similar to their own. They compared the reference speaker's and the learner's pitch and speed, each talking in their respective native language (L1), to measure voice similarity. Arguably, the most similar voice possible is the learner's own voice. Indeed, Bissiri & Pfitzinger [4] found that imitating one's own voice yields better results than imitating a different voice.

A reason why using the learner's voice for auditory feedback works well is this: When hearing one voice with a correct pronunciation and another voice with an erroneous pronunciation (which is usually the case when a teacher corrects a learner's speech), it is hard to tell apart which of the acoustic deviations should be attributed to the voice difference and which account for the errors. It is therefore useful to hear erroneous and correct pronunciations in the same voice. One voice that learners will always hear with erroneous pronunciation is their own (until they have completely overcome their accent). To make the learners aware of the actual errors, it therefore makes sense to use that voice – their own – for the correct pronunciation as well. This of course requires technical manipulations, which are challenging but today seem feasible.

1.2. Syllable Duration as a Proxy for Speech Rhythm

Speech rhythm is a complex linguistic category that is influenced by a number of phonological properties, resulting in very different language-specific phonetic correlates.

Dauer [5] describes a set of (mostly phonological) features that define a language's speech rhythm. These include syllable structure (permissible syllable types as well as their frequency), stress, phonetic and phonological vowel reduction, phonetic and phonological vowel length.

All of these features vary across languages and lead to a language-specific phonetic pattern that is perceived as speech rhythm. Interestingly, all of the factors Dauer [5] assumes to define speech rhythm, also contribute to syllable duration. There-

fore, if learners transfer any of these phonological properties from one language to another, one major phonetic correlate of their accentedness must be syllable duration. This is why we chose syllable duration as a proxy for the linguistic category speech rhythm.

1.3. Software in Use

To assess the effectiveness of this and other feedback methods for CAPT, we have implemented a browser-based experiment tool called Captain. The tool and its source code are publicly available (see section 5). Existing tools such as SpeechRecorder [6] or PsychoPy [7] did not provide enough flexibility to implement our study design. There were two major reasons for developing a completely new tool, rather than using or extending one of the existing tools. First, we wanted to provide a framework for the evaluation of a range of CAPT feedback methods. It seems reasonable to have a tool at hand designed for the specific purpose. Second, we wanted the feedback methods implemented for scientific evaluation to be available for actual learning applications, preferably browser-based. This could be achieved quite easily by designing Captain in a modular way.

Captain is written in Typescript. The version used in the study was 0.1.1 and it was run in Chromium (version 51). It uses the WebAudio API [8] to make speech recordings, it is capable of running WebMAUS [9] and PSOLA [1] during the experiment and exploiting the results right away. Results are stored in emuDB format [10] on a server.

1.4. Hypotheses

In this study, we hypothesize that a learner's performance improves in terms of syllable duration, when the learner imitates pre-recorded native utterances and receives feedback according to the method we are proposing (H1). We also hypothesize that the performance gain is higher when receiving this kind of feedback than in a trivial learning paradigm where the learner listens to the same recording several times without receiving feedback (H2).

2. Method

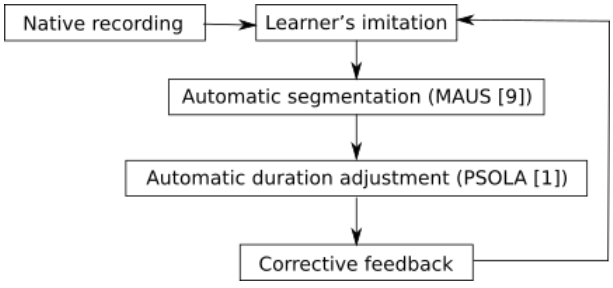


Figure 1: **Diagram of the manipulation method.**

We recorded six learners of German with various L1 backgrounds in a listen-and-repeat paradigm to test whether the feedback method we are proposing can successfully and intuitively be exploited by learners of a language (participant details are given in table 1). We chose 20 German recordings (one sentence each) by a male native speaker[1] and randomly distributed them among two experimental conditions: feedback condition (FC) and control condition (CC). The distribution was the same for all six speakers, but the order in which the 20 sentences were presented was randomised differently for every speaker.

[1]Speaker M65A from the PhonDat1 corpus, as provided by the Bavarian Archive for Speech Signals [11].

Table 1: **Details of the participants.** The proficiency level was subjectively assessed by the first author of this study.
G: Gender; AoL: Age of learning; AoA: Age of arrival; LoR: Length of residence in months; P: Proficiency

ID	L1	G	Birth	AoL	AoA	LoR	P
3	Azerbaijani	F	1987	18	28	9	B2
5	Polish	M	1991	23	23	18	B2
6	Turkish	M	1977	19	38	12	B1/B2
7	Italian	F	1994	15	21	5	B1
9	Russian	F	1955	12	-	-	B2/C1
10	Malagasy	F	1965	15	-	-	B2

In both conditions, each sentence was presented five times. The first trial was to present the native speaker's recording and ask participants to imitate it as closely as possible. The succeeding four trials were different across the two conditions. In control condition, all five trials were the same. This was to see how well speakers would improve with no feedback at all but with five chances.

In feedback condition, every recording of the learner was transformed on-the-fly (see figure 1): It was processed with WebMAUS [9], Wrassp's [12] implementation of the KSV pitch tracker and an own implementation of the PSOLA [1] algorithm. The result was a manipulated version of the learner's recording that matched the native speaker's syllable durations. This manipulated version was played back to the learner as the stimulus for the succeeding trial.

Each attempt of the participant was scored by means of the "mean deviation in syllable duration in milliseconds". To achieve this, the learner's token was compared with the reference recording syllable-wise. For each pair of syllables, the absolute value of the duration difference was calculated and the mean of those absolute values was taken. This is the performance indicator, with lower values indicating better performance. Performance gains are therefore indicated by falling scores between trials.

3. Results

Of the 120 test sentences this pilot study should have yielded (six participants with 20 sentences each; five repetitions per test sentence), only 84 could be included in the analysis. The most common reason for exclusion was the participant failing to properly imitate the sentence they had heard.

Figure 2 compares the participants' performance across the two experimental conditions. It shows a number of interesting points: (1) In both conditions, learners improve between the first and last trial. (2) The improvement is the largest over the first three trials. (3) The improvement is larger in feedback condition. (4) Performance in the first trial is worse in feedback condition (the median is only slightly higher, but variability is much larger). (1), (2), and (3) were expected, and (1) and (3), if confirmed statistically, would lend support to both our hypotheses. (4), however, was not expected at all. The two experimental conditions use the same kind of reference stimulus in the first trial. Systematic differences between the conditions should only

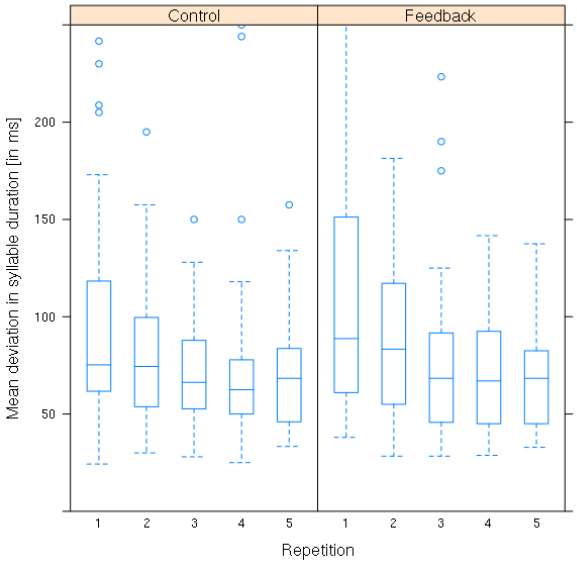


Figure 2: **Learning effect over the course of five repetitions**. For both control and feedback condition, we see how well learners performed in each of the five repetitions of every sentence. The plot has been limited at 250 ms for better scaling of the more common values. The boxes represent the first and third quartiles, the horizontal bars represent the respective median.

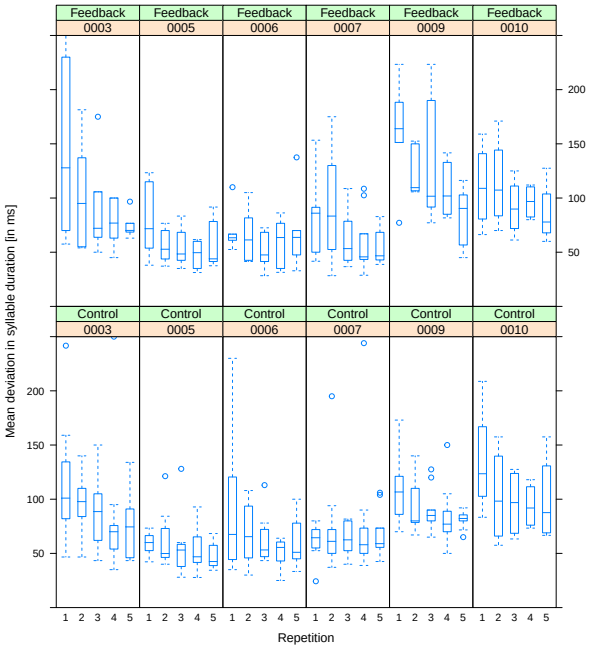


Figure 3: **Comparison of individual subjects.** Each subject's performance is compared across the two conditions and the five repetitions. Data are presented in the same way as in figure 2.

arise in trials 2–5.

Figure 3 shows the same comparison, but individually for all six participants. Table 2 gives a numeric model of that comparison: intercept and slope of a linear function fitted on each of the 12 graphs (logistic regression). The intercept models initial performance in trial 1 (or to be more exact, it models the imaginary trial 0). The slope models the performance gain between each trial. An ANOVA performed on this numeric model yielded a low significance ($p < 0.05, F = 8.94$) for condition, suggesting that the feedback condition did indeed yield better performance gains for learners than did the control condition.

Figure 4 shows pre-test results and reveals two more interesting effects. First, in showing the performance of the first author of this study (who is a native speaker of German), it establishes a baseline of what scores can be taken to be a "100 % achievement" (note, though, that the first author is from southern Germany while the reference speaker is from northern Germany). Second, it shows results from German L2 speakers who are at B2 level (non-native phonetician 1) and C1 level (non-native phonetician 2), respectively. However, being phoneticians and being very experienced with speech experiments seems to help them achieve native-like results. Neither of the three phoneticians has any slope between the first and fifth trial.

4. Discussion

In this paper, we have presented a first, small-scale evaluation of a feedback method for computer-assisted pronunciation training (CAPT). We compared six participants' performance under two conditions: First, when they heard an accent-corrected version of their own voice; second, when they heard the same native utterance several times.

Our results suggest that learners can successfully exploit both conditions, and that the difference between the conditions varies between participants. The feedback condition exhibits a much steeper slope (i. e. higher performance gain) for most participants. This seems to support our hypothesis H2, indicating that our feedback method is not only useful; it is also more useful than the control method, which is a trivial learning paradigm (listen to the same recording again and again).

However, the steeper slope does not seem to result in better scores in the fifth trial. It rather seems to result from worse scores in the first trials. This was completely unexpected since the two experimental conditions do not differ in the first trial. The reason why it occurred anyway is unknown at this point and might be a confound for the present results. It might have to do with the different sentences used in the two conditions (of the 20 sentences used, 10 were assigned to FC and 10 to CC; this assignment was the same for all participants).

We would like to point out two methodological shortcomings in this pilot study. First, we did not include a familiarization phase for the participants. We did show them screenshots of what the experiment would look like (it basically consisted of a white screen with two buttons), but they had no way to try it out themselves. Indeed, we had to exclude the first test sentences from analysis for some subjects, because they did not properly handle the interface.

The second issue lies within the scoring method (mean deviation in syllable duration). It does not normalize for speech rate, which means that talking faster or slower than the reference speaker leads to lower scores, even if the speech rhythm is perfectly acceptable. Score gains might therefore indicate adaptations in speech rate rather than actual rhythm improvements. This might be counter-acted by comparing the overall duration of reference and participant token and multiplying the score with their quotient.

Table 2: **Linear functions fitted to the learners' performance gain over the five trials.** Compare figure 3.

Subject	Condition	Slope	Intercept
0003	Feedback	-12.2	143
	Control	-7.1	112
0005	Feedback	-4.1	73
	Control	-2.4	65
0006	Feedback	-4.5	77
	Control	-3.9	80
0007	Feedback	-6.0	90
	Control	-3.9	80
0009	Feedback	-14.3	162
	Control	-7.1	112
0010	Feedback	-10.4	128
	Control	-8.6	126

Of course the automatic manipulation does not leave the signal without noise. The influence of this noise remains to be investigated. Evaluations of manipulation results have been carried out based on perception tests with human listeners [13] and based on machine learning [14].

We are confident that these preliminary results show that automatically reducing the learner's foreign accent and using their own voice as a kind of corrective feedback is feasible both from a technological and from a pedagogic point of view. It will be very interesting to see if it can be employed in learner-directed software and extended to cover other aspects of pronunciation than speech rhythm.

5. Online resources

The learner recordings as well as the automatically generated manipulations have been published with the Bavarian Archive for Speech Signals and can be retrieved at http://hdl.handle.net/11022/1009-0000-0001-3141-E.

An online demo of Captain including the feedback method described will be made available at http://www.phonetik.uni-muenchen.de/apps/captain. The source code of the software has been released under an open source license and is available from https://gitlab.lrz.de/groups/ips-lmu.

6. Bibliography

[1] F. Charpentier and E. Moulines, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones." in *Proceedings of Eurospeech 89*, vol. 2, pp. 13–19.

[2] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training. 15 (5), 441–467." vol. 15, no. 5, pp. 441–467.

[3] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors – In search of the golden speaker," vol. 37, pp. 161–173. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639301000097

[4] M. P. Bissiri and H. R. Pfitzinger, "Italian speakers learn lexical stress of German morphologically complex words," vol. 51, no. 10, pp. 933–947. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639309000363

[5] R. M. Dauer, "Stress-timing and syllable-timing reanalyzed," vol. 11, no. 1, pp. 51–62.

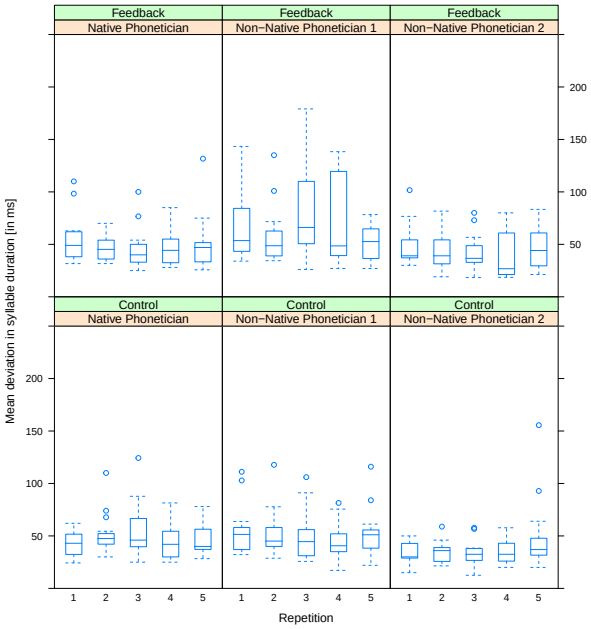


Figure 4: **Native speaker baseline and phoneticians' achievement.** Like in figure 3, we see individual comparison results, but this time from pre-tests. "Native phonetician" is the first author of this study, to establish a baseline how much even very good learners can be expected to deviate. The two non-native phoneticians' results suggest that the task is easier for phoneticians, independent from their L2 proficiency.

[6] C. Draxler and K. Jänsch, "SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software," in *Proc. of the IV. International Conference on Language Resources and Evaluation*, pp. 559–562.

[7] J. W. Peirce, "PsychoPy—Psychophysics software in Python," vol. 162, pp. 8–13. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165027006005772

[8] W3C, Audio Working Group, "Web Audio API," https://www.w3.org/TR/webaudio/. [Online]. Available: https://www.w3.org/TR/webaudio/

[9] T. Kisler, U. Reichel, F. Schiel, C. Draxler, B. Jackl, and N. Pörner, "Bas speech science web services - an update of current developments," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA), 2016.

[10] R. Winkelmann. The emuDB format. [Online]. Available: https://cran.r-project.org/web/packages/emuR/vignettes/emuDB.html

[11] Bavarian Archive for Speech Signals. Phondat1 speech corpus. [Online]. Available: http://hdl.handle.net/11022/1009-0000-0000-CAB2-3

[12] R. Winkelmann, L. Bombien, and M. Scheffers. wrassp. [Online]. Available: https://cran.r-project.org/web/packages/wrassp/

[13] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," vol. 51, no. 10, pp. 920–932. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639308001763

[14] M. Peabody and S. Seneff, "Towards Automatic Tone Correction in Non-native Mandarin," in *Chinese Spoken Language Processing*, ser. Lecture Notes in Computer Science, Q. Huo, B. Ma, E.-S. Chng, and H. Li, Eds. Springer, no. 4274, pp. 602–613.

Speech in Interaction – The Zurich Tangram Corpus

Yshai Kalmanovitch

University of Zurich, Switzerland
Yshai.kalmanovitch@uzh.ch

Abstract

The Zurich Tangram Corpus (ZTC) was design primarily to investigate the relationship between interactional experience and inter-speaker phonetic convergence. It is aimed to look primarily at phonetic convergence between interlocutors as a factor of interactional intensity and of mutual interactional experience over time. It contains recordings of dyad interaction in semi-natural discourse. Subjects were required to work together on an order reconstruction task in three sessions one week apart from each other. In addition to audio data, video data as well as eye-tracking data were collected.

Index Terms: phonetic convergence, interactional intensity, speech in interaction.

1. Introduction

Pickering and Garrod [1,2] describe imitation and/or convergence of communicational means in interpersonal interaction as an automatic process motivated by the interlocutors' need to facilitate communication by reducing variability in and increasing predictability of the immediate perceptual environment. According to their model, different perceived communicative behaviour patterns prime different sets of assumptions based on their association with past interactional experience of the interlocutors, while at the same time they inhibit other sets of assumptions [1]. The automatic nature of cognitive processing of perceived inputs leads to a potential conflict between the input perceived from other interlocutors' speech and the one resulting from the speaker's own speech, if those are too different from each other, as different inputs prime different sets of assumptions and at the same time mutually inhibit each other [2, 4]. Thus, in speech production interlocutors are motivated to converge to each other to reduce and avoid such conflicts [2].

Since the conflict predicted by [2] is expected to increase with greater interactional intensity (in terms of frequency of meaningful turn change in interaction), a logical assumption is that greater interactional intensity will lead to greater convergence between the interacting parties. It is also expected that mutual interactional experience will be carried to future interactions and learned speech patterns will be primed and activated when the same interactional conditions reoccur.

The corpus presented in this paper was designed to investigate these assumptions. It was initially designed to look primarily at the phonetic level, but is well suited to look at other linguistic levels, as well as at further phenomena of speech and communication in interaction.

2. Subjects

Subjects were recruited among students of different linguistic departments at the University of Zurich. Most of the students recruited are native speakers of Swiss German dialects and Swiss Standard German. They were then divided into dyads, matching speakers from different departments in each dyad in order to assure there was no prior acquaintance between the subjects in each dyad.

As of now, 40 subjects were recorded, and 32 of them are females. The vast majority of the subjects are between 19.5 to 24 years old (range: 18.5 to 30, median=22.25).

The subjects were divided into 20 dyads. With a very imbalanced gender distribution among the applying subjects, 13 dyads are all-female dyads, 1 dyad is an all-male dyad and 6 dyads are mixed gender dyads. Further recordings are still planned with hope to collect more data from male speakers.

The subjects were recorded normally once a week over three sessions on the same day and the same time. Each subject was paid 90 CHF upon the completion of all three recording sessions.

3. Task

The task used in this study is partly based on a task described by Clark and Wilkes-Gibbs [3]. In this task, one of the subjects (the instructor) has a table with 4X6 Tangram-figures (see fig.1), while the other subject (the receiver) is given an envelope containing cards carrying the corresponding figures. A line at the bottom of each card indicates the bottom of the figure. The receiver is expected to reconstruct the order of the figures as appearing on the table held by the instructor.

The subjects were verbally instructed to work together, so that the instructor had to describe the figures and the receiver could ask direct question or otherwise negotiate the instructor's explanation until the relevant figure was identified. Other strategies were not forbidden, with the only restriction being not to point directly at the figure in question. Figures could reappear in multiple tasks with the subjects in both roles.

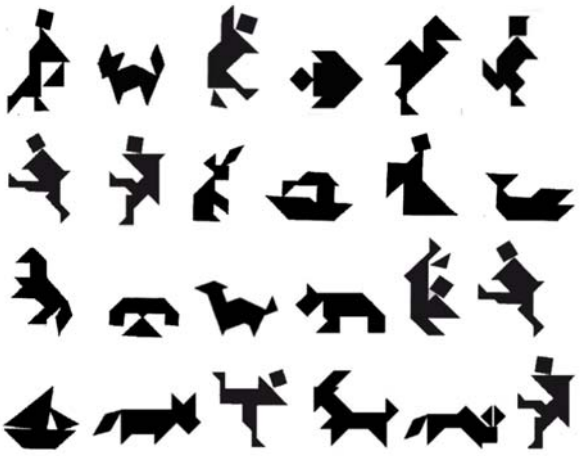


Figure 1: *A table with a matrix of 4X6 tangram figures*

In each session the subjects were required to work on four tasks, alternating roles between consecutive tasks. In order to maintain interest, the tasks varied slightly from one session to the other and were made more complex. Thus, in the first session the receiver was given cards with 24 figures corresponding exactly to those on the table of the instructor. In the second session one or two figures showing on the instructor's table were missing and some extra cards were added. In the third session the subjects had to add the remaining cards at the end of each task to those in the following task, thus increasing the number of extra cards in each task. At the same time the participants "collected" the cards to complete the cards missing in the first task at the end of the session.
The subjects were asked to use Standard German. No explicit indication as to the specific variety of standard German was given.

4. Collecting and processing the data

4.1. Recordings

All the recordings took place at the working space of the Phonetics Laboratory of the University of Zurich. Thus, the recording environment was quiet but still acoustically natural (the echo-attenuated booth at the lab was actually used by the recording technician who could watch the subjects from it during each task). The subjects were sat at around table facing each other (see figure 2.). The sitting order was kept from one recording session to the other so that the subject sitting on the left was always the first one to assume the role of the instructor at the first task. While the tables with the tangram figures were placed on the table during the whole recording session (2 tables for each subject), an envelope with the cards was provided by the recording technician prior to each task.

Figure 2: *Setting of the communicative task with the instructor sat on the left and the receiver on the right (in the upper photos they are wearing eye trackers).*

For the audio recording, a portable recorder (Marantz PMD661) and two external omni-directional lavalier microphones (Sennheiser MKE 2-PC) were used. The microphones were adjusted using a neckband approximately 2-3 cm from the subject's mouth. The two participants were recorded in stereo at a sampling rate of 44.1 KHz. Gain levels were readjusted manually by the recording technician at the beginning of each session.
Although the microphones did slightly change their position due to head movements of the subjects during the recordings, this setting still allowed a relatively clean, high quality recording in the natural surroundings. In addition, it allowed a rather neat segregation of the individual speakers' channels, such that the interference from one channel to the other is relatively small, as can be seen from fig. 3

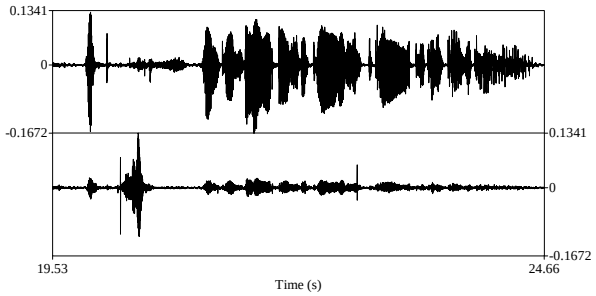


Figure 3: *Wave forms of the speech of two speakers recorded simultaneously in stereo*

In addition to the audio data, video data were collected for 17 of the dyads, using 2 digital video cameras with HD resolution (1080 x 1920 px), 25 fps (frames per second), converting recordings to MPEG-4 format.
Eye-tracker data were collected for 4 dyads using two binocular head-mounted eye-tracking devices (Dikablis Professional, Ergoneers) in full-HD field camera (1080 x 1920 px) at 60Hz eye-tracking frequency. More eye-tracking data are intended to be collected in further recordings planned, especially as technical problems were often encountered that could not be fixed in real time in order not to interrupt the natural course of the recorded interaction.

4.2. Storing and further processing of the recordings

The Audio data were stored as WAV files on a portable 2TB hard disk (Seagate Technologies) as well as on the servers of VideoLab and the Phonetics Laboratory of the University of Zurich.
At the current state the corpus entails a total of 60 (3*20) Audio recordings ranging from approximately 22 to 53 minutes (due to a technical problem, the audio recording in the second session of one of the dyads terminated after only 8 minutes, but an audio recording of this session is nevertheless available from the video data, though naturally of a lesser acoustic quality).
The mean duration of a recorded session in the data currently included in the corpus is about 33 minutes. This mean duration however refers to the raw data, including inter-task pauses due to the preparation for the next task. In the next stage therefore the recordings are cut to units representing single tasks in each session (a total of 12 tasks per dyad). This will allow analysis of accommodation processes within each session as well as between parallel stages in different sessions. In particular, we aim at discovering processes of long- and short-term cognitive learning and forgetting of speech patterns acquired within each session as well as from one session to the other (e.g. one could expect interlocutors to show more convergence at the first task of the second session compared to the first task of the first session, but not necessarily compared to the last task of the first session).

4.3. Transcription and annotation of interactional intensity

The recordings are to be supplied with an orthographic transcription using two text grid tiers (one for each subject in a dyad). The tiers will be segmented to mark the beginning and the end of each linguistic contribution by each interlocutor. This method of annotation will supply measures for the components defining interactional intensity.

Interactional intensity here refers to different features of the speech signal in the immediate perceptual environment that can increase or decrease the motivation of phonetic convergence. Such features are responsible for creating and strengthening predictions about the expected acoustic input both from other interlocutors as well as the result of the speaker's own speech production, and for invoking conflicts between those predictions and the actual speech signal in the immediate perceptual environment [2, 5]. Consequently, such features affect cognitive short-term learning and forgetting processes and allow for more accurate or less accurate sensorimotor adaptation.

In the analysis, we intend to use five features of interactional intensity.

- Turn taking frequency or inter-turn interval (ITI) is given by calculating the time interval between the beginning of a meaningful linguistic contribution by one interlocutor and the beginning of the next meaningful contribution by the other interlocutor (hence interjection, fillers etc. are excluded unless embedded in a meaningful contribution). The shorter the ITI is, the more frequent speakers can compare their predictions to the actual acoustic input and the more chances they have to calibrate and adapt their articulation [5]. Thus, the shorter the ITI is, the greater is the assumed interactional intensity.
- The actual duration of the linguistic contribution (TD) is given by calculating the time between the beginning of the turn and the actual end of the meaningful linguistic contribution. The longer the TD is, the more data in the speech signal are available in the immediate perceptual environment, allowing predictions about the expected linguistic input. Hence, the longer the TD is, the stronger are those predictions and the greater is the assumed interactional intensity.
- The information supplied by the TD and the ITI also allows access to information about latencies (TL) between turns and/or about parallel speaking (which can be regarded as negative latencies). These can be easily obtained by comparing the TD with the ITI (since ITIs are marked at the beginning of a turn, the TL always refer only to latencies documented at the end of a turn, i.e. between two consecutive turns). Longer latencies decrease predictions' strength of the preceding inputs by allowing short-term forgetting processes, while shorter latencies increase the conflicts between the expected and the actual acoustic input [6]. Thus, the shorter TL are, the greater is the assumed interactional intensity.
- Interactional bilaterality (IB) is the quantitative ratio between the contributions of both interlocutors in temporal units. Since the perceptual and motoric predictive models of the different speech patterns perceived from the acoustic input (from other interlocutor's speech as well as from own speech) mutually inhibit each other [2, 4], the lesser dominance one of those models has over others in the current interactional context, the greater are the conflicts invoked by processing them. Thus, the closer the IB-ratio is to 1, the greater is the assumed interactional intensity.
- The temporal variability of the actual durations of the meaningful contributions (TV) is simply the standard deviation of the actual duration of the linguistic contributions. This factor affects conflicts between different models in the same way as interactional bilaterality, but in the short-term. Thus, the smaller the TV is, the greater is the assumed interactional intensity.

The different components of interactional intensivity described above can be used either as components of a formula to give an interactional intensity index or can be fed independently into a regression model.

4.4. Metadata

At the end of the third session the participants were asked to fill in a questionnaire about their biographical background (place of birth, origin of the parents and linguistic knowledge and experience) and about their evaluation on a range from 1 to 5 of the tasks with respect to enjoyability (median=4) and difficulty of each task (median=2, 2 and 3 respectively).

4.5. Public use

After its completion the corpus is intended to be accessible to other researchers without any further conditions apart from the standard academic ethics.

5. Preliminary observations

5.1. Lexical alignment

A very preliminary and somewhat superficial observation of the data confirms that the results of the study of lexical alignment by Clarks and Wilkes-Gibbs [3] are likely to be replicated. Thus, although different names were used to indicate the same figures by different subjects (e.g. the fifth figure from the left in the second row in figure 1 was referred to as either an *angel*, a *nun* or a *woman with long sleeves*), subjects within the same dyad used always the same name, possibly after some negotiation. This is in particular advantageous, as it allows direct comparison of phonetic productions of identical tokens without manipulating or priming those tokens outside the actual context of the task, decreasing the chance of influence of the experimenter himself due to his social dominance in the context of the experiment.
In addition, the names used by the subjects to refer to reoccurring figures became shorter and definite with the time, as found also by Clarks and Wilkes-Gibbs [3].

5.2. Spatial alignment

Another interesting observation refers to the spatial conceptualization of the task itself. Thus, most of dyads chose a straightforward solution of ordering the cards in a 4X6 matrix, as appearing on the table of the instructor, going one row after the other. Yet, other dyads chose very different strategies which were used by both subjects within the same dyad, mostly without overtly negotiating it. In one dyad the receiver simply put the cards one on top of the other (in this

case the “first” receiver informed the instructor of his intention and the instructor imitated him in his turn without mentioning it again, despite the fact that he was not obliged to choose the same strategy). In other dyads both subjects ordered the cards in columns rather than rows (see the bottom pair of photos in figure 2) or in what seems to be quite a random matrix-pattern (e.g. 3X7+3) etc.

6. Research perspectives

The primary interest behind the ZTC remains the investigation of phonetic phenomena – primarily interpersonal phonetic convergence – in interaction. However, the corpus is well suitable for the investigation of further aspects of communication and convergence in interaction, assisted also by the video data which were not thoroughly discussed in this report. Indeed, the task often calls for the interlocutors to synchronize relative spatial concepts so that those could be understood by both interlocutors. Thus, the relative directional reference *left* could be understood as referring either to the speaker’s left, the listener’s left, the instructor’s figure’s left or the receiver’s figure’s left.

As mentioned, most of the participants recorded for the corpus are natives of Swiss German from different geographical and dialectal regions in Switzerland who in the experiment speak the Swiss variety of Standard German. Thus, the TCZ can also serve as a source for studying this particular language variety.

7. Acknowledgements

The corpus is composed as part of my dissertation under the supervision of Stephan Schmid, head of the Phonetics Laboratory at the University of Zurich, in cooperation with PD Dr. Klaus Wolfgang Kesselheim, Head of the VideoLab at the University Research Priority Project "Language and Space" at the University of Zurich.

The project was made available thanks to the funding of the Doctoral Program in Linguistics (DPL) of the University of Zurich.

8. References

[1] Pickering, M. J., Garrod, S. “Toward a mechanistic psychology of dialogue.” *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169-190, 2004.

[2] Pickering, M. J., Garrod, S. “An integrated theory of language production and comprehension.” *Behavioral and Brain Sciences*, vol. 36, no. 4, pp 329-347, 2013.

[3] Clark, H. H., Wilkes-Gibbs, D. "Referring as a collaborative process." *Cognition*, vo. 22, no. 1, pp. 1-39, 1986.

[4] Grossberg, S. “Resonant neural dynamics of speech perception.” *Journal of Phonetics*, vol. 31, no. 3, pp. 423-445, 2003.

[5] Houde, J. F., Nagarajan, S. S., Sekihara, K., Merzenich, M. M. “Modulation of the auditory cortex during speech: an MEG study.” *Journal of cognitive neuroscience*, vol. 14, no. 8, pp. 1125-1138, 2002.

[6] Goldinger, S. D. “Echoes of echoes? An episodic theory of lexical access.” *Psychological review*, vol. 105, no. 2, pp. 251-279, 1998.

PATSY-I: A Corpus on Non-Native English Air Traffic Communication

Caroline Kaufhold[1], *Christine Martindale*[1], *Axel Horndasch*[1]
Klaus Reinhard[2], *Elmar Nöth*[1]

[1]Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nuremberg, Germany
[2]e.sigma Technology GmbH, Ilmenau, Germany
{caroline.kaufhold, christine.f.martindale, axel.horndasch, elmar.noeth}@fau.de,
kreinhard@esigma-technology.com

Abstract

In many global tasks English is used as an international language. As a consequence, non-native speakers of the English language often communicate with other non-native speakers. An example is the Air Traffic Control (ATC) service which directs aircrafts on the ground and through controlled airspace. It is of course essential that there is a perfect understanding between the pilot and the ground-based controller to prevent collisions and to organize air traffic efficiently. Aviation English already accommodates non-native speakers of English by providing guidelines for wording and phraseology. To avoid confusion, for example, letters and numbers are spelled according to the international spelling alphabet provided by the International Civil Aviation Organization (ICAO). However, the ability to speak and understand English still has a high impact on communication success.

In this paper, we present a corpus that was recorded in the context of the ATC phraseology training system PATSY, the prototype of which was presented at the Show&Tell session at Interspeech 2015 [1]. The corpus consists of basic ATC utterances by speakers of 16 different mother tongues. Furthermore, "Please Call Stella" [2] and part of "The Rainbow Passage" [3] were recorded twice for every speaker with different biases. We plan on using this data to study the entrainment effect, which was observed for conversations by Levitan and Hirschberg [4]. Preliminary results on basic ATC utterances show a moderate correlation between the speakers' self-assessment and the GoP (Goodness of Pronunciation) score.
Index Terms: English as Lingua Franca (ELF), Computer Assisted Pronunciation Training (CAPT), Accent Entrainment

1. Introduction

A multitude of circumstances arise in daily life where people with different mother tongues must communicate with each other. However, depending on the context, the consequences of these interactions vary greatly regarding severity [5]. This project investigates the use of English as an international language, or as the Lingua Franca (ELF), in the context of air traffic control, in which flawless communication is of course of utmost importance.

Communication between speakers where at least one party's mother tongue is not English can be classified into two distinct categories: firstly, one of the interlocutors is a native speaker (NS) and the other is a non-native speaker (NNS) and secondly, both interlocutors are non-native speakers. While miscommunication is likely if only one of the interlocutors is a NNS, communication success relies even more on the intelligibility of the parties' spoken English if both are non-native speakers.

The paper is structured as follows: section 2 gives a general introduction to the work-in-progress phraseology training system PATSY [1] which was initially presented at the Show&Tell session at Interspeech 2015. The recording setup, structure of the recordings and the collected data is then discussed in section 3. The part of the corpus which contains basic ATC wordings is described in more detail in section 3.2. The other part, which will be used for investigating the effect of accent entrainment is described in section 3.3. Our experimental setup and preliminary results for automatically assessing pronunciation proficiency on the PATSY-I corpus using the Goodness of Pronunciation (GoP) score [6] are discussed in section 4. Section 5 summarizes the PATSY-I corpus and concludes the paper.

2. PATSY Project

PATSY is an abbreviation for the German name of the project "Piloten/ATC Trainingssystem für den Sprechfunk" which translates to "Pilot/Air Traffic Controller (ATCO) Training System for radio communication". Air Traffic Control (ATC) is the service provided by the ground-based controller or ATCO, who navigates the aircraft on the ground and in the controlled airspace. Pilots and ATCO maintain radio contact and communicate in the English language. In order to compensate for difficulties in understanding due to the language barrier, the International Civil Aviation Organization (ICAO) introduced special spelling rules for letters and digits. Phraseology and the special pronunciation rules are taught in flight school and it is planned that PATSY will become part of the training for new pilots. On the one hand PATSY shall help the trainees to internalize the vocabulary and syntax used in ATC communication. On the other hand, it will give direct feedback regarding the user's pronunciation and intelligibility.

In our first prototype, which was shown on Interspeech 2015 [1], we use the Goodness of Pronunciation (GoP) score [6] to assess pronunciation skills. During the assessment, the user can listen to the play-back of a reference speaker after each turn. By re-recording the utterance he/she can improve his/her pronunciation. PATSY will then compute the new pronunciation score and update the evaluation presented to the user. Over the last year, we have continued to collect recordings from as many different mother tongues as possible. The "PATSY-I" corpus, which contains the data collected thus far, is discussed in the following section.

3. The PATSY-I Corpus

The PATSY-I corpus is the first set of recordings collected in the course of the PATSY project. The corpus recordings can be split into two parts: recordings of type "Flight Basics" and recordings of type "Read Paragraphs". Basic ATC recordings comprise ICAO spelling alphabet words and numbers with and without special pronunciation rules. They will be referred to as "Flight Basics". The other part of the corpus consists of recordings of the two paragraphs: "Please Call Stella" [2] and a part of "The Rainbow Passage". These recordings will be referred to as "Read Paragraphs" [3].

Our goal was to gather recordings from speakers with many different mother tongues with a reasonably high level of proficiency in English. Therefore, many of the participants were PhD students working at the Pattern Recognition Laboratory of the Friedrich-Alexander University Erlangen-Nuremberg. Recordings of 70 speakers were collected of whom 47 were working in research and 23 were students or graduates at the time of the recordings. As shown in in Table 1, there are 34 German and 17 Chinese speakers among the 70 participants. The remaining 19 have a wide range of mother tongues represented by three or less speakers. 54 of the 70 speakers are male and 16 are female. The recording setup and a description of the reading tasks is given in the following section.

$\sum$	**de**	**cn**	**ar-sy**	**es**	**fa**	**tr**	**en**	**en-zw**
62	34	17	3	3	2	1	1	1
$\sum$	**es-ar**	**ml**	**hi**	**hr**	**it**	**id**	**pt**	**ru**
8	1	1	1	1	1	1	1	1
70								

Table 1: Distribution of mother tongues (L1). German (*de*), Chinese (*cn*), Arabic (Syria) (*ar-sy*), Spanish (*es*), Persian (Farsi) (*fa*), Turkish (*tr*), English (*en*), English (Zimbabwe) (*en-zw*), Spanish (Argentina) (*es-ar*), Malayalam (*ml*), Hindi (*hi*), Croatian (*hr*), Italian (*it*), Indonesian (*id*), Portuguese (*pt*), Russian (*ru*)

3.1. Recording Conditions

The SpeechRecorder software [7] offered by the Clarin-D website was used to collect the PATSY-I corpus. 59 speakers were recorded in a non-noisy environment using a Plantronics headset microphone. The remaining 11 recordings were done mostly in China at the participant's home without additional guidance. To some extent these recordings are affected by background noise and/or bad recording equipment.

Before the recording session, speakers were asked to rate statements about their personality taken from the Big Five Inventory-10 (BFI-10) by Rammsted and John [8] which is a short form of the Big Five Inventory-44 (BFI-44) by John and Srivastava [9]. The BFI-44 describes human personality in terms of 5 factors: "Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience". It consists of 44 statements which are rated on a 5-point Likert scale anchored at 1 = disagree strongly to 5 = agree strongly. To save time, Rammsted and John reduced the number of statements from 44 items in total to 2 items per factor, such that our participants had to rate 10 personality statements. All participants who were recorded for the Patsy-I corpus rated the English version of the statements (both BFI-10 and BFI-44 have also been published in German).

Furthermore, speakers were presented with a form to collect demographic data. The information they were asked to provide was: name, age, gender, place of birth and residence, native language, other languages spoken, English learning method, age of English onset, year of last English lesson, duration of residence in an English speaking country and own English proficiency judgement. Also date, location and recording hardware were collected for each session.

3.2. "Flight Basics" Recordings

To cope with the demands of a phraseology training system for future pilots and ATCOs, we focused on two basic topics of their flight education: call signs and radio (channel) frequencies. Call signs are used for the identification of an aircraft and represent a unique name which consists of a combination of the airline's identification number and the flight number. The ICAO spelling alphabet assigns a unique word to each letter of the alphabet in order to reduce misunderstandings due to poor audio conditions, radio interferences or differing pronunciations. For example, "*a*" is spoken as "*alpha*" and "*z*" is spoken as "*zulu*". For the same reasons, there is also an ICAO pronunciation given for numbers. The "*th*" sound is avoided, such that "*three*" and "*thousand*" become "*tree*" and "*tausand*", "*four*" and "*nine*" are pronounced as "*fower*" and "*niner*". Radio frequencies are six digits long and the period after the first three digits is pronounced as "*decimal*"; for example, 129.775 is spelled as "one two niner decimal seven seven five".

The recording session was structured as follows: first, the participant was asked to read nine sequences of three ICAO spelling alphabet words (e.g. "*bravo romeo mike*"). Then all speakers were shown nine sequences of digits which were between two to six digits long and written as words. After presenting the ICAO pronunciation rules for numbers "*three*", "*four*", "*nine*" and "*thousand*", eleven prompts showing written sequences of the same length as before were displayed. Numbers which should be pronounced in a special way were shown according to the ICAO pronunciation rules (e.g. "tree" instead of "three"). Finally, the speaker was asked to read ten call signs and five frequencies aloud. Digits were written as words and the period was written as "*decimal*".

For each recording, the study participants were shown a prompt which they read aloud. After a fixed number of seconds, the next prompt was shown such that recordings of all speakers were of the same length. Every speaker recorded 44 basic ATC utterances the distribution of which is shown in Table 2. For this part of the corpus, every speaker contributed 4.02 minutes of recorded speech.

	Vocabulary	**Recordings**
Flight Basics	ICAO Spelling Alphabet	9
	English Numbers	9
	ICAO Numbers	11
	Call Signs	10
	Radio Frequencies	5
Number of recordings per speaker in total:		44

Table 2: Number of recordings of type "Flight Basics" per speech task, per speaker.

3.3. "Read Paragraphs" Recordings

In PATSY, the user is asked to answer specific ATC questions in English. PATSY then checks if the input is correct (the right words were uttered) and computes a pronunciation score. If that score is below a certain threshold, the system assumes that the intelligibility of the utterance is not sufficient and asks the user to record it again. To support an improvement of the pronunciation score, the user's utterance as well as the words' correct pronunciation spoken by a reference speaker are played back so he/she can perceive the difference.

When designing the PATSY system, the authors were confronted with the question of whether both the pronunciation and the accent of the user will become more similar to the reference speaker. Levitan and Hirschberg observed this effect which they call "entrainment" for conversational speech. They showed that conversational partners become more similar, for example, in terms of turn-taking behavior [10] or speech phenomena triggering backchannels [11] which listeners use in order to signal continued interest and understanding [12]. For PATSY, users do not take part in a conversation, however, we collected recordings in order to analyse the potential effect of "accent entrainment".

The recording session was structured as follows: first, all 70 speakers were asked to read a part of "The Rainbow Passage" (RR) followed by the paragraph "Please Call Stella" (RS). Then, the participants were shown the prompts for doing the "Flight Basics" recordings described above (see Table 2). Next, the speakers listened to a recording of a reference speaker reading the same part of "The Rainbow Passage" he/she had read before and then they were asked to read it again (L-RR). Finally, for every sentence of "Please Call Stella" the speaker listened to a recording of a reference speaker after which he/she was asked to read it again (L-RS).

Read Paragraphs	**Reading without listening:**		
	"The Rainbow Passage"	(RR)	1
	"Please Call Stella"	(RS)	1
	Reading after listening:		
	"The Rainbow Passage"	(L-RR)	1
	"Please Call Stella"	(L-RS)	8
Number of recordings per speaker in total:			11

Table 3: Number of recordings of type "Read Paragraph" per speech task, per speaker. The ID in brackets behind the name of the paragraph indicates which paragraph was read: "R" for the part of the "The Rainbow Passage" and "S" for "Please Call Stella". The prefix states whether the user read the paragraph without listening to a reference speaker ("R") or he/she listened to a recording before reading the paragraph ("L-R"). In the "L-R" case for "Please Call Stella" each sentence was recorded separately.

The participants listened either to reference speakers with American English (AE) or British English (BE) accent. The AE and BE versions of "Please Call Stella" were taken from the Speech Accent Archive [13]. The AE reference recording and the BE reference recording for the part of "The Rainbow Passage" were taken from the IDEA corpus [14]. All participants of the PATSY recording sessions were either part of the AE or the BE group. As a consequence they were either confronted with reference speakers who had an American (36 participants) or a British accent (34 participants). Eleven recordings were done by every speaker. Table 3 shows an overview with respect to the number of recordings per speaker and reading task. The mean, minimum and maximum recording duration in seconds for every "Read Paragraphs" recording is shown in Table 4. It is interesting to see that on average speakers needed less time when they read "The Rainbow Passage" the second time. The short time span needed for recordings of type "L-RS" is due to the fact that all 8 sentences of "Please Call Stella" were read and recorded separately.

	Duration		
ID	mean	min	max
RR	43.20	31.70	64.15
L-RR	32.86	22.48	67.67
RS	41.28	4.23	88.03
L-RS	6.26	3.40	30.22

Table 4: Statistics regarding the "Read Paragraph" recordings: mean, minimum and maximum duration in seconds.

4. Experiments

In this section we present preliminary results of our experiments on the "ICAO Spelling Alphabet" recordings of the PATSY-I corpus. The goal was to investigate the suitability of the GoP score for automatic pronunciation scoring.

The data used for the experiments comprises 630 recordings (9 per speaker). The computation of the GoP score according to Witt [6] was done using a Julius speech recogniser [15]. The recogniser was trained on recordings of American English speakers and was provided by our cooperation partner e.sigma. We used the recogniser for word and phoneme recognition. The vocabulary of the word recogniser includes all 26 words of the ICAO spelling alphabet and additional pronunciation variants for 18 of these words which results in 54 entries in the pronunciation lexicon in total. For the vocabulary of the phoneme recogniser the 44 phonemes of the English language were used.

The GoP score is a well-known approach for identifying mispronounced phonemes. This similarity measure describes the distance between reference phonemes, based on a recognizer trained with native speech, and the actual phonemes of a user utterance [16], which are in our case the PATSY-I recordings. The score is "0" if there is no difference between the reference phoneme "known" to the recogniser and the phoneme produced by the speaker. A GoP score greater than zero denotes a difference between the actual realisation of the phoneme and its reference.

The mean GoP score was computed for each speaker of the PATSY-I corpus. The statistics on the resulting 70 GoP scores are as follows: a minimum GoP score of 0.41 and a maximum GoP score of 3.35 were observed. The median GoP score was 0.76 and the mean of all GoP scores was 0.87. The great difference between the maximum GoP score and the mean GoP score is due to the high GoP scores of some speakers who were recorded with background noise and/or bad recording equipment as already mentioned in section 3.1. However, within the speakers recorded under lab conditions, the highest GoP score is 1.26.

These results show that the recording conditions have an effect on the GoP score and consequently on the speakers' per-

ceived pronunciation. While this outcome is comprehensible – due to the background noise the accuracy of the phoneme recognizer decreases and the GoP score is worse – it also indicates that differences in recording conditions should be accounted for when comparing GoP scores. However, since these are only preliminary results, this outcome has to be analysed in more detail.

To examine the usability of the GoP score for pronunciation proficiency, we related the computed GoP scores to the self-assessment of English proficiency given by each speaker (see section 3.1). Each speaker therefore rated his/her own English proficiency on a scale of 1 to 6, where "1" represented "Very Good" while "6" was "Very Bad". Values between two integers were also valid. The complete distribution is shown in Table 5. Comparing the speaker proficiency grades of the self-assessment with the computed GoP scores, we obtained a moderate Pearson correlation of $r = 0.49$ (p-value: 0.00002) and a weak Spearman correlation of $\rho = 0.32$ (p-value: 0.00778). These preliminary results show that there is a correlation between the speakers' own perception of their English proficiency and the computed GoP score. However, in order to make more reliable statements about the speakers' intelligibility, we plan on using the word error rate (WER) of a speech recognizer. Another approach on our roadmap is to examine the effect of additive noise on speaker intelligibility.

Self-assessment	1	1.5	2	2.5	3	4	5	6
#speakers	8	2	34	6	15	4	1	0

Table 5: Distribution of English proficiency ratings given by all speakers of the PATSY-I corpus.

5. Summary

The PATSY-I corpus contains speech recordings of 70 speakers of 16 different nationalities. Every speaker made 55 recordings 44 of which contain standard ICAO wordings and 11 recordings are either one or more sentences of the paragraph "Please Call Stella" or part of "The Rainbow Passage". The focus of the PATSY-I corpus is on the pronunciation skills of non-native English speakers based on Air Traffic Control flight communication wording. A key aspect of our approach is to look at the difference between pronouncing English "as the speaker is used to it" and pronouncing English as recommended by the ICAO pronunciation rules, which were designed to increase a speaker's intelligibility. Another goal when recording the PATSY-I corpus was to get data regarding the influence of a reference speaker's accent on the participant. To achieve this goal we made one group of participants listen to a reference speaker with an American English accent and the other group had to listen to a reference speaker with a British English accent before reading the text themselves. In summary, we have on average 4 minutes of ATC flight communication wording and 2.8 minutes of read paragraphs per speaker. We presented a database spanning more than 8 hours of data from 70 different speakers. Our experiments regarding the use of the GoP score for automatically assessing the pronunciation proficiency already show promising results which we will further analyse in the near future. We are also discussing internally the possibility to provide the PATSY-I corpus to the BAS Repository.

6. Acknowledgements

This project was funded by the Federal Ministry for Economic Affairs and Energy's ZIM (Central SME Innovation) program. Some of the recordings used in this project are used by special permission of the International Dialects of English Archive, online at http://wwww.dialectsarchive.com.

7. Bibliography

[1] C. Kaufhold, V. Gamidov, A. Kiessling, K. Reinhard, and E. Nöth, "PATSY — It's All About Pronunciation!" in *in Proc. of Interspeech 2015*, Dresden, Germany, 2015, pp. 1068–1069.

[2] Weinberger, Steven H and Kunath, Stephen A., "The speech accent archive: towards a typology of english accents," in *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Brill, 2011, pp. 265–281.

[3] G. Fairbanks, *Experimental phonetics: selected articles.* University of Illinois Press, 1966.

[4] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." in *in Proc. of Interspeech 2011*, Florence, Italy, 2011, pp. 3081–3084.

[5] J. M. Levis, "Guidelines for promoting intelligibility." in *in Proc. of International TESOL Conference*, Seattle, WA, 2007. [Online]. Available: http://jlevis.public.iastate.edu/intelligibility.ppt, visited 2016-09-15.

[6] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning." *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[7] C. Draxler and K. Jänsch, "SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software." in *in Proc. of LREC 2004*, Lisbon, Portugal, 2004.

[8] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German." *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[9] O. P. John and S. Srivastava, "The Big Five trait taxonomy: History, measurement, and theoretical perspectives." *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[10] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue." in *in Proc. of AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, Palo Alto, California, 2015.

[11] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels." in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 113–117.

[12] E. A. Schegloff, "Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences." *Analyzing discourse: Text and talk*, vol. 71, p. 93, 1982.

[13] S. Weinberger. (2015) Speech Accent Archive. George Mason University. [Online]. Available: http://accent.gmu.edu, visited 2016-09-15.

[14] P. Meier, "International dialects of English archive." 1997. [Online]. Available: http://www.dialectsarchive.com/, visited 2016-09-15.

[15] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," 2001.

[16] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation," in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 21–30.

Der optionale Komplementierer im Deutschen – ein Fall prosodischer Syntax

Gerrit Kentner[1], *Isabelle Franz*[1], *Christian Dück*[1]

[1]Goethe-Universität Frankfurt

kentner@lingua.uni-frankfurt.de, i.franz@em.uni-frankfurt.de, chrisd14@gmx.net

Abstract

Wir untersuchen das Vorkommen von eingebetteten Verbzweit und Verbend-Sätzen (letztere mit dem Komplementierer "dass") in Abhängigkeit der rhythmisch-prosodischen Struktur der Satzspitze des eingebetteten Satzes. Die Korpusdaten legen einen deutlichen Einfluss des linguistischen Rhythmus auf die Syntax nahe. Bemerkenswert ist dabei, dass der Rhythmus nicht nur die Wortabfolge beeinflusst, sondern bereits die Wahl der syntaktischen Struktur (Haupt- vs. Nebensatz) bedingt. Für Modelle der Sprachproduktion bedeutet dieses Ergebnis, dass die Prosodie nicht nur die syntaktische Struktur reflektiert, sondern die Wahl der syntaktischen Struktur mitbestimmt.

Schlüsselbegriffe: Prosodische Syntax, optionaler Komplementierer, Rhythmus, Sprachproduktion

1. Einleitung

Theorien der Sprachproduktion (Levelt 1989 [1], Dell et al. 1997 [2]) gehen einvernehmlich davon aus, dass bei der Formulierung von Sätzen die Erzeugung der Satzstruktur ("grammatische Enkodierung") vor der phonologischen Verarbeitung beginnt; entsprechend sind syntaktische Einflüsse auf die phonologische Struktur von Äußerungen erwartbar und gut dokumentiert (Speer et al. 2011 [3], Wagner 2005 [4], u.v.a.). Strittig dagegen ist, inwieweit phonologische Faktoren ihrerseits Einfluss auf die Planung des Satzbaus nehmen können. Tatsächlich gibt es deutliche Hinweise auf ein gewisses Maß an wechselseitiger Beeinflussung von Phonologie und Syntax: Sprecher bilden, wenn sie die Wahl haben, eher prosodisch besonders wohlgeformte Sätze als denkbare syntaktische Alternativen, deren lautliche Struktur suboptimal ist (Behaghel 1930 [5], Schlüter 2005 [6] und andere). Allerdings ist der Rahmen, in dem die Phonologie Einfluss auf die Satzstruktur nimmt, weder abgesteckt noch hinreichend modelliert worden. Es ist weitgehend unklar, welche prosodischen Wohlgeformtheitsbedinungen auf welchen Stufen der grammatischen Enkodierung wirksam sein können. Die Problematik der psycholinguistischen Modellierung wird besonders daran deutlich, dass prominente Grammatikmodelle generativer Prägung postulieren, dass das zur Erzeugung der Satzstruktur nötige syntaktische Wissen von phonologischen Bedingungen weitgehend unberührt bleibt – die Phonologie gilt demnach als rein interpretierende Komponente, die nicht unabhängig satzbaurelevant sein kann (Zwicky & Pullum 1986 [7], Scheer 2010 [8]). Vor dem Hintergrund der unproblematischen Annahme, dass bei der Sprachproduktion auf grammatisches Wissen zurückgegriffen wird, stellt sich die Frage, wie die mentale Grammatik – hier insbesondere die Schnittstelle von Syntax und Phonologie betreffend - beschaffen sein muss, um phonologische Einflüsse auf den Satzbau in der Sprachproduktion erklären zu können (für einen rezenten Überblick vgl. Anttila 2016 [9]).

2. Korpusstudie

Hier untersuchen wir den Einfluss des phonologischen Faktors "linguistischer Rhythmus", der sich aus der Abfolge lexikalisch betonter und unbetonter Silben ergibt, auf die Struktur eingebetteter Sätze. Im Deutschen ist, ähnlich wie im Englischen, der Komplementierer "dass" zur Einleitung von subordinierten Sätzen in bestimmten Kontexten optional. Anders als im Englischen ist im Deutschen mit der Auslassung des Komplementierers allerdings ein deutlicher Unterschied in der syntaktischen Strukturierung des eingebetteten Satzes verbunden: während die Struktur mit Komplementierer die Verb-End-Stellung erfordert (1-a), sind asyndetisch eingebettete Sätze Verb-Zweit-Sätze (1-b).

(1) a. Manche sagen, dass die Radikalen sich den Sozialisten annähern.
b. Manche sagen, die Radikalen nähern sich den Sozialisten an.

In einem Sprachproduktionsexperiment haben Lee und Gibbons (2007) [10] für das Englische nachgewiesen, dass der optionale Komplementierer *that* von der wortprosodischen Struktur des eingebetteten Subjekts abhängt. Für den Fall, dass das Subjekt auf der ersten Silbe unbetont ist, wird zur Vermeidung von *stress lapse* der ebenfalls unbetonte Komplementierer signifikant häufiger ausgelassen, als wenn das eingebettete Subjekt mit einer betonten Silbe beginnt. In einem Produktionseexperiment replizieren wir die Studie von Lee und Gibbons für das Deutsche. Hier berichten wir eine begleitende Korpusstudie (schriftsprachliche Korpora), die bereits einen signifikanten Einfluss der Prosodie auf die An- bzw. Abwesenheit des optionalen Komplementierers belegt. Dazu haben wir das TüPP-D/Z Korpus (Zeitungstexte mit 11 Mio. Sätzen bzw. 204 Mio Tokens) nach Vorkommen von acht einbettenden Verben durchsucht, bei denen der Komplementierer optional ist.

		Struktur des eingebetteten Satzes	
		Verbzweit	Verbend (mit *dass*)
Satzspitze	Name	1675	1538
	Artikel	17168	9926

Tabelle 1: *Vorkommen von eingebetteten Sätzen mit Eigenname oder Artikel an der Satzspitze, aufgeschlüsselt nach Nebensatzstruktur. Einbettende Verben:* `finden, glauben, sagen, wissen, denken, meinen, hören, hoffen.`

Wir vergleichen das Vorkommen von "dass" bei eingebetteten Sätze, deren Satzspitze entweder mit einem Eigennamen besetzt ist (der unserer Annahme zufolge meist initial betont ist) oder mit einem (unbetonten) definiten Artikel (*der, die, das*). Die Daten zeigen deutlich, dass An- oder Abwesenheit

von “dass” nicht unahbängig von der Besetzung der Satzspitze (Eigenname vs Artikel) ist (χ^2=154,1, p<0,001). Insbesondere Abfolgen von Komplementierer und Determinierer sind erwartungsgemäß unterrepräsentiert - sie führen zur prosodisch unvorteilhaften Abfolge von mindestens zwei unbetonten Silben (Lapse).

3. Diskussion

Diese Daten legen einen deutlichen Einfluss des linguistischen Rhythmus' auf die Syntax nahe. Bemerkenswert ist, dass der Rhythmus nicht nur die Wortabfolge beeinflusst, sondern bereits die Wahl der syntaktischen Struktur (Haupt- vs. Nebensatz) bedingt. Für Modelle der Sprachproduktion bedeutet dieses Ergebnis (zumindest soweit es zulässig ist, von geschriebener Sprache auf spontansprachliche Produktion zu schließen), dass die Prosodie nicht nur die syntaktische Struktur reflektiert, sondern die Wahl der syntaktischen Struktur mitbestimmt. Modelle, die eine parallele Verarbeitung von Prosodie und Syntax erlauben (z.B. Jackendoff 2002 [11]), werden den Daten deutlich leichter gerecht, als unidirektionale Modelle, in denen prosodische Verarbeitung syntaktische Vorverarbeitung voraussetzt.

4. Literaturverzeichnis

[1] W. Levelt, *Speaking. From intention to articulation*. Cambridge, MA: MIT Press.

[2] G. Dell, L. Burger, W. Svec, “Language production and serial order: A functional analysis and a model”, *Psychological Review*, vol. 104, no. 1, pp. 123–147, 1997.

[3] S. Speer et al., “Situationally independent prosodic phrasing”, *Laboratory Phonology*, vol. 2, no. 1, pp. 35–98, 2011.

[4] M. Wagner, *Prosody and recursion*. Doctoral Dissertation, MIT, 2005.

[5] O. Behaghel, “Zur Wortstellung des Deutschen”, *Language*, vol. 6, no. 4, 1930.

[6] J. Schlüter, *Rhythmic Grammar*, Berlin: De Gruyter, 2005.

[7] A. Zwicky, G. Pullum, *The Principle of Phonology-Free Syntax: introductory remarks*, Columbus: OSU Working Papers in Linguistics, 1986.

[8] T. Scheer, *A guide to morphosyntax-phonology interface theories: how extra-phonological information is treated in phonology since Trubetzkoy's Grenzsignale.* Berlin: De Gruyter, 2010.

[9] A. Anttila, “Phonological Effects on Syntactic Variation”, *Annual Review of Linguistics*, vol. 2, pp. 115-137, 2016

[10] M. Lee, J. Gibbons, “Rhythmic alternation and the optional complementiser in English: New evidence of phonological influence on grammatical encoding”, *Cognition*, vol. 105, no. 2, 2007.

[11] R. Jackendoff, *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: OUP, 2002.

New evidence for prosodic parallelism in German(ic) morphophonology

Gerrit Kentner[1]

[1]Goethe-Universität Frankfurt
kentner@lingua.uni-frankfurt.de

Abstract

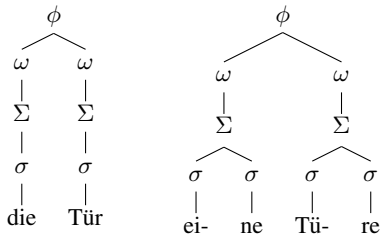
This paper presents two studies that make the case for prosodic parallelism as a factor in German(ic) word formation.

Index Terms: prosodic parallelism, prosody, morphology, prosodic morphology, Germanic

1. Introduction

In their recent contribution, Wiese & Speyer [1] (henceforth W&S) come forward with a very interesting proposal regarding the effect of supra-lexical prosody on word prosodic structure. The proposal, in nutshell, is this: when given the choice, speakers strive for a rendition that maximizes prosodic parallelism; for two words that are prosodic phrase mates, the foot structures are preferably parallel, i.e. the feet have the same number of syllables and stress pattern. W&S build their account of prosodic parallelism on the analysis of optional schwa, examining a large corpus of written German. Among other things, they investigated several cases of nouns with apparently freely alternating monosyllabic and disyllabic variants like *Tür* ∼ *Türe* ('door') or *Tags* ∼ *Tages* ('day$_{Gen}$') in the context of (preceding) monosyllabic or disyllabic determiners.

(1)



Using chi-square tests on bigram frequencies, they disprove statistical independence of the prosodic shapes of co-occuring determiner and noun. The results suggest that, more often than not, the number of syllables in the alternating noun corresponds to the number of syllables in the determiner as in (1), in line with the assumption of a constraint on prosodic parallelism.

In a response to W&S, I pointed out several problems concerning the case of determiner-noun sequences and the use of written corpora to ascertain the effect of prosodic parallelism [3]. Specifically, referring to common reduction phenomena in spoken speech, I questioned W&S's assumption that the determiner corresponds to a prosodic foot. A subsequent study [4] on the alternating adverbs *gern∼gerne, selbst∼selber, lang∼lange* ('happily, oneself, for a long time') preceding various verb forms suggested that avoidance of stress lapse and stress clash, but not prosodic parallelism, account for the presence or absence of the schwa syllable on the adverb. Correspondingly, as it stands, the case for prosodic parallelism as a constraint on word or phrasal prosody appears to be weak.

Here, I present two case studies providing fresh evidence for the role of prosodic parallelism in German(ic) morphophonology. The cases suggest that a constraint on prosodic parallelism, albeit weak, is active on the word and phrasal level.

2. Parallel reduplication in German

The first case concerns rhyme and ablaut reduplications in German (2-a). These word formations are prime examples of prosodic morphology in that reduplication is only licit when native prosodic feet are involved [2]. Although rhyme and ablaut reduplication are mainly found in playful or facetious registers, they are subject to clear restrictions: Firstly, while reduplication is possible on the basis of monosyllables or trochees, reduplication with non-native feet or more complex foot structures are ungrammatical or at least clearly degraded (**Yvónnepivònne* < *Yvónne, *Manuélapanuèla* < *Manuéla*). Secondly, rhyme and ablaut reduplications observe a strict non-identity requirement regarding the segmental structure; base and reduplicant need to differ minimally, yielding the characteristic ablaut or rhyme. Crucially, however, non-identity on the prosodic level (2-b), (2-c) is illicit – the two feet involved in reduplication have to be strictly symmetric, i.e. parallel in shape: if the base is monosyllabic, the reduplicant must be monosyllabic. Conversely, when the base is disyllabic, the reduplicant has to be disyllabic, too.

(2) a. Mischmasch, Hickhack, Krimskrams, Schickimicki, Ilsebilse, doppelmoppel
'mishmash, bickering, bric-a-brac, posh person, proper_name-RED, double-RED'

b. ??Mischemasch, ??Hickehack, ??Krimsekrams, ??Schickimick, *Ilsebils, *doppelmopp

c. *Mischmasche, *Hickhacke, *Krimskramse, *Schickmicki, *Ilsbilse, *doppmoppel

This requirement on reduplication is best captured with the constraint on prosodic parallelism. The data thus constitute evidence for its validity in German morphophonology.

3. Prosodic parallelism in coinages

For the second case study, (mostly English) coinages for musical genres from the website `everynoise.com` were examined. These coinages are names and as such a suitable test case. In contrast to generic words, names are not as open to morphological processes like inflection or derivation which would potentially alter the prosodic rendering.

Besides simplex words (e.g. *pixie*), these coinages are either phrases (e.g. *swedish metal*), or compounds/blends (e.g. *trip hop*). To ascertain the effect of prosodic parallelism, all dyadic coinages (n=714) listed in `everynoise.com` were scrutinised. While the majority of these was non-parallel in na-

ture (e.g. *chicago house*), the subset involving only monosyllables and trochees as members of the dyad (n=498) did show a significant influence of prosodic parallelism (cf. Table 1) over and beyond a strong preference for monosyllabic constituents.

		right constituent	
		monosyll	trochaic
left constituent	monosyll	221	48
	trochaic	164	65

Table 1: *Cross-tabulation of coinages by prosodic shape of left and right constituent.*

A general linear model with binomial link function that was applied to this subset confirms that the prosodic shape of the left member of the dyad (usually the morphological or syntactic dependent) is not independent of the prosodic makeup of the morphological head in the right member (z=2.611, p=0.009). Moreover, the morphosyntactic status of the dyad (compound or phrase) significantly affected the prosodic shape of the left member (z=5.364, $p<0.001$) with a higher number of trochees in the case of phrases.

4. Conclusions

The two case studies suggest that, even though the effect of prosodic parallelism on optional schwa appears to be limited, it nevertheless systematically conditions the phonological makeup of complex words and phrases – at least as long as native prosodic feet (i.e. monosyllables or trochees) are involved.

5. Acknowledgements

Thanks are due to Marc Schwab who helped sieving and annotating the `everynoise.com` corpus.
Note that parts of the introduction are taken verbatim from [3].

6. Bibliography

[1] R. Wiese and A. Speyer, "Prosodic parallelism explaining morphophonological variation in German," *Linguistics*, vol. 53, no. 3, pp. 525–559, 2015.

[2] G. Kentner, "On the emergence of reduplication in German morphophonology," *Zeitschrift für Sprachwissenschaft*, vol. 36, no. 2, to appear.

[3] G. Kentner, "Problems of prosodic parallelism: A reply to Wiese and Speyer (2015)," *Linguistics*, vol. 53, no. 5, pp. 1233–1241, 2015.

[4] G. Kentner, *Schwa optionality and the prosodic shape of words and phrases*. Ms. Uni Frankfurt, 2016.

Relation between articulatory and acoustic information in phonemic representations

Eugen Klein[1], *Jana Brunner*[1], *Phil Hoole*[2]

[1] Humboldt University of Berlin
[2] Ludwig Maximilian University of Munich
eugen.klein@hu-berlin.de, jana.brunner@hu-berlin.de, hoole@phonetik.uni-muenchen.de

Abstract

The Direct Realism approach to speech perception [5] assumes that the acoustic signal produced by a speaker is interpreted by the listener in terms of articulation. This instantaneous mapping between acoustic properties of a speech sound and the articulatory configuration which produced it, assumes a strong causal relation between the two. This relation was investigated here by prompting native Russian speakers to produce the sound /ɨ/ by means of two distinct articulatory configurations resembling Russian steady-state vowels /i/ and /u/. The dissociation between articulation and acoustics is achieved through systematic changes in the direction of auditory perturbation of the second formant (F2) produced by the participants. We report first results of a study which extends empirical knowledge about the role of articulatory information in speakers' phonemic representations.

Index Terms: speech perception, phonemic representations, acoustic and articulatory dissociation, auditory perturbation

1. Introduction

One of the first theoretical approaches to phonemic representations involved the notion of "distinctive features", which were used to derive a descriptive classification of speech sounds [6]. Each speech sound was defined as a set of features, though the authors stayed rather agnostic concerning their nature as these were located on the "motor, acoustic, and auditory level" [6, p. 8]. On the other hand, proponents of Articulatory Phonology defined the articulatory events, gestures in their own terms, as phonological primitives themselves [3]. In accordance to the ideas of the Articulatory Phonology, the Direct Realism (DR) approach to speech perception [5] reinforces the importance of articulatory events as part of speakers' phonological representations as it assumes a strong causal relation between articulatory configurations and corresponding acoustic properties of speech sounds, which essentially allows the listener to interpret the acoustic signal in terms of articulation. In the current study, we investigate the nature of this relation by attempting to dissociate between articulatory configuration of a speaker and her/his own acoustic perception of the corresponding speech sound. This is achieved by means of real-time formant perturbation of speaker's auditory feedback during his/her vowel production.

In a perturbation study, participants hear their own speech production via insert headphones. On each trial, participants are prompted to produce short utterances while formant frequencies (F1 and/or F2) of the contained vowels are shifted up or down. Results of previous perturbation studies using real-time formant perturbation consistently show that participants compensate for the applied shifts in F1 and F2 dimensions [7, 8]. That means that in order to adjust for the formant shifts in the received feedback, participants start to produce higher or lower F1 and/or F2 values. These results suggest that participants try to match their perception of the produced sound with a particular acoustic goal rather than a specific articulatory configuration which leads to articulatory adjustments.

A related but broader question concerns the mapping between the articulatory configuration which a speaker uses to produce a sound and the acoustics of that sound which in most cases, including previous perturbation studies, constitutes a one-to-one relationship. The current study, on the other hand, has the goal to modify participants' articulation-acoustics mapping into a two-to-one relationship in order to assess the causal nature which is assumed by DR.

During the perturbation phase of the current experiment, participants will be asked to produce the sound /ɨ/ embedded in front or back consonantal contexts (i.e., /dɨ/ or /gɨ/). Every time a participant produces /dɨ/, the F2 value of the vowel will be auditorily perturbed to resemble acoustically the vowel /u/. On the other hand, when a participant produces the syllable /gɨ/, the vowel will be perturbed in the direction of /i/ within the F1-F2 vowel space. Based on the findings of the previous perturbation studies, participants are expected to compensate for the effects of the auditory perturbation depending on its direction. Additionally, the two different consonantal contexts should induce participants to produce /ɨ/ by means of the two distinct articulatory configurations they normally use to produce /i/ and /u/. If the causal relation between acoustics and articulation plays a role for perception and speech planning, as predicted by DR, participants should not use the two distinct articulatory configurations to produce the sound /ɨ/. In that case they are expected to choose a single articulatory configuration to produce /ɨ/ in both consonantal contexts.

Two acoustical signals are recorded during the study: i) the auditorily unperturbed signal produced by the participant and ii) the auditorily perturbed signal, which is perceived by the participant. In case participants produce /ɨ/ with two distinct articulatory configurations, there should be a significant difference in F2 values of auditorily unperturbed speech sounds produced in front and back consonantal contexts. On the other hand, the F2 difference between /dɨ/ and /gɨ/ in the auditorily perturbed signal should be small, assuming that /ɨ/ in both syllables has the same acoustic goal for the participant.

2. Methods

2.1. Participants

Seven native speakers of Russian (5 female, 2 male) without reported hearing or speaking disorders participated in the study. The mean age of the group was 27.4 years. Participants have spent on average 4.5 years in Germany at the moment of the recordings. The data collection was carried out at the phonetics lab of the Humboldt University of Berlin.

2.2. Apparatus

The current study made use of the Audapter software package [4] to manipulate formant frequencies produced by participants in real-time. The overall delay of the feedback loop is approximately 14 ms. Figure 1 illustrates the general experimental set-up.

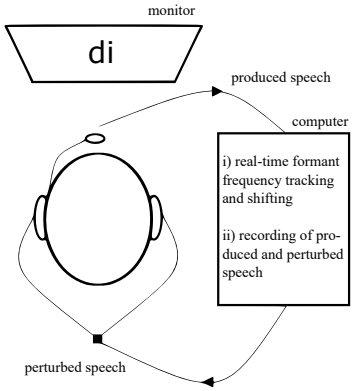


Figure 1: *Schematic diagram of the experimental set-up.*

2.3. Procedure

During an experimental session, participants were seated in front of a computer monitor wearing a Sennheiser MZA 900P neck-worn microphone and E-A-RTONE 3A insert headphones. Each recording session consisted of four phases and lasted for approximately 20-25 minutes (cf. Table 1). On each trial, which had an approximate duration of 2 seconds, participants were visually prompted to produce CV syllables with prolonged vowel portions. The prolongation of the vowel segments made for one thing the formant tracking more reliable and for the other maximized the amount of experimental time during which participants were exposed to perturbed vowels. The interstimulus interval was approximately 1.5 seconds long. All stimuli and experimental instructions were presented in Russian using a Cyrillic font.

Table 1. *Experimental phases and conditions*.

Phase	Stimuli	Perturbation	Trials
Baseline	/di/, /di̵/, /gi̵/, /gu/	no	60
Perturbation	/di̵/, /gi̵/	–/+ 220 Hz	50
		–/+ 370 Hz	50
		–/+ 520 Hz	50

During the baseline phase, participants' production of CV syllables /di/, /di̵/, /gi̵/, and /gu/ was recorded, while participants were able to familiarize themselves with the experimental situation involving auditory feedback delivered over headphones. Note that no auditory perturbation was applied during the baseline phase. After the baseline phase, the first two formants (F1 and F2) were extracted from the baseline recordings to assure the quality of formant tracking and to adjust the tracking settings if needed.

During the three perturbation phases, participants were visually prompted to produce /i̵/ in front or back consonantal contexts (i.e., /di̵/ or /gi̵/). Depending on the context, the second formant (F2) of the vowel /i̵/ was shifted by the perturbation software down or up. The magnitude of perturbation increased in the course of the experiment from 220 to 520 Hz in 150 Hz steps. That means that every time when a participant produced /di̵/ or /gi̵/, she or he was hearing herself/himself producing rather /du/ or /gi/. Participants were not told about the application of auditory perturbation during the experiment. After each session, participants were asked if they had experienced unusual acoustics during the experiment. The majority of the participants reported that their pronunciation changed in the last perturbation phase, which suggests that participants became aware of strong F2 perturbations (520 Hz). However, the analyzed data suggest that participants' partial awareness of auditory perturbation had no impact on their performance.

The signal produced by the participants in the course of the experiment was analyzed with respect to its F2 values to evaluate the main hypothesis of the study, namely whether participants would produce the sound /i̵/ by means of two distinct articulatory configurations.

3. Results

The total recording of seven participants amounted to 1470 trials. Along with audio recordings, the perturbation software stored data files containing the formant values (F1 and F2) tracked on each trial. The formant values were extracted from each file by means of an automatic algorithm. First, the algorithm identified the vowel part within the participant's response by finding the longest continuous part of the formant vector with the smallest standard deviation (SD). Additionally, each trial was manually inspected and the vowel boundaries were adjusted if the algorithm's suggestion was inaccurate. Then, the middle 50% of the identified vowel part was used to calculate the median values for each formant (cf. Figure 2). The same extraction procedure was applied to perturbed and unperturbed formant vectors. 41 trials in total were discarded due to erroneous or missing responses.

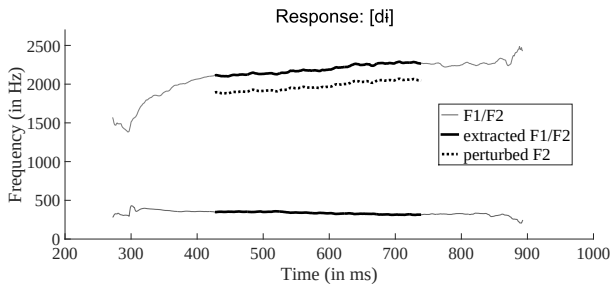


Figure 2*: F1 and F2 formant vectors of participant's response used to extract and compute F1 and F2 median values.*

3.1. Effect of auditory perturbation

In this section, we consider the effect of F2 perturbation on the production of /i̱/ over the course of the experimental phases. The results for all produced syllables during the baseline and perturbation phases are shown in Figure 3. Using R's lme4 package [2], a linear mixed-effects model was fitted for baseline data including the produced response as a fixed effect and F2 frequency as the dependent variable. Random intercepts were modeled for every participant with random slopes for the produced response. The p-values were obtained with lmerTest [1]. The model indicated a significant difference between F2 values for /gu/ and /di/ ($t = -32.558$, $p < 0.01$). Also, F2 value for /gi̱/ was significantly different from /gu/ ($t = 28.982$, $p < 0.01$). Importantly, there was no significant difference of the F2 values between /gi̱/ and /di̱/ ($t = 3.536$, $p > 0.01$), supporting the view that the effect of the consonantal context alone was too weak to dissociate between the two /Ci̱/ syllables.

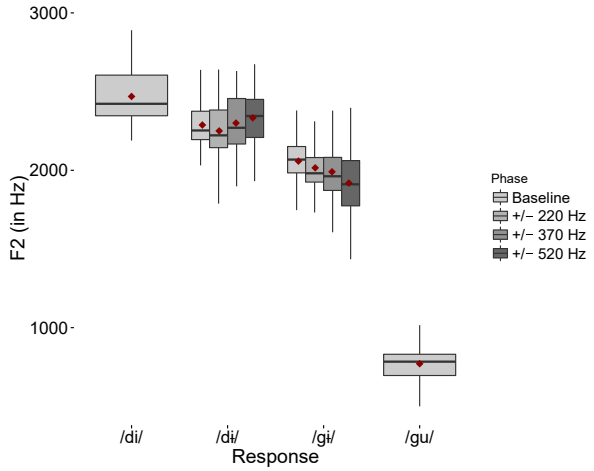


Figure 3: *F2 values produced during the baseline and perturbation phases split by response. Data is pooled across all participants. Red dots represent sample means. For /di̱/, F2 is perturbed downwards, whereas F2 is perturbed upwards for /gi̱/.*

Next, we examine the change in production of /i̱/ in the course of the three perturbation phases. For each response (/di̱/ and /gi̱/), a linear mixed-effects model was fitted including the experimental phase as a fixed effect and F2 frequency as the dependent variable. Both models included random intercepts for every participant with random slopes for the produced responses. As seen in Figure 3, there is a trend of increasing compensation of F2 over the course of the three perturbation phases for /di̱/ and /gi̱/. However, the statistical results support this observation only partially. Compared to the baseline phase, the produced F2 was significantly higher for /di̱/ when perturbation of 370 and 520 Hz was applied ($t = 4.625$, $p < 0.01$; $t = 3.064$, $p < 0.01$). In case of /gi̱/, the produced F2 was significantly lower compared to the baseline phase only when perturbation of 520 Hz was applied ($t = -4.742$, $p < 0.01$). The data indicates that F2 values for /gi̱/ were overall more variable compared to /di̱/.

Finally, we consider the F2 changes in the feedback signal perceived by the participants during the perturbation phases. The same models were fitted as in the previous paragraphs, but were now applied to the perturbed signal. As seen in Figure 4, despite the compensatory adjustments, perceived F2 significantly decreased for /di̱/ ($t = -21.125$, $p < 0.01$; $t = -9.624$, $p < 0.01$; $t = -11.248$, $p < 0.01$) and increased for /gi̱/ ($t = 10.087$, $p < 0.01$; $t = 8.419$, $p < 0.01$; $t = 5.207$, $p < 0.01$) over the course of the perturbation phases. These results suggest that the magnitude of the applied F2 perturbation (220-520 Hz) was too extreme for the participants to compensate for the shifts completely.

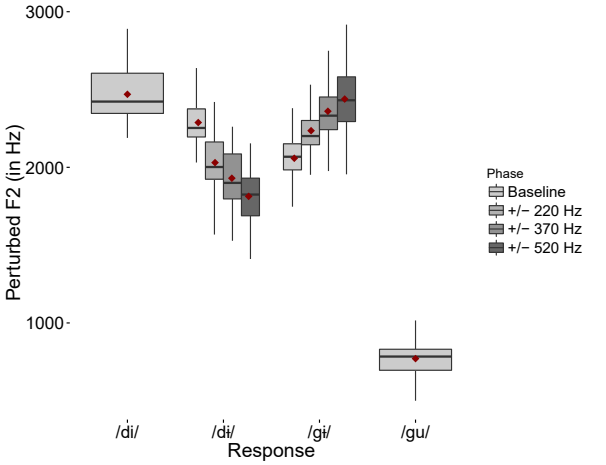


Figure 4: *F2 values perceived by participants during the baseline and perturbation phases split by response. Data is pooled across all participants. Red dots represent sample means. For /di̱/, F2 is perturbed downwards, whereas F2 is perturbed upwards for /gi̱/.*

3.2. Compensation patterns

Box plots with F2 frequencies, plotted separately for each participant, were manually inspected to assess the individual magnitudes of compensation. Comparing the F2 frequencies produced during the baseline phase with F2 frequencies produced during the perturbation phases, we came to the conclusion that different participants applied different strategies to compensate for auditory perturbations.

Two participants compensated for both perturbation directions where compensation in /di̱/ and /gi̱/ productions was approximately 25%, respectively. The relative magnitude of compensation was independent of the magnitude and direction of the perturbation. It is important to point out that during the perturbation phases these participants produced /di̱/ with F2 values as high as or even higher than F2 of their baseline /di/ productions. This finding suggests that during their productions the two participants strongly relied on the acoustic feedback they perceived and less on the articulatory information accessible to them during those trials.

Four participants exhibited a more complex pattern where the F2 frequency for both /Ci̱/ syllables drifted in opposite directions over the course of the perturbation phases. However, with regard to their respective baseline recordings, these four participants produced /di̱/ with approximately the same or lower absolute F2 values during all three perturbation phases. In case of /gi̱/, they compensated approximately 20-55% of the applied perturbation. These results suggest that these four participants tried to keep the produced F2

frequencies of /di̶/ and /gi̶/ similar rather than to aim for absolute F2 values which they had produced during the baseline phase.

One participant exhibited compensation behavior only for the syllable /di̶/ with a magnitude of approximately 40%, thus producing F2 values as high as for the syllable /di/ in the baseline phase. On the other hand, the same participant followed the perturbation direction for the syllable /gi̶/. One possible explanation for this pattern would be that this particular participant employed articulatory information as a distance measure between /di̶/ and /gi̶/ rather than acoustic distance seen in the four participants discussed in the previous paragraph.

4. Discussion

Keeping in mind the rather small study sample, the experimental results reported here are in line with previous perturbation experiments which consistently report compensatory adjustments of the perturbed formant frequencies [4, 7, 8]. As in the study by MacDonald et al. [7], two of our participants produced higher F2 frequencies when their feedback was altered such that they heard themselves producing lower F2. The results reported here extend that general finding in several respects, one of which concerns the absolute compensation magnitude of F2 frequency shifts which was approximately twice as high compared to MacDonald et al.'s study reaching over 200 Hz. Differently from previous perturbation studies, in the current experiment the F2 shifts were applied simultaneously in both directions along the frequency scale depending on the consonantal context of the produced vowel. The first observation was that participants were able to adapt rapidly their compensation strategies to produce the target sound /i̶/ despite the opposite shifts in F2. In order to do this, the majority of the tested participants were using essentially two different articulatory configurations, as their productions of /di̶/ and /gi̶/ kept drifting apart with increasing magnitude of F2 perturbation. However, there were substantial differences in compensation behavior between the two syllables regarding F2 amplitudes. One possibility to explain the observed tendency of the syllable /di̶/ to have more stable F2 values as well as the on average substantial difference in compensation magnitudes between /di̶/ and /gi̶/ are different physiological constraints associated with the two syllables. In the case of /di̶/, the forward movement of the tongue required to compensate for the applied perturbation was limited by the alveolar ridge and the upper incisors. On the other hand, compensation movement in the case of /gi̶/ was directed towards pharynx allowing the tongue to travel a farther distance along the palate. Also, it is worth mentioning that perturbation of /di̶/ in the direction of /du/ resulted in a qualitatively different percept than when /gi̶/ was perturbed in the direction of /gi/. That difference in perception between /di̶/ and /gi̶/ may have additionally contributed to different compensation strategies for these two syllables. That difference is most pronounced in the case of one speaker who compensated for the perturbation in the case of /di̶/ but shifted F2 in the same direction during her production of /gi̶/. It seems that in the case of /gi̶/, this participant directed her compensation towards an acoustic goal but employed articulatory information during the production of /gi̶/ keeping the articulatory distance between the two syllables constant across the baseline and the perturbation phases.

Taken together, these findings suggest, on the one hand, that the absolute magnitude of compensation and the compensation strategies depend more strongly on (individual) physiological and perceptual constraints rather than on the general capabilities of the feedback mechanism *per se*. On the other hand, use of different articulatory configurations along with different compensation strategies for /di̶/ and /gi̶/ speaks against the idea central to the theory of DR that specific articulatory configurations and the corresponding speech sounds are causally linked. On the contrary, the current study demonstrates that this relation is highly flexible within and across speakers. The results further show that the majority of tested speakers strongly rely on the acoustic information to produce the intended speech sound. An intriguing finding in this respect is the observation that instead of acoustic goals in terms of absolute frequency values, a portion of participants seems rather to employ a notion of acoustic similarity when compensating for auditory perturbations.

A central follow-up question concerns the agile and robust ability of native speakers to produce target speech sounds despite the extreme articulatory changes required due to auditory perturbation. A plausible and verifiable hypothesis is that this ability strongly depends on speakers' experience with speech production in their target language. We will address this hypothesis in our future research by investigating compensatory strategies employed by L2 learners of Russian.

5. Acknowledgements

We gratefully acknowledge support by DFG grant 220199 to JB. We thank all participants who took part in the study.

6. References

[1] Kuznetsova, A. Brockhoff, P.B., Bojesen Christensen, R. H., 2016. lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-32. https://CRAN.R-project.org/package=lmerTest

[2] Bates, D., Maechler, M., Bolker, B., Walker, S. 2016. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-12. http://CRAN.R-project.org/package=lme4

[3] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica*, 49, 155-180.

[4] Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., Perkell, J. S. 2008. A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In: Sock, R., Fuchs, S., Laprie, Y. (eds), *Proceedings of the 8th Intl. Seminar on Speech Production.* 65-68.

[5] Fowler, C. A. 1996. Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am.*, 99, 1730-1741.

[6] Jakobson, R., Halle, M. 1956. *Fundamentals of language.* Den Haag: Mouton.

[7] MacDonald, E. N., Goldberg, R., Munhall, K. G. Compensations in response to real-time formant perturbations of different magnitudes. *J. Acoust. Soc. Am.*, 127, 1059-1068.

[8] Villacorta, V. M., Perkell, J. S., Guenther, F. H. 2007. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.*, 122, 2306-2319.

Der Einfluss der F0-Kontur als akustischer *cue* für chinesische Deutschlerner in konkurrierenden Kontexten

Nicola Klingler

Institut für Schallforschung – Österreichische Akademie der Wissenschaften

nicola.klingler@gmail.com

Abstract

Der lexikalische Ton prägt die chinesischen Sprachen nicht nur intralingual, sondern stellt für chinesische Sprecher darüber hinaus das salienteste Merkmal zur perzeptiven Identifikation von Silben auch außerhalb der chinesischen Sprachen dar (siehe bspw. [1] für das Englische). Dass dabei der sprachspezifische Gebrauch der F0-Kontur in der *target language* (TL) eher unwichtig ist, zeigen die Ergebnisse des durchgeführten Experiments, welches die Perzeption des deutschen Wortakzent untersuchte: Obwohl die F0-Kontur im Deutschen meist nur zur Markierung der Intonation – und somit zur pragmatischen Differenzierung – oder zur Markierung des Fokusakzents in den Vordergrund tritt [2], vertrauten die in diesem Perzeptionsexperiment untersuchten chinesischen Sprecher so stark auf den *cue* F0-Kontur, dass dieser, auch wenn er in Konkurrenz mit divergierenden *cues* (hier: Vokallänge, Vokalqualität und Intensität) stand, einen marginal signifikanten Einfluss auf die Wahrnehmung hatte.

Schlüsselbegriffe: Fremdspracherwerb, Phonetik, Phonologie, Prosodie

1. Einleitung

Transfererscheinungen aus einer Erstsprache (L1) auf eine TL wurden bisher meist auf segmentaler, seltener jedoch auf suprasegmentaler Ebene betrachtet (siehe jedoch [3], [4], [1], [5] und [6]). Im Folgenden wird die Perzeption des deutschen Wortakzents durch Mandarin-Chinesisch Sprecher (MA L1-Sprecher) untersucht, da einige Ausspracheabweichungen von MA L1-Sprechern vermutlich auf deren Perzeption [1], und dabei speziell auf die suprasegmentalen Eigenschaften der Erstsprache zurückzuführen sind. Dabei wird der Fokus auf den Einfluss der F0-Kontur gelegt, da diese im Mandarin-Chinesischen primär für die Produktion des lexikalischen Tons – welcher somit ein inhärentes Merkmal der Silbe darstellt [7] – im Deutschen hingegen auf einer Ebene über der Silbe, zur Gliederung des Satzes (prosodische Phrasen) und zur Konstituierung des Fokusakzents genutzt wird [8]. Die hier vorgestellten Daten wurden durch ein, an [1] angelehntes Online-Perzeptionsexperiments erhoben.

2. Theoretischer Hintergrund

Der deutsche Wortakzent konstituiert sich aus mehreren akustischen Korrelaten (Vokallänge, Vokalqualität, Intensität und Stimmtonverlauf) die in verschiedenem Maße relevant für die Wahrnehmung desselben sind ([9] und [7]). So führen [2] und [10] an, dass eine höhere Intensität und die Eigenschaften des Vokals von Deutsch L1-Sprechern (D L1-Sprecher) eher zur Determinierung des Wortakzents genutzt werden, als der Stimmtonverlauf. Im Mandarin-Chinesischen hat jedoch die Vokalquantität keine distinkitve Funktion [11] und auch die Vokalqualität ist gegenüber dem Stimmtonverlauf untergeordnet [11]. Der lexikalische Ton nimmt die prominenteste Position innerhalb der (supra-)segmentalen Merkmale im Mandarin ein; somit ist es, vor dem Hintergrund von bspw. der Interlanguage-Hypothesis [12][1] oder des Perceptual Assimilation Model [14] plausibel anzunehmen, dass sich an dieser Stelle Probleme für MA L1-Sprecher ergeben, wenn sie Deutsch als Zweitsprache erlernen. Es wird angenommen, dass Sprecher einer Tonsprache aufgrund ihrer L1 sensibler auf Veränderungen der F0-Kontur in der TL reagieren und dies zu Perzeptionsproblemen in einer Sprache wie dem Deutschen führen kann.

Wie bereits bei [1] deutlich wurde, sind MA L1-Sprecher nicht *stress-deaf* und sollten die Unterschiede zwischen den betonten und unbetonten Silben wahrnehmen und interpretieren können. Aufseiten der Logatome wird allerdings erwartet, dass es zu Abweichungen in der Erkennbarkeit kommt. Aufgrund der Ergebnisse von [1] liegt die Vermutung nahe, dass die MA L1-Sprecher sensibler auf Veränderungen in der F0-Kontur reagieren und diese dann fälschlicherweise als betonte Silben interpretieren, d. h., dass die Antworten der MA L1-Sprecher direkt mit der Manipulation der F0-Kontur korrelieren. Falls sich herausstellt, dass die F0-Kontur als wichtigster *cue* für die chinesischen Sprecher zu werten ist, würde sich dieses Ergebnis mit den Erkenntnissen aus [15], [1] und [6] decken. Wenn die F0-Kontur dazu auf die D L1-Sprecher keinen signifikanten Einfluss hat, so würde dieses Ergebnis für die These sprechen, dass Sprecher einer Tonsprache sensibler auf Veränderungen der Grundfrequenz reagieren und diese auch dann als *cue* nutzen, wenn sie in der TL andere Funktionen bedient.

3. Experiment

Das hier durchgeführte Perzeptionsexperiment sollte gezielt untersuchen, auf welche akustischen *cues* MA L1-Sprecher am Stärksten vertrauen, wenn sie keine lexikalischen Informationen haben, um die Position des Wortakzents festzustellen. Dabei wurden sowohl Logatome als auch echte Minimalpaare[2] des Deutschen von einem D L1-Sprecher produziert, manipuliert und dann im Experiment abgefragt. Die Manipulationen betrafen dabei die phonetischen Merkmale (Vokaldauer, Vokalqua-

[1] Zur Erweiterung der Interlanguage-Hypothesis auf phonetische Phänomene siehe bspw. [13].

[2] Auch bei den Minimalpaaren ist die lexikalische Information elidiert, da sich die Minimalpaare (bspw. Kon*stanz* vs.*Kon*stanz) nur durch die Position des Akzents unterscheiden. Die hierzu erhobenen Daten bestätigten, dass die MA L1-Sprecher nicht *stress-deaf* sind, werden allerdings im Folgenden nicht genauer diskutiert.

lität und F0-Kontur) der Logatome, welche mit den akustischen *cues* für den deutschen Wortakzent in Fokusposition korrelieren.[3]

3.1. Manipulation der Stimuli

Die Logatome bestanden aus zwei Silben mit Onsetclustern die der deutschen, nicht jedoch der Mandarin-Chinesischen Phonologie entsprachen. Die Silben wurden dann aneinander gefügt und von einem D L1-Sprecher in einem Trägersatz, in verschiedenen Betonungsvarianten eingesprochen. Somit ergab sich ein Set aus vier Stimuli deren *cues* mit Praat [16] manipuliert wurden.

statro		trosta	
A	B	**B**	A
A	**B**	B	**A**

Tabelle 1: Betonungsvarianten der Testwörter

Zur Manipulation der F0-Kontur[4] wurden die Stimuli in Praat im *Manipulation Editor* geöffnet. Dann wurde durch *stylize (2st)* die vorhandene F0-Kontur reduziert, sodass nur noch ein *peak* oder ein Tiefpunkt[5] vorhanden war. Diesen Punkten wurden die Werte (in Hertz) gegenübergestellt und gleichmäßige Schritte zwischen H* und L* berechnet. Somit ergibt sich aus jedem Testwort ein Set aus fünf Schritten:

- original Stimuli mit durch Praat vereinfachter, aber nicht veränderter F0-Kontur (H* liegt auf der betonten Silbe, L* auf der unbetonten)
- 1. Schritt: H* wird um einen Schritt reduziert
- 2. Schritt: *mean*-Wert, die F0-Kontur ist somit eine horizontale Linie
- 3. Schritt: die F0-Kontur des original-Stimulus wird invertiert, so dass H* und L* vertauscht werden
- 4. Schritt: die F0-Kontur von Schritt 1 wird invertiert[6]

Diese Manipulationen wurden, um eine ausbalancierte Testgrundlage zu erzielen, auf alle vier Testwörter angewandt. Dabei bewirkt die Manipulation der F0-Kontur keine Veränderungen an der Intensität oder an anderen Eigenschaften des Signals, das bedeutet, dass hier die übrigen *cues* in Konkurrenz mit dem manipulierten *cue* stehen. Insgesamt entstanden so 46 Stimuli (4*5 F0, 2(4*3) Vokalqualität und Vokallänge, 2 unmanipulierte Stimuli).

3.2. Partizipanten

Die Partizipantengruppe bestand aus zehn Sprechern, die nach eigenen Angaben Mandarin-Chinesisch als Muttersprache sprechen. Allerdings haben einige (n=2) Sprecher darüber hinaus angegeben, dass sie die örtlichen Dialekte, bspw. Taiwanisch sprechen.[7] Vor dem Experiment musste ein Fragebogen ausgefüllt werden, welcher biographische als auch sprachspezifische Daten abgefragt hat; bspw. wo die Partizipanten aufgewachsen sind, welche Sprachen sie in welcher Reihenfolge gelernt haben, in welchem Umfeld die Partizipanten Deutsch erlernt haben und ob sie länger als drei Monate in Deutschland gelebt haben. Neben den chinesischen Sprechern wurden auch Daten von fünf Deutsch L1-Sprechern (D L1-Sprechern) als Kontrollgruppe erhoben. Die Ergebnisse dieser Gruppe stellen die *baseline* dar, mit der die Daten der MA L1-Sprecher verglichen werden. Die Stimuli wurden den Partizipanten auditiv präsentiert [18]. Dabei mussten sie entscheiden, welche Silbe für sie betont klingt, die Antwortmöglichkeiten bestanden dabei aus: erste Silbe betont (S1) und zweite Silbe betont (S2). Das Audiosignal konnte mehrfach wiederholt werden, allerdings wurden die Partizipanten angewiesen, sich die Aufnahme (nur) zweimal anzuhören. Das Ende des Experiments wurde durch eine abschließende Seite angezeigt. Alle erhobenen Daten wurden automatisch anonymisiert. Jeder Stimulus wurde dreimal wiederholt, somit bewerteten die Partizipanten 60 (3(4*5)) Stimuli mit einer manipulierten F0-Kontur.

4. Ergebnisse

4.1. Analyse mit R

Die Analyse der Daten wurde mit R [19] durchgeführt. Da die Partizipantenanzahl zu keiner ausbalancierten Datenlage führte, konnten die Bedingungen nicht gegeneinander getestet werden, sondern wurden stattdessen einzeln betrachtet. Durch die Manipulation der Stimuli konnte a priori festgehalten werden, welche Silbe – phonetisch gesprochen – markiert ist und somit als betont wahrgenommen werden sollte. Aus diesem Grund wurde für die Darstellung der Bedingung F0 mit Hilfe des harmonischen Mittels die Genauigkeit der gegebenen Antwortet berechnet.

$$harmonisches\ Mittel = 2 * \frac{precision * recall}{precision + recall} \quad (1)$$

Das harmonische Mittel (F) ist ein Indexwert, der die Fehlerwerte der Daten ausgleicht und auf die Werte *recall* und *precision* angewandt wird.

$$recall = \frac{relevant + found}{relevant} \quad (2)$$

$$precision = \frac{relevant + found}{found} \quad (3)$$

Diese stellen dar, wie groß die Schnittmenge der gewählten Antworten mit den möglichen Antworten und den richtigen Antworten ist.

In Abbildung 1 ist dargestellt, wie sich die Daten der Partizipanten über die verschiedenen Schritte in der Bedingung F0 in Bezug zu F verhalten haben. Partizipant t487 weist einen Wert von 0.5 auf, was bedeutet, dass dieser in 50% der Fälle richtig geantwortet hat. Bei den übrigen Partizipanten sieht man, dass die Ergebnisse auf der ersten Stufe alle auf oder über 50% liegen, wohingegen bei den Stufen drei und vier die Präzision insgesamt nachgelassen hat und mehrere Partizipanten eher niedrige Werte erreicht haben. Interessant ist, dass die Ergebnisse

[3] Auf eine Manipulation der Intensität wurde verzichtet, da dies im gegebenen Experimentdesign nicht kontrolliert abgefragt werden konnte.

[4] Im Laufe dieses Experiments wurden auch die *cues* Vokallänge und Vokalqualität getestet. Die Ergebnisse werden hier nicht separat aufgeführt, fließen jedoch in die Diskussion der Daten mit ein.

[5] Im Folgenden wird auf die jeweiligen *peaks* und Tiefpunkte mit H* bzw. L* verwiesen (in Anlehnung an [7] und [17].

[6] Ursprünglich wurden sechs Schritte für diese Bedingung konstruiert, zur Untersuchung jedoch nur die aufgeführten Schritte herangezogen. Die nicht untersuchten Schritte beeinhalteten level-Konturen, die auf Höhe des jeweiligen H* und auf Höhe des jeweiligen L* lagen.

[7] Diese Sprecher wurden trotzdem in der untersuchten Gruppe behalten, da sich die Untersuchung in erster Linie auf die Annahme stützt, dass Sprecher einer Tonsprache mit der Verwendung der F0-Kontur im Deutschen Probleme haben.

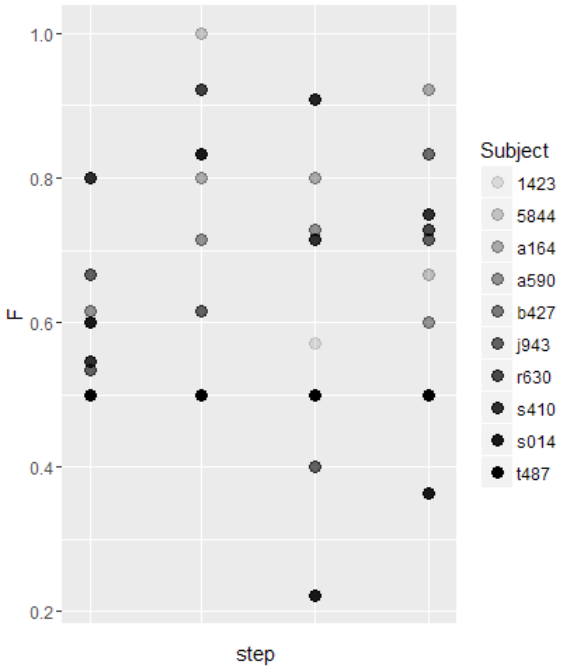


Abbildung 1: Bedingung F0, aufgeschlüsselt nach den einzelnen Partizipanten

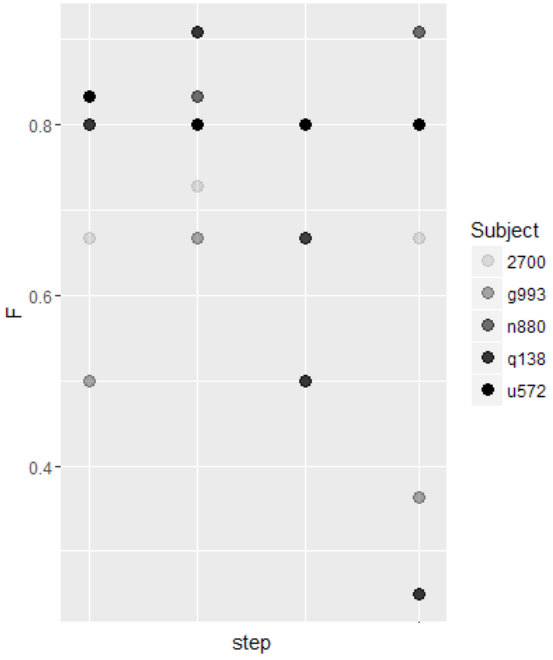


Abbildung 2: Bedingung F0, aufgeschlüsselt nach den einzelnen Partizipanten (Kontrollgruppe)

für die zweite Stufe deutlich besser sind (fünf Sprecher, deren Ergebnisse über 70% lagen), als für die übrigen Stufen, obwohl die zweite Stufe die *mean* F0-Kontur kodiert. Da die Ergebnisse innerhalb der Bedingung annähernd normalverteilt sind, wurde eine ANOVA durchgeführt, die für diese Bedingung keine signifikanten Ergebnisse bestätigt ($p = 0.058$) und somit der *cue* F0-Kontur keinen signifikanten Einfluss auf die Partizipanten zeigt.

In Abbildung 2 sind ebenfalls die Ergebnisse der Bedingung F0 in Bezug zu F dargestellt. Wie zu sehen ist, hatten die D L1-Sprecher insgesamt weniger Probleme die richtige Betonung festzulegen. Allerdings ist interessant, dass die Präzision auf den Schritten 3 und 4 stark nachlässt. Besonders die große Streuung auf Schritt 4 deutet darauf hin, dass hier Probleme auftraten.[8] Auch in diesem Fall wurde eine ANOVA durchgeführt, die zeigt, dass der Einfluss der f0 auf den einzelnen Schritten im Kontinuum nicht signifikant ist ($p = 0.26$).

4.2. Diskussion der Ergebnisse

Abbildung 2 zeigt, dass die D L1-Sprecher den *cue* F0 wahrnehmen und nutzen können, und dessen Abwesenheit möglicherweise einen Einfluss – wenngleich nur minimal – auf die Erkennbarkeit der Akzentuierung hat. Dies könnte darauf zurückzuführen sein, dass ausschließlich der Fokusakzent im Deutschen mit der F0-Kontur angezeigt wird und somit die Ergebnisse der Bedingung F0 auf den Schritten 3 und 4 in der Kontrollgruppe vor allem zeigen, dass die zuverlässige Determiniation der Betonungsposition durch die Diskrepanz der verschiedenen *cues* gestört wird. Im Gegensatz dazu verdeutlicht Abbildung 1, dass die MA L1-Sprecher sensibler auf die Veränderungen der F0-Kontur reagiert haben und hier der Einfluss der F0-Kontur zu marginal signifikanten Ergebnissen geführt hat. Hierbei ist interessant, dass die MA L1-Sprecher höhere Erkennungswerte auf dem zweiten Schritt im Kontinuum aufweisen. Dies könnte in Anlehnung an die Ergebnisse der Kontrollgruppe darauf hinweisen, dass die Perzeption zwar größtenteils von der F0-Kontur beeinflusst wird, bei Fehlen derselben jedoch auf die anderen vorhandenen, übereinstimmenden *cues* zurückgegriffen werden kann. Das könnte bedeuten, dass die MA L1-Sprecher die *cues* Vokallänge und Vokalqualität vor allem dann nutzen können, wenn sie übereinstimmend die Position des Wortakzents kodieren.

Die Ergebnisse sind dabei durchaus mit den Daten von [1] vergleichbar und zeigen deutlich, dass die F0-Kontur einen starken Einfluss auf die Perzeption der MA L1-Sprecher hat und sich dies in der Wahrnehmung und Interpretation von suprasegmentalen Eigenschaften der TL manifestiert. Darüber hinaus zeigt sich hier, dass der Einfluss auch in konkurrierenden Kontexten aufzuzeigen ist und gegen hierarchisch wichtigere *cues* in der TL durchgesetzt werden kann.

Zu diskutieren bleibt, wie zu erklären ist, dass die Partizipanten auf den Schritten drei und vier eine reduzierte Erkennungspräzision aufweisen. Hierbei ist, übereinstimmend mit den Ergebnissen von [1], zu bemerken, dass bei den MA L1-Sprechern keine Präferenz für eine initiale Betonung nachzuweisen ist.[9] So ist die erste Silbe über die Bedingungen hinweg nicht nur seltener als betont erkannt worden, es wurde darüber hinaus auch auffällig häufig die Mitte des Kontinu-

[8] Das Fehlen einiger Datenpunkte ist auf die zugrundeliegende Rechnung zurück zu führen. Wenn einer der Werte für *recall* oder *precision* bei 0 liegt, so kann F nicht berechnet werden.

[9] Auch der Einfluss der Testworte (statro vs. trosta), sowie der *Length of Residence* und der *Age of Acquisition* wurde überprüft und zeigte, dass keine Effekte vorzufinden waren.

ums (bei den Bedingungen Vokallänge und Vokalqualität) als final betont wahrgenommen, obwohl hier zu Erwarten wäre, dass keine eindeutige Zuordnung möglich ist. Dies könnte darauf zurückzuführen sein, dass in den Bedingungen Vokalqualität und Vokallänge die F0-Kontur keine akustische Information übertragen kann, da sie auf dem jeweiligen *mean*-Wert steht. In Anlehnung an [1] wird deshalb festgehalten, dass die MA L1-Sprecher ihre Entscheidung, welche Silbe betont ist, nicht nur auf Basis des Signals getroffen haben können, sondern die Übertragung eines Betonungsmusters aus der L1 plausibel scheint. Dies würde, in Anlehnung an [20] bedeuten, dass die MA L1-Sprecher die finale Betonung des Mandarin-Chinesischen übertragen. [21] und [22] haben jedoch gezeigt, dass das Betonungsmuster für bisyllabische Worte im Mandarin-Chinesisch aus zwei normalbetonten Silben oder — in selteneren Fällen -– aus einer betonten und einer schwach bzw. unbetonten Silbe besteht. Vor diesem Hintergrund betrachtet, zeigen die Ergebnisse nicht, dass die MA L1-Sprecher die erste Silbe als unbetont interpretieren oder eine finale Betonung des Mandarin-Chinesischen übertragen. Da im Mandarin-Chinesisch nur die zweite Silbe unbetont sein kann, ist es verständlich, dass die chinesischen Partizipanten auch nur hier die optionale Möglichkeit zur Betonung sehen. Deshalb zeigen die Ergebnisse vor allem, dass die MA L1-Sprecher bei einer reduzierten Markierung (d.i., nicht alle *cues* markieren eine betonte Silbe) auf ihr L1-Betonungsmuster (die erste Silbe ist immer betont) zurückgreifen und entscheiden, ob die zweite Silbe möglicherweise betont ist; bspw. aufgrund einer höheren Intensität[10] oder einer ausgeprägteren F0-Kontur, als im Mandarin-Chinesischen für unbetonte Silben üblich. Die Ergebnisse der Bedingung F0 gehen damit einher, da dort die ersten Schritte im Kontinuum mit höherer Präzision eingeordnet wurden, als die Schritte auf denen sich die *cues* widersprechen.

5. Fazit

Das hier durchgeführte Experiment untersuchte die Perzeption des deutschen Wortakzents durch MA L1-Sprecher. Dabei wurden den Partizipanten manipulierte Stimuli präsentiert, in welchen einige *cues* des deutschen Wortakzent manipuliert und in Konkurrenz zu einander gestellt wurden. Es kann festgehalten werden, dass Sprecher einer chinesischen Tonsprache Probleme mit der Wahrnehmung des deutschen Wortakzents haben, sobald mindestens ein akustischer *cue* desselben in Diskrepanz mit den übrigen steht. Darüber hinaus hatte auf die hier untersuchten Sprecher die Veränderung der F0-Kontur den größten Einfluss ($0.5 > p > 0.05$), woraus abgeleitet werden kann, dass die Interlanguage der Sprecher auf suprasegmentaler Ebene stark von der L1 geprägt ist, da die Sensibilität gegenüber der F0-Kontur vermutlich auf das lexikalische Tonsystem der L1 zurückzuführen ist. Dies deutet weiterführend darauf hin, dass die Gewichtung der *cues* für prominente oder betonte Silben in der Perzeption von MA L1-Sprechern dahingehend von D L1-Sprechern abweicht, als dass die F0-Kontur an erster Stelle und die Intensität (vermutlich) an zweiter Stelle steht. Die übrigen *cues* (Vokalqualität und Vokaldauer) sind dagegen weniger wichtig.

[10] Diese wurde, wie bereits gesagt, im durchgeführten Experiment nicht manipuliert und markiert somit immer die betonte Silbe.

6. Bibliographie

[1] Q. Wang, "L2 stress perception: The reliance on different acoustic cues," in *Proceedings of the 4th Conference on Spreech Prosody*, Campinas, 2008, pp. 6–9.

[2] A. M. Kijak, "How stressful is L2 stress? A cross-linguistic study of L2 perception and production of metrical systems," Ph.D. dissertation, Universität Utrecht, Lot, 2009.

[3] H. Ding, O. Jokisch, and R. Hoffmann, "F0 Analysis of Chinese Accented German Speech," in *Proceedings of International Symposium on Chinese Spoken Language Processing*, Singapur, 2006.

[4] H. Ding and R. Hoffmann, "An Investigation of Prosodic Features in the German Speech of Chinese Speakers," in *Prosody and Language in Contact. L2 Acquisition, Attrition and Languages in Multilingual Situations*, ser. Prosody, Phonology and Phonetics, E. Delais-Roussarie, S. Herment, and M. Avanzi, Eds., 2015, pp. 221–241.

[5] Y. Zhang, S. L. Nissen, and A. L. Francis, "Acoustic characteristics of English lexical stress produced by native Mandarin speakers," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4498–4513, 2008.

[6] M. Ploquin, "Prosodic Transfer: From Chinese Lexical tone to English Pitch Accent," vol. 4, no. 1, pp. 68–77, 2013.

[7] C. Gussenhoven, *The phonology of tone and intonation*. Cambridge University Press, 2004.

[8] R. Wiese, *The phonology of German*, reprinted ed., ser. The phonology of the world's languages. Oxford: Univ. Press, 1996.

[9] T. Becker, *Einführung in die Phonetik und Phonologie des Deutschen*. Darmstadt: WBG, 2012.

[10] M. Jessen, "Word stress in West-Germanic languages: German," in *Word Prosodic Systems in the Languages of Europe*, H. Van der Hulst, Ed. Berlin: Mouton de Gruyter, 1999, pp. 515–545.

[11] C. Hunold, *Untersuchungen zu segmentalen und suprasegmentalen Ausspracheabweichungen chinesischer Deutschlerner*, ser. Hallesche Schriften zur Sprechwissenschaft und Phonetik. Peter Lang, 2009, vol. 28.

[12] L. Selinker, "INTERLANGUAGE," *IRAL - International Review of Applied Linguistics in Language Teaching*, vol. 10, 1972.

[13] K. J. Keys, "Interlanguage phonology. Theoretical questions and empirical data," *Linguagem & Ensino*, vol. 5, no. 1, pp. 75–91, 2002.

[14] C. Best, "A direct realist perspective on cross-language speech perception." in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research*. York: Timonium, MD, 1995, pp. 171–204.

[15] Y. Chen, "From Tone to Accent: The Tonal Transfer Strategy fo Chinese L2 learners of Spanish," in *Proceedings of the 15th ICPhS Conference*, Barcelona, 2007, pp. 1043–1048.

[16] P. Boersma and D. Weenik, "Praat: Doing phonetics by computer," 2014. [Online]. Available: http://www. fon. hum. uva. nl/praat/.

[17] M. Grice, S. Baumann, and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology," in *Prosodic typology*. Oxford Univ. Press, 2005, pp. 55–83.

[18] E. Onea and A. Syring, *Courant Research Centre "Text Structures"*, 2011. [Online]. Available: https://onexp.textstrukturen.uni-goettingen.de/

[19] R Development Core Team, "A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Wien, Österreich, Tech. Rep., 2008. [Online]. Available: http://www.R-project.org.

[20] J. Norman, *Chinese*, ser. Cambridge Language Surveys. Cambridge University Press, 2002.

[21] L. Wang, *Chinese Phonology*. Beijing: Zhonghua Shuju, 2008.

[22] C. Ni, W. Liu, and B. Xu, "From English pitch accent detection to Mandarin stress detection, where is the difference?" *Computer Speech & Language*, vol. 26, no. 3, pp. 127–148, Jun. 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0885230811000477

Big Data for analyses of small-scale regional variation: A case study on sound change in Swiss German

Adrian Leemann[1], *Marie-José Kolly*[2]

[1]Phonetics Lab., Department of Theoretical and Applied Linguistics, University of Cambridge
[2]Phonetics Lab., Department of Comparative Linguistics, University of Zurich
al764@cam.ac.uk, marie-jose.kolly@uzh.ch

Abstract

In this case study we examine sound change of *Altoberdeutsch* <iu> in Swiss German dialects. We used contemporary dialect data from nearly 60,000 speakers – collected with the smartphone app *Dialäkt Äpp* – and compared it to historical *Atlas* data from the 1950s. Results revealed hierarchical and contra-hierarchical diffusion patterns for some dialectal variants, while other variants remained virtually unchanged over the course of seven decades. We further report change in apparent time, with older speakers using traditional variants more frequently than younger speakers. Using this case study as a model, future work using the *Dialäkt Äpp* corpus will reveal patterns of feature diffusion and dialect leveling on a larger scale.

Index Terms: sound change; dialect leveling; crowdsourcing; dialectology; Swiss German

1. Introduction

The most recent large-scale study on Swiss German (hereafter SwG) dialects – the *Sprachatlas der Deutschen Schweiz* (*Atlas* for short, [1]) – dates back 60–70 years. It documents the linguistic situation of German-speaking Switzerland in the first half of the 20th century for 566 localities. Anecdotal evidence and previous, mostly small-scale studies, revealed that dialects have changed considerably since then. Yet our understanding of how dialects have changed on a regional level remains patchy. In this contribution, we will contribute to fill this gap with a case study using Big Data that was crowdsourced with the free iOS app *Dialäkt Äpp* (*DÄ* for short, [2]). *DÄ*'s main function is the prediction of the user's dialect [3]. For 16 variables, users select their dialect variant from a drop-down menu. For the variable *Bett* 'bed', for example, they choose from the variants [bɐt] (as used in Western SwG) or [bet] (Eastern SwG). At the end of the quiz, *DÄ* guesses which dialect the user speaks. Underlying this prediction are 16 maps from the *Atlas*. Following dialect prediction, users can evaluate the result and indicate their actual dialect. With this information, the 16 variables can be assessed for language change (*Atlas* vs. *DÄ* data). A first pilot revealed global patterns of language change in SwG [4]. The large bulk of this corpus, however, is yet to be analyzed, especially with regard to in depth, small-scale analyses of regional variation and change. The objective of the present proof of concept study is the analysis of small-scale, regional diachronic variation in SwG dialects in the variable *Altoberdeutsch* <iu>. Because of its historical nature, we use grapheme symbols to present this variable.

1.1. Previous studies

Only a few studies have examined how SwG dialects have changed since the *Atlas*. [5] and [6] reported change in the lexicon. The latter found convergence tendencies towards Standard German and showed that younger speakers deviated from the *Atlas* more than older speakers in lexical features. Similarly, [5] conducted an online survey with 9000 participants. Based on this study, [7] as well as [8] presented dialect maps that corroborate tendencies of leveling in the lexicon for some of the variables examined. For morphosyntax, [9] found that only little change has occurred for the constructions examined. A number of studies have further reported sound change over the past decades, such as [10], [11] and [12]. The two latter investigations documented significant change for Aarau. On the whole, variants that were documented in the *Atlas* are still in use in Aarau, yet they co-exist alongside additional, more frequently used variants. /l/-vocalization in particular has received much attention in the literature. A number of studies report the spread of vocalized /l/ to regions not previously attested as vocalizing in the *Atlas*: towards Luzern [13, 14], Fribourg [15], Central Switzerland [16], and the Bernese Oberland [17, 18]. Our own research revealed change between the *Atlas* and today: a recently conducted study applying a rapid anonymous survey framework [19] indicates that /l/-vocalization has spread in a southerly, westerly, and central direction within German-speaking Switzerland. [4], using the *DÄ* corpus, revealed that phonetic variables demonstrated most agreement with the *Atlas* (67%), followed by the morphological (59%), and the lexical variable (53%). Until today, however, we have not investigated small-scale regional patterns of language change to the level of detail required, using *DÄ* data.

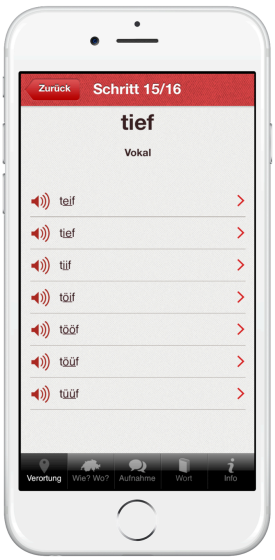
1.2. Research questions

In the present contribution, we provide a case study for small-scale, diachronic analyses of one variable, *Altoberdeutsch* <iu>, which – in most cases – stems from (Proto-)Germanic <eu> [20]. Both [20] and [21] claim <eu> to be one of the most complex variables with regard to how the sound has changed over time. (Proto-)Germanic <eu> developed towards *Altoberdeutsch* (the southern varieties of Old High German) <iu> in words such as *tief* 'deep' or *Fliege* 'fly' while varieties further north featured <oi> [20]. In Middle High German <iu> began changing into three spatially distributed variants in Switzerland: (i) a Northeastern variant <üü>, (ii) a Northwestern diphthongized variant <ie>, and (iii) a Southwestern group of variants which underlyingly trace back to <öü> (cf. Figure 3; [21]).

2. Methods

Section 2.1. describes the prediction and evaluation functionality of *DÄ* that enables analyses of language change; in 2.2. we describe the users and localities of the *DÄ* corpus, and 2.3. presents the distribution of *tief* variants as represented in the *Atlas* – the reference point for analyses of sound change.

2.1. Dialäkt Äpp & Procedures



DÄ's main function is the prediction of the user's dialect, which is based on 16 discriminative maps from the *Atlas*. The app prompts users to select their pronunciation variant from a list of each of the 16 variables by tapping on the screen. Given that SwG does not have a standard writing system, variants are spelled close to pronunciation – and feature additional IPA transcriptions where necessary. All variants are accompanied with sounds for users to listen to; see the prompt for *tief*, Figure 1.

Figure 1: *tief* and its dialectal variants for users to select.

When users arrive at the end of the quiz, the app presents a list of five localities – out of 550 adapted from the *Atlas* (16 / 566 original localities have merged, [11]) – that best corresponds to the user's dialect. Users are then asked to evaluate the predicted dialect (Figure 2, left). In case of an accurate prediction, they type in age and gender and send off the data anonymously (Figure 2, center left). In case of an incorrect prediction, they indicate their dialect by choosing from a list of localities (Figure 2, center right), which correspond to those used for the dialect prediction; users select age and gender and send off their information (Figure 2, right).

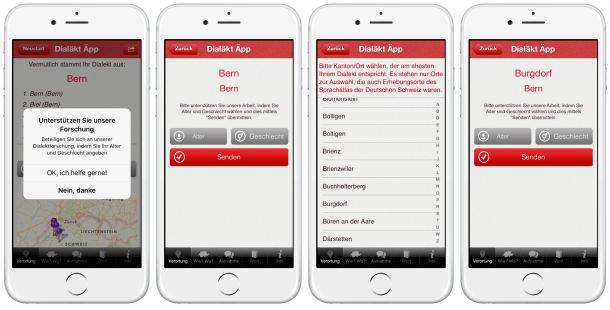


Figure 2: Evaluation of dialect prediction by users.

We then compare the users' values to those in the *Atlas*. *DÄ* predicted Bern for the fictive user of Figure 2 (center left). If s/he in fact speaks the Bern dialect, s/he would enter age and gender, and send off the data. The 16 variants for Bern as indicated in the *Atlas* are then compared to the values entered by the user. In the case shown here, it is likely that there is little discrepancy between the *Atlas'* and contemporary values, since *DÄ* predicted the correct locality. If, however, this user claimed to speak the dialect of Burgdorf, s/he would indicate this (as shown in Figure 2, center right and right) and send off the data. In this case we obtain a greater difference between the *Atlas'* and the contemporary data, which means that the dialect of this speaker from Burgdorf has become more like the dialect of Bern.

2.2. *DÄ* corpus

The corpus consists of data from 58,923 users from effectively all localities in the *Atlas*. Only three *Atlas* localities were not represented in the DÄ corpus: Mutten (Grisons), Obergoms (Valais), and Sternenberg (Zurich). For all other localities, there was at least one respondent, with a median of 48 respondents per locality. 42% of the users were females, 58% males. On average, users were 31.5 years old (MD=27; SD=15.5). 30% of the users were predicted in the correct locality, and 65% in the right canton [4].

2.3. Reference material

For the sake of showing results on analyses of sound change (cf. 3.2. & 3.3.), Figure 3 shows the *Atlas* variants.

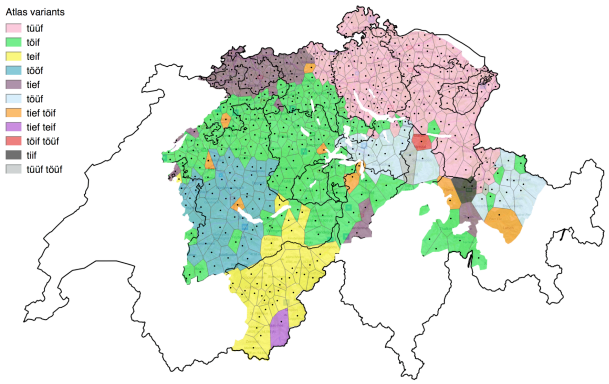


Figure 3: Variants of *Altoberdeutsch* <iu> shown in the *Atlas*.

Northeastern SwG had <üü>, Northwestern SwG <ie>, and in the multi-variant region (cf. 1.2) we find <öi> in the Western Midlands and Central Switzerland, and <öö> in much of the Southwest [21]. The *Atlas* further indicated large areas of <öü> in parts of Graubünden and in Central Switzerland. Fribourg and the Southwestern part of Bern are characterized by monopthongized <öö>. <ei> was reported in Southeastern Bern and in Valais. A pocket in Uri featured <ie>, which otherwise is dominant in the Northeast. Some of the variants shown in the *Atlas* were categorized for *DÄ* (e.g. <täüf> was included in <töüf>), and a number of localities demonstrated two variants (see Figure 5). Space prevents a comprehensive review of this categorization procedure; it was conducted using plausible historical linguistic rationales.

3. Results

3.1. Number of respondents

Figure 4 shows the number of respondents per locality, broken into ten natural classes (Jenks). We used *Voronoi* polygons for each locality (ten buffer) in Figures 3–6. Midland localities (e.g. Zurich N=3119, Bern N=2736, Basel N=1842) show the

highest number of respondents; Alpine localities, on the contrary, frequently feature 1–40 respondents only (e.g. Habkern (Bern) N=11, Betten (Valais) N=9, Sisikon (Uri) N=16).

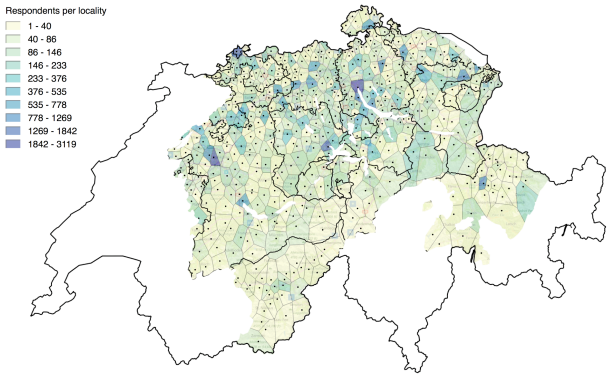


Figure 4: Number of respondents by locality.

3.2. Agreement with Atlas

Figure 5 shows the *DÄ–Atlas* agreement scores – 0 (red) showing no agreement, 1 (green) showing full agreement.

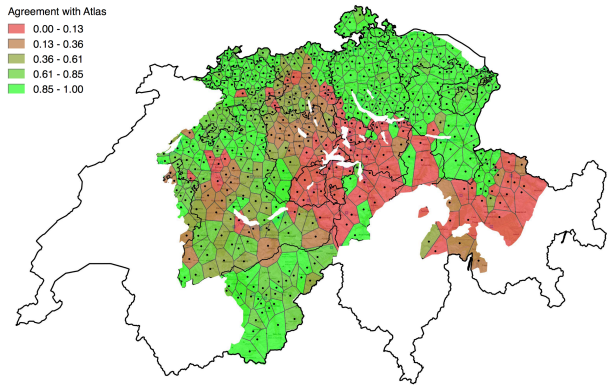


Figure 5: Agreement of *DÄ* variants with *Atlas* variants.

Much of Northeastern Switzerland reveals high agreement scores (green), e.g. the cantons of Zurich (cantonal mean M=.94), Thurgau (M=.96), St. Gallen (M=.92), as well as both Appenzell cantons (AI, M=.75; AR, M=.93). Much of Central Switzerland (e.g. Nidwalden, M=.015), some localities in the canton of Bern (M=.61), many localities of the cantons of Aargau (M=.54) and Graubünden (M=.35) reveal high disagreement scores.

3.3. *DÄ* variants

Figure 6 illustrates the variants indicated in *DÄ*. The broad geographical patterns attested in [20], [21], and in the *Atlas* (cf. Figure 3) remain largely intact: <üü> in the Northeast, <ie> in the Northwest, and a multi-variant region in the Southwest. The isoglosses of <ie> in the Northwest appear to have remained stable; <üü> has gained considerable terrain, spreading towards the Southwest, where it replaced <öü> in Graubünden, Glarus and Schwyz, and pushed aside <öi> in most of Aargau and Luzern. The geographical distribution of unrounded <ei>, mostly present in the Wallis, remained stable. Quite strikingly, <ie> – a typical feature of Basel German, but also, to a small extent, found in Uri (cf. Figure 3) – has diffused towards numerous Central Swiss localities, replacing <öi> in the Cantons of Uri, and Unterwalden. This phenomenon is also illustrated in Figure 5 where many of the Central Swiss localities are colored in red (indicating much disagreement with the *Atlas*). In Southern Bern <öi> replaced the monophthongized variant <öö>. For some localities, *DÄ* data included the same proportion of speakers for two different variants. <öi> and <öö>, for example, were equally reported in one locality nestled between the <öi> / <öö> isogloss, see Figure 6 (dark blue).

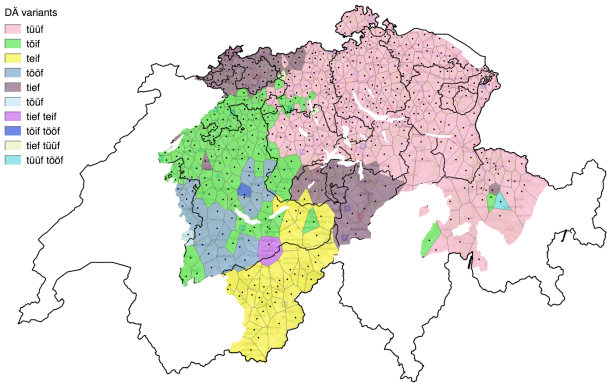


Figure 6: Variants of *Altoberdeutsch* <iu> as used in *DÄ*.

3.4. Change in apparent time

Figure 7 shows the *Atlas* agreement scores by age group. We divided the speakers into five natural breaks (Jenks) according to their age: 10–22, 23–32, 33–43, 44–57, 58–90. To test for an effect of *age*, we ran a GLM that included *sex*, *age* and *dialect* as factors (α=.05). A standard likelihood ratio test revealed a significant effect of *age* ($\chi^2(4)$=-481.07, p<.0001). Figure 7 shows that the oldest group (purple) used the variants indexed in the *Atlas* the most (high agreement), the youngest group (red) the least (low agreement).

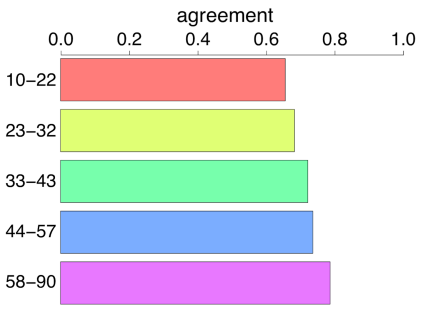


Figure 7: Proportion of *DÄ* speakers who indicated the same variant as indicated in the *Atlas*.

For the youngest age group, we found that 66% of answers were identical with the *Atlas* (red), followed by 68% (yellow), 72% (green), and 74% (blue). The oldest group (purple) demonstrated the highest *Atlas* agreement scores with 79% (purple).

4. Discussion

4.1. Regional variation and change in apparent time

A relatively recent theme of work in dialectology is leveling: the loss of minority dialectal variants and regional convergence towards majority features (cf. [6, 7, 8]). We find such leveling tendencies for *tief*: the Northeastern variant <üü> has spread considerably in southerly, westerly, and southeasterly direction – resulting in a convergence of traditional forms used in the 1950s towards the majority form <üü>. Reasons for this change are multifactorial and allow only for speculations on our part. The change in Central Switzerland may have to do with a substantial figure of the students from Schwyz and Glarus commuting to Zurich for training [23]. This spread may be evidence for an example of hierarchical diffusion, where the majority form spreads to increasingly smaller settlements [24]. Secondly, the diffusion of <üü> in southeasterly direction may be explained with residents from rural Grisons settlements commuting to work in Chur [25, 23], a city that serves as a transportation and cultural hub of the area and which – most importantly – was reported as using <üü> in the *Atlas* (see Figure 3). In addition to this potential diffusion from within the canton, Southeastern Switzerland is a popular summer and winter holiday destination for residents from the canton of Zurich [26]. Thirdly, we find a diffusion of <üü> in a westerly direction towards the canton of Aargau, pushing back local variants such <öi> and <ie> in Aarau, Lenzburg, Bremgarten, and Muri. Only the Bernese Aargau (Zofingen district) remained relatively stable in the southern periphery. Previous studies have shown that this canton in particular has proven to be in flux and in a zone of dialectal instability, nestled between the two major linguistic radii of Bern and Zurich [11].

Looking at the change of other variants, the monophthongized variant <öö> has lost substantial ground in Western German-speaking Switzerland and seems to be becoming replaced by <öi>, possibly under the influence of the linguistic radius of the urban regions of Bern (cf. [19]). What stands out synchronically, however, is the Bern city-specific variant <ie> in contemporary data, despite being surrounded by localities reporting <öi>. This Bern city-specific variant does not seem to spread to its urban regions. Another noticeable process of diffusion appears in Central Switzerland, where <ie> has diffused substantially towards both Ob- and Nidwalden, as well as to virtually all localities elicited in the canton of Uri. Presumably this is an example of contra-hierarchical diffusion (cf. [27]), where changes spread from the rural region – e.g. Andermatt (Uri) – to smaller towns and finally to larger towns (e.g. Sarnen, Stans).

The *age* effect reported in 3.4. fits in nicely with our intuitions on sound change in apparent time in SwG: older speakers show different speech patterns than younger speakers, which can be understood as an instance of sound change taking place in our sample, as the older speakers' variants agree with those reported in the *Atlas* more often. This trend may, however, also be an artifact of the large data set we are using (cf. Kilgarriff 2005).

4.2. Methodological caveats

There are methodological caveats that need to be kept in mind when we compare data that has been crowdsourced or collected with traditional dialectological methods (cf. [4]). In the *Atlas*, researchers elicited data directly from the speakers. For *DÄ*, there was no researcher present – giving subjects substantial freedom of interpreting the instructions given. For the *Atlas*, speaker selection criteria were stringent – as was typical for dialectology at the time [28]; in *DÄ*, users had different linguistic backgrounds, educational levels, and mobility habits. The two databases further differ in the distribution of speaker age: *Atlas'* subjects ranged between 51 and 80 [29]; in *DÄ* the median age is 27 [4]. Further factors that contribute to noise in the data are the users' self-declaration of dialects when evaluating the result, where users may have imitated a 'model', perhaps more nostalgic form of a dialect when doing so; potential multiple submissions [30], and potential biases stemming from the user interface (variants presented at the top of the drop-down menu may have been clicked more often than those at the bottom). In addition, users essentially perform a speech perception test when selecting a variant, as the interface allows them to listen to pre-canned recordings for all variants of *tief*. The degree to which this self-evaluation of phonetic realizations affects their choice of variants deserves scrutiny in future work. Despite this noise, previous research has shown that traditional dialectological methods reveal very similar diffusion patterns to those found through app crowdsourcing [4].

5. Conclusions

We presented a case study of how Big Data, crowdsourced through a smartphone app, can be used to study small-scale regional diachronic variation. From a methodological viewpoint, this dataset provides a novel way of studying language change due to the new sampling technique: dialectological methodology embodies a notion of the 'authentic' speaker; it has been biased towards population groups associated with maintaining the most distinctive regional varieties, i.e. NORMs [28, 29] or speakers of the 'vernacular' [30]. By changing data collection methods and giving up control over sampling, our approach avoids these biases. This approach is not meant to replace existing techniques for the collection of dialectological data, but simply wishes to highlight the power and added value of crowdsourced Big Data as a way of complementing established methods. Using this case study as a model, future studies using this corpus will reveal in greater detail which areas have undergone most change and which variants have spread or been replaced in the past 60–70 years.

6. Acknowledgements

We thank Daniel Wanitsch for server-side technical assistance and 65 backers who made *DÄ* possible through crowdfunding.

7. References

[1] Sprachatlas der deutschen Schweiz. (1962–2003). Bern: Francke
(Vols. 1–6), Basel: Francke (Vols. 7, 8).

[2] Leemann, A., & Kolly, M.J. (2013). Dialäkt Äpp. https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8 (accessed 30.06.2016).

[3] Kolly, M.-J., Leemann, A. (2015). Dialäkt Äpp: communicating dialectology to the public – crowdsourcing dialects from the public, in: Leemann, A., Kolly, M.-J., Schmid, S., Dellwo, V. (Eds.), Trends in Phonetics and Phonology. Studies from German-speaking Europe (pp. 271–285). Bern etc.: Lang.

[4] Leemann, A., Kolly, M.-J., Purves, R., Britain, D., Glaser, E. (2016). Crowdsourcing language change with smartphone apps. PLoS ONE 11/1: e0143060.
[5] Glaser, E. (2008). Der Wortschatz des Schweizerdeutschen. http://www.ds.uzh.ch/Forschung/Projekte/Schweizer_Dialekte/index.php (accessed 30.06.2016).
[6] Juska-Bacher, B. (2010). Wortgeographischer Wandel im Schweizerdeutschen. Sommersprossen, Küchenzwiebel und Schmetterling 70 Jahre nach dem SDS. Linguistik online, 42, 19–42.
[7] Christen, H., Glaser, E., Friedli, M. (Eds.). (2015). Kleiner Sprachatlas der Deutschen Schweiz. 6th ed. Huber: Frauenfeld.
[8] Glaser, E. (2016). Weiterführung der Online-Umfrage 2008. http://www.ksds.uzh.ch/de/onlineumfrage2008.html
[9] Glaser, E. (2014). Wandel und Variation in der Morphosyntax der schweizerdeutschen Dialekte. Taal en Tongval, 66(1), 21–64.
[10] Christen, H. (1998). Dialekt im Alltag: Eine empirische Untersuchung zur lokalen Komponente heutiger schweizerdeutscher Varietäten. Tübingen: Niemeyer.
[11] Siebenhaar, B. (2000). Sprachvariation, Sprachwandel und Einstellung. Der Dialekt der Stadt Aarau in der Labilitätszone zwischen Zürcher und Berner Mundartraum. Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte 108).
[12] Siebenhaar, B. (2002). Dialektwandel und Einstellung – Das Beispiel der Aarauer Stadtmundart, in: Berns, J., van Marle, J. (Eds.), Present-day Dialectology: Problems and Findings (pp. 313–332). Berlin, New York: Mouton de Gruyter (Trends in Linguistics 137).
[13] Haas, W. (1973). Zur l-Vokalisierung im westlichen Schweizerdeutschen, in: Bausinger, H. (Ed.), Dialekt als Sprachbarriere: Ergebnisbericht einer Tagung zur alemannischen Dialektforschung (pp. 63–70). Tübingen: Tübinger Vereinigung für Volkskunde.
[14] Christen, H. (1988). Sprachliche Variation in der deutschsprachigen Schweiz. Dargestellt am Beispiel der l-Vokalisierung in der Gemeinde Knutwil und in der Stadt Luzern. Stuttgart: Steiner (=Zeitschrift für Dialektologie und Linguistik. Beiheft 58).
[15] Piller, A. (1997). Sprachwandel im Sensebezirk dargestellt am Beispiel der /l/-Vokalisierung und der Rundung der Palatalvokale. Licentiate thesis, University of Fribourg.
[16] Christen, H. (2001). Ein Dialektmarker auf Erfolgskurs: Die /l/-Vokalisierung in der deutschsprachigen Schweiz. Zeitschrift für Dialektologie und Linguistik, 1, 16–26.
[17] Flury, A. (2002). l-Vokalisierung und nd-Velarisierung in Spiez: Eine empirische Untersuchung. Licenciate thesis, University of Bern.
[18] Matter, M., Ender, A. (2006). Datenerhebung mit einer Rapid Anonymous Study am Beispiel der l-Vokalisierung. Talk at 4. Tage der Schweizer Linguistik, 20 November 2006, Basel.
[19] Leemann, A., Kolly, M.-J., Werlen, I., Britain, D., Studer-Joho, D. (2014). The diffusion of /l/-vocalization in Swiss German. Language Variation and Change 26(2), 191–218.
[20] Wiesinger, P. (1970). Phonetisch-phonologische Untersuchungen zur Vokalentwicklung in den deutschen Dialekten. Vol. II: Die Diphthonge im Hochdeutschen. Berlin: de Gruyter.
[21] Hotzenköcherle, R. (1986). Dialektstrukturen im Wandel. Gesammelte Aufsätze zur Dialektologie der deutschen Schweiz und der Walsergebiete Oberitaliens. Aarau, Frankfurt am Main, Salzburg: Sauerländer.
[22] BFS=Bundesamt für Statistik (2016). Amtliches Gemeindeverzeichnis der Schweiz. www.bfs.admin.ch (accessed 30.06.2016).
[23] BFS=Bundesamt für Statistik (2015). Pendlermobilität der Schweiz 2013. Neuchâtel.
[24] Bailey, G., Wikle, T., Tillery, J., Sand, L. (1993). Some patterns of linguistic diffusion. Language variation and change, 5(3), 359–390.
[25] Amt für Raumentwicklung Graubünden (2007). Siedlungsbericht Graubünden. Analyse der Siedlungsentwicklung seit 1980. Chur.
[26] Amt für Raumentwicklung Graubünden (2012). Zweitwohnungen in Graubünden. Canobbio.
[27] Wikle, T. (1997). The spatial diffusion of linguistic features in Oklahoma. Proceedings of the Oklahoma Academy of Science 77, 1–15.
[28] Chambers, J. K., Trudgill, P. (1998). Dialectology. Cambridge University Press: Cambridge.
[29] Hotzenköcherle, R. (1984). Die Sprachlandschaften der deutschen Schweiz. Edited by Bigler, N., Schläpfer, R. Aarau: Sauerländer.
[30] Birnbaum, M. H. (2004). Human research and data collection via the Internet. Psychology, 55(1), 803.
[29] Bucholtz, M. (2003). Sociolinguistic nostalgia and the authentication of identity. Journal of Sociolinguistics, 7, 398–416.
[30] Eckert, P. (2003). Elephants in the room. Journal of Sociolinguistics, 7, 392–397.

(Mor-)phonotactic consonant clusters in Standard Austrian German and Standard German German

Hannah Leykum[1], *Sylvia Moosmüller*[1]

[1] Acoustics Research Institute, Austrian Academy of Sciences, Austria

hannah.leykum@oeaw.ac.at, sylvia.moosmueller@oeaw.ac.at

Abstract

Consonant clusters occur within morphemes (phonotactic clusters) as well as across morpheme boundaries (morphonotactic clusters). Since morphonotactic clusters contain morphological information, differences between the two types of clusters are expected in speech production. Previous studies on language acquisition, speech processing, and computer simulations proved a different treatment of these types of clusters.

Our previous analyses of Standard Austrian German (SAG) speakers showed no significant difference in the production of the two types of clusters. We interpreted this as a result of specific timing relations in SAG, which might impede a different treatment of the two types of clusters. In order to prove this hypothesis, we compared word-final phonotactic and morphonotactic consonant clusters in homophonous word-pairs produced by speakers of SAG, speakers of Standard German German living in Germany (SGG), and speakers of Standard German German living in Austria (SGGA).

The analyses revealed that, as expected, the speakers of SAG did not differentiate between the two types of clusters. Whereas, the speakers of SGG and SGGA as well, did no differentiation between phonotactic and morphonotactic clusters in speech production. Therefore, the hypothesis on an influence of the prosody could not be confirmed.

Index Terms: morphonotactics, phonotactics, consonant clusters, Standard Austrian German, Standard German German

1. Introduction

Morphonotactic consonant clusters are specified by a morpheme boundary within the consonant cluster, whereas phonotactic consonant clusters are defined as clusters occurring within a morpheme. In the current study, consonant clusters which exist both within a morpheme as well as across morpheme boundaries are investigated.

The morphological function of a consonant cluster is assumed to facilitate the processing of morphonotactic consonant clusters [1]. Speech processing experiments showed that morphonotactic clusters are detected faster in a visual word recognition task compared to homophonous phonotactic clusters [1]. A cluster modification task revealed that a morpheme boundary could have a positive impact on the processing of morphonotactic consonant clusters compared to the processing of homophonous phonotactic consonant clusters [2]. However, in an auditory cluster detection task, phonotactic clusters were favoured [2].

A computer simulation has shown that the cognitive representation of morphonotactic clusters differs from the representation of phonotactic clusters [3]. Investigations on first language acquisition show mixed results. Some studies revealed that children learn to produce morphonotactic clusters prior to phonotactic cluster [4, 5], others concluded that children learn both types of clusters at the same time [6], and still others found that only purely or prevailing morphonotactic clusters are acquired prior to phonotactic clusters [7]. In speech production, some results on word-final coronal stop deletion [8, 9] indicate that morphonotactic clusters are less susceptible to deletions and reductions, whereas other investigations point to the high importance of the phonological context [10, 11], and the frequency and predictability of a word [12, 13] on stop deletions.

As shown, several investigations pointed out that in language acquisition, speech processing and computer simulation differences between the two types of clusters could exist. Therefore, an extension of the Strong Morphonotactic Hypothesis [14] predicts that these differences also exist in speech production: Morphonotactic clusters are expected to be less susceptible to deletion and reduction processes than phonotactic clusters.

In previous investigations on consonant clusters produced by SAG speakers [15–17], no differences between phonotactic and morphonotactic clusters emerged. We interpreted this result as a consequence of specific timing characteristics of SAG. These imply that SAG, in contrast to SGG, rather aligns with quantifying languages by imposing constraints on the relative duration of vowel + consonant sequences [18, 19]. Thus, SAG constitutes a prosodically mixed type to be found between true quantifying languages and word languages, as defined by Auer [20], whereas SGG has, in the course of its history, changed into a true word language [21]. For this reason, we concluded that the temporal pattern might be disturbed by distinguishing between phonotactic and morphonotactic clusters.

In order to test this hypothesis, we compared morphonotactic and phonotactic consonant clusters produced by SAG speakers with the production of SGG and SGGA speakers. It is hypothesised that SAG speakers will show no differences between morphonotactic and phonotactic consonant clusters, whereas differences are expected in the production of SGG speakers. SGGA speakers act as a control group. Whenever differences between SAG and SGG speakers occur, the direction SGGA speakers will take can help interpret the results.

2. Material and method

2.1. Subjects

Recordings of 16 speakers aged between 20 – 25 years were conducted. All subjects were either speakers of Standard

Austrian German (SAG) or speakers of Standard German German (SGG). Half of the speakers of SGG have been living in Vienna (Austria) since at least one year (hereinafter SGGA). The SAG speakers (as defined by [22]) were all born and raised in Vienna, with at least one parent born and raised in Vienna as well. The subjects were students and at least one parent has an academic education or both parents have a qualification for university entrance. The SGG and SGGA speakers were defined as speakers born and raised in the northern part of Germany (north of the Benrath line), showing no dialectal influence. Within each group, subjects were balanced for gender.

2.2. Material and procedure

Four different word-final consonant clusters consisting of two or four consonants were selected. For these consonant clusters, homophonous word-pairs exist, of which the consonant cluster either has a phonotactic or morphonotactic status. This results in 12 monosyllabic German words (nouns and conjugated verbs in the present tense) as target words. The selected clusters and the corresponding target words are presented in Table 1.

Table 1: *Consonant clusters and target words*

	phonotactic	morphonotactic
[st]	Hast *(hurry)*	hasst *(you hate)*
		hasst *(he/she hates)*
	Mist *(dung, rubbish)*	misst *(you measure)*
		misst *(he/she measures)*
[ft]	Schaft *(stem)*	schafft *(he/she creates, works)*
[xt]	Macht *(power)*	macht *(he/she makes)*
[ŋkst]	Hengst *(stallion)*	hängst *(you hang)*

The target words were embedded in sentences in a post-focal position. To ensure high comparability, carrier phrases of the following structure were designed:

Zu mir? - Ich habe zu Peter „die Hast" gesagt, glaube ich.

(To me? - I said to Peter "the hurry", I think.)

By structuring the sentences this way, the focus is on the name 'Peter', whereas our target word ("Hast") is in a post-focal position. The sentences with the target words were part of a larger set of sentences. The 16 participants were asked to read all sentences twice. This resulted in a total of 384 target words.

The recordings of the speakers were segmented, annotated, and transcribed manually. Thereafter, measurements and semi-automatic extraction of the following acoustic parameters were carried out: relative duration and intensity (RMS amplitude) of the words, clusters, individual consonants of the clusters, and phonemes surrounding the clusters.

In order to eliminate influences such as the grammatical category of a word [23], word frequency [24], or individual differences in speaking rate, we calculated the relative duration of the clusters in relation to the duration of the word. The articulation rate (in syllables per second) was calculated in dividing the number of syllables by the sum of the durations of the target word and the word preceding and following the target word. Pauses were not included in the calculation. Log transformed word frequency values (extracted from: http://wortschatz.uni-leipzig.de/) were included as a control variable in the analyses to account for a possible impact of the word frequency on the articulation. For words containing a morphonotactic cluster, the word frequency of the conjugated verb was extracted.

The measurements of the parameters were analysed statistically with R by using mixed effect models (with subject and word as random factors). Where necessary, Tukey post-hoc tests with p-value adjustment were carried out.

Erroneously, some target words were stressed or followed by a pause. Theses clusters were nevertheless included in the analysis. They functioned as additional control variables.

As concerns the cluster /st/ (see Table 1), the morphonotactic target words are homophonous conjugated verbs in the 2nd and 3rd person singular, present tense. They were separated in order to verify potential differences between these two forms. However, the statistical analysis revealed no significant differences in duration or intensity of the two forms. Therefore, the two forms are not distinguished in the subsequent analyses.

Also, the two groups of SGG-speakers were tested for differences between the groups. Whenever no differences turned out, the two groups were lumped together.

Significant main effects concerning differences between the pairs of homophonous words will not be reported in the following section.

3. Results

3.1. Duration of the consonant clusters

Concerning the relative duration (in % of word duration) of the consonant clusters, a mixed effect model including random effects for "subject" and "word" was fit. The analysis revealed neither an influence of the variety, nor an influence of the type of cluster on the relative duration of the cluster. However, a significant interaction between gender and word-pair emerged ($p<0.05$) (see Figure 1).

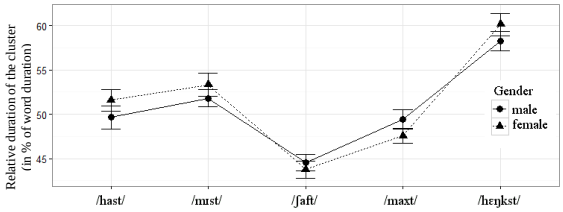


Figure 1: *Gender*word-pair interaction*

Regarding the differences between speakers of SAG, SGG and SGGA, no significant results concerning the relative and absolute duration of the cluster, the word and the vowel preceding the cluster emerged. However, we found a significant vowel*variety interaction ($p<0.05$) in the relative duration of the vowel preceding the cluster, revealing longer vowels for the SAG speakers for the vowels /ɪ/ and /a/, but not for /ɛ/.

The statistical analysis of the articulation rate revealed no significant influences of gender or variety of the speaker.

3.2. Intensity of the consonant clusters

The relative intensity of the cluster was calculated by comparing the absolute intensity (RMS) with the intensity of the vowel preceding the cluster to normalise the intensity values. A mixed effect model including random effects for

"subject" and "word" revealed a significant interaction between word-pair and variety ($p<0.01$, see Figure 2). However, the type of cluster had no influence on the relative intensity of the cluster.

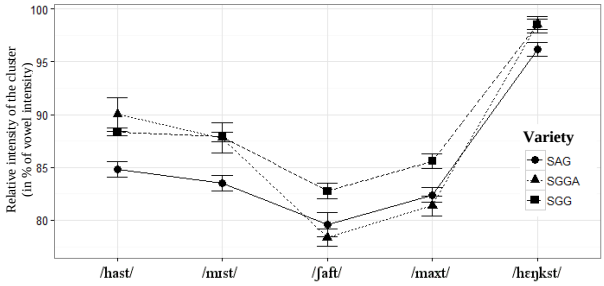


Figure 2: *Interaction between word-pair and variety*

3.3. Final /t/-deletion and reduction

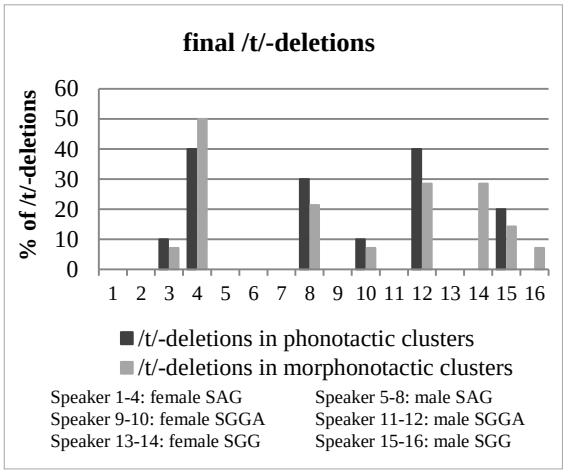


Figure 3: *Deletions of the word-final /t/*

In the realisations of the consonant clusters, several deletions of the cluster-final stop occurred. A large inter-speaker variability could be observed: half of the speakers never deleted the final /t/, whereas others showed up to 50% of deletions (see Figure 3). The statistical analysis revealed neither a significant interaction between the occurrence of deletions and the type of cluster, nor between deletions and variety, nor between deletions and gender, nor a higher order interaction between these variables. At least, the articulation rate was significantly higher ($p<0.001$) when the final stop was deleted compared to fully realised clusters.

In duration and intensity of the final consonant (when realised), no differences occurred in terms of the type of the cluster, the variety and/or gender of the subjects.

The absolute and relative duration of the fricative preceding the word-final stop was significantly longer in clusters with an acoustically deleted final /t/ ($p<0.001$), but the absolute and relative duration of the clusters with deleted /t/ were significantly shorter compared to clusters with a realised final consonant. However, neither with regard to the distinction between phonotactic and morphonotactic clusters nor between varieties, significant effects emerged.

4. Discussion

The aim of the present study was to examine whether in speech production, the impact of prosody impedes a differentiation between phonotactic and morphonotactic consonant clusters in SAG. Therefore, recordings of speakers of SAG were compared with speakers of SGG living in Germany and Austria. The hypothesis stated that, on the one hand, SAG speakers will produce no differences between phonotactic and morphonotactic clusters, because of the higher importance of the prosody in SAG, whereas on the other hand, we expected that these differences will occur in the realisations of SGG (and SGGA) speakers, since SGG is a true word language. The investigated data corroborate the first part of the hypothesis; speakers of SAG did not differentiate between the two types of clusters. The second part of the hypothesis, however, could not be confirmed, since, in the same way as SAG speakers, speakers of SGG/SGGA did not differentiate the two types of clusters. Neither in the relative duration, nor in the relative intensity, an interaction between variety and type of cluster could be found.

No significant variety-specific or type-of-cluster-specific differences emerged neither in the reduction and deletion of word-final /t/ nor in the duration of the reduced vs. non-reduced clusters. In addition, regarding the number of deletions of the final stop, no differences between speakers of SAG and speakers of SGG, male and female speakers or between the deletions in phonotactic or morphonotactic clusters occurred.

With regard to the duration and intensity of the clusters, no significant differences concerning the type of cluster emerged. However, the statistic evaluation revealed a significant variety*word-pair interaction for the relative intensity of the clusters. As this type of interaction is not in the focus of the current study, the interesting result (see Figure 2) should be further investigated in future research.

Moreover, the revealed gender*word-pair interaction in the relative duration of the cluster as well as the vowel*variety interaction in the relative duration of the vowel preceding the cluster deserve further investigation within studies on e.g. gender or variety differences.

Since no interaction between the variety and the type of cluster could be found in none of the investigated parameters, our hypothesis on a possible influence of prosody on cluster productions could not be confirmed.

The design of our study was highly controlled; therefore, the phonological context could not have influenced the results, since it was held constant by using homophone target words embedded in carrier phrases. Moreover, the articulation rate, as well as the insertion of pauses or the erroneous stress of the target word were included in the analysis as control variables. Likewise, the influence of word frequency and the grammatical category was reduced by calculating the relative duration of the clusters. However, comparing words of different absolute durations is problematic anyway. The calculation of the relative duration of the clusters reduces the impact of these effects. Yet, differences in the absolute duration of the words are possibly also affecting the relative duration of parts of the words.

A further possibly interfering variable is the predictability of a speech segment on deletion and reduction processes [12, 13, 25–27]. The redundancy of the information contained in the cluster is higher in the morphonotactic clusters, because

the verbs containing the clusters were directly preceded by the subject pronoun, making the information concerning the morpheme boundary highly predictable. Nevertheless, by none of the speaker groups, the morphonotactic clusters were reduced to a higher extent than the phonotactic clusters. Yet, with the investigated material, it is not possible to preclude opposing processes eliminating each other.

Since in our previous investigations no differences between the two types of clusters emerged for speakers of SAG, and since in the present study, these results were confirmed and could be extended to speakers of SGG, the question arises, whether the results could be explained by the morphological richness of a language. Therefore, further languages have to be investigated.

In future research, it is planned to investigate additional clusters and words, also non-homophonous words, embedded in carrier phrases as well as elicited in more natural speech within a semi-spontaneous speaking task. Moreover, the investigation will be extended to German and French consonant clusters in word-medial position.

Acknowledgements

The current investigation was performed within the project I 1394-G23 “Human Behavior and Machine Simulation in the Processing of (Mor)Phonotactics”, funded by the FWF and the project “Die österreichische Standardaussprache Wiens in Kontakt mit der deutschen Standardaussprache“, funded by Kultur Wien.

References

[1] K. Korecky-Kröll, W. U. Dressler, E. M. Freiberger, E. Reinisch, K. Mörth, and G. Libben, “Morphonotactic and phonotactic processing in German-speaking adults,” *Language Sciences*, vol. 46, pp. 48–58, 2014.

[2] C. Celata, K. Korecky-Kröll, I. Ricci, and W. U. Dressler, “Phonotactic processing and morpheme boundaries: word-final /Cst/ clusters in German,” *Italian Journal of Linguistics*, vol. 27, no. 1, pp. 85–110, 2015.

[3] B. Calderone, C. Celata, K. Korecky-Kröll, and W. U. Dressler, “A computational approach to morphonotactics: Evidence from German,” *Language Sciences*, vol. 46, pp. 59–70, 2014.

[4] L. Kamandulyté, “The Acquisition of Morphonotactics in Lithuanian,” *Wiener Linguistische Gazette*, vol. 73, pp. 88–96, 2006.

[5] P. Zydorowicz, “Polish morphonotactics in first language acquisition,” *Wiener Linguistische Gazette*, vol. 74, pp. 22–44, 2007.

[6] E. Freiberger, “Morphonotaktik im Erstspracherwerb des Deutschen,” *Wiener Linguistische Gazette*, vol. 74, pp. 1–23, 2007.

[7] L. Kamandulyté-Merfeldienė, “Morphonotactics in L1 acquisition of Lithuanian: TD vs. SLI,” *Eesti Rakenduslingvistika Ühingu aastaraamat - Estonian Papers in Applied Linguistics*, vol. 11, p. 95, 2015.

[8] G. R. Guy, “Explanation in variable phonology: An exponential model of morphological constraints,” *Language Variation and Change*, vol. 3, pp. 1–22, 1991.

[9] G. R. Guy, “Form and Function in Linguistic Variation,” in *Towards a Social Science of Language: Papers in honor of William Labov. Volume 1: Variation and change in language and society*, G. R. Guy, C. Feagin, D. Schiffrin, and J. Baugh, Eds, Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 221–252, 1996.

[10] B. Schuppler, W. van Dommelen, J. Koreman, and M. Ernestus, “Word-Final [t]-Deletion: An Analysis on the Segmental and Sub-Segmental Level,” *Proceedings of the 10th annual conference of International Speech Communication Association (INTERSPEECH)*, pp. 2275–2278, 2009.

[11] F. Zimmerer, M. Scharinger, and H. Reetz, “When BEAT becomes HOUSE: Factors of word final /t/-deletion in German,” *Speech Communication*, vol. 53, no. 6, pp. 941–954, 2011.

[12] B. Schuppler, W. A. van Dommelen, J. Koreman, and M. Ernestus, “How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level,” *Journal of Phonetics*, vol. 40, no. 4, pp. 595–607, 2012.

[13] R. J. J. H. van Son and L. C. W. Pols, “Information structure and efficiency in speech production,” *Proceedings of Eurospeech 2003*, pp. 769–772, 2003.

[14] W. U. Dressler and K. Dziubalska-Kołaczyk, “Proposing Morphonotactics,” *Italian Journal of Linguistics*, vol. 18, no. 2, pp. 249–266, 2006.

[15] H. Leykum, S. Moosmüller, and W. U. Dressler, “Homophonous Phonotactic and Morphonotactic Consonant Clusters in Word-Final Position,” *Proceedings of the 16th annual conference of International Speech Communication Association (INTERSPEECH)*, pp. 1685–1689, 2015.

[16] H. Leykum, S. Moosmüller, and W. U. Dressler, “Word-final (mor-)phonotactic consonant clusters in Standard Austrian German,” *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, 2015.

[17] H. Leykum and S. Moosmüller, “Poster: Das (mor-)phonotaktische Konsonantencluster /st/ in wortmedialer und wortfinaler Position in homophonen Wortpaaren,” *11. Tagung Phonetik und Phonologie im deutschsprachigen Raum*. Marburg, Oct. 2015.

[18] R. Bannert, *Mittelbairische Phonologie auf akustischer und perzeptorischer Grundlage*. Lund: Gleerup, 1976.

[19] S. Moosmüller and J. Brandstätter, “Phonotactic information in the temporal organization of Standard Austrian German and the Viennese dialect,” *Language Sciences*, vol. 46, pp. 84–95, 2014.

[20] P. Auer, “Silben- und akzentzählende Sprachen,” in *Language typology and language universals: An international handbook*, M. Haspelmath, E. König, W. Oesterreicher, and W. Raible, Eds, Berlin, New York: W. de Gruyter, 2001.

[21] R. Szczepaniak, *Der phonologisch-typologische Wandel des Deutschen von einer Silben- zu einer Wortsprache*. Berlin: Walter de Gruyter, 2007.

[22] S. Moosmüller, *Hochsprache und Dialekt in Österreich: Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck*. Wien: Böhlau, 1991.

[23] J. M. Sorensen, W. E. Cooper, and J. M. Paccia, “Speech timing of grammatical categories,” *Cognition*, vol. 6, pp. 135–153, 1978.

[24] M. Pluymaekers, M. Ernestus, and R. H. Baayen, “Lexical frequency and acoustic reduction in spoken Dutch,” *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2561–2569, 2005.

[25] M. Pluymaekers, M. Ernestus, and R. H. Baayen, “Articulatory planning is continuous and sensitive to informational redundancy,” *Phonetica*, vol. 62, no. 2-4, pp. 146–159, 2005.

[26] M. Pluymaekers, M. Ernestus, R. H. Baayen, and G. Booij, “Morphological effects on fine phonetic detail: The case of Dutch -igheid,” in *Laboratory phonology 10*, C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée, Eds, Berlin, New York: Mouton de Gruyter, pp. 511–532, 2010.

[27] M. E. Iris Hanique, “Final /t/ reduction in Dutch past-participles: the role of word predictability and morphological decomposability,” *Proceedings of the 12th annual conference of International Speech Communication Association (INTERSPEECH)*, pp. 2849–2852, 2011.

[28] I. Hanique and M. Ernestus, “The Role of Morphology in Acoustic Reduction,” *Lingue e linguaggio*, vol. 2, pp. 147–164, 2012.

The interplay of prominence and boundary strength: a comparative study

Katalin Mády[1], *Felicitas Kleber*[2], *Uwe D. Reichel*[1], *Ádám Szalontai*[1]

[1]Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
[2]Ludwig-Maximilians-Universität Munich, Germany
{mady.katalin|uwe.reichel|szalontai.adam}@nytud.mta.hu, kleber@phonetik.uni-muenchen.de

Abstract

Hungarian is a language with left-headed head-/edge-prominence. Our goal was to investigate if prominence in Hungarian can be increased by inserting or strengthening phrase boundaries before emphasised words. German, being a right-headed head-prominence language was the basis for the comparison. Since prominence marking in Hungarian is highly dependent on syntax, a list of fruits differing in size was used. Participants were asked to utter fruit names so that someone else can guess if a fruit was small or large. We hypothesised that Hungarian speakers would use boundary signals *preceding* a large fruit, whereas Germans would either insert boundaries *after* a large fruit or not make use of final lengthening and pauses. Results show that Hungarians use more pauses than Germans in all positions, and the occurrence of pauses is used to enhance prominence. While pre-boundary lengthening was only observed preceding a large fruit in Hungarian, it was present for speakers of both languages in the final syllable of the large fruit itself. Pause occurrences after a large fruit did not depend on fruit size in any of the languages.

Index Terms: prosodic boundary, prosodic prominence, final lengthening, Hungarian, German

1. Introduction

Prosodic prominence can be marked in various ways depending on the prosodic typology of the language: it can be marked by prominence cues on the head or phrase, by the edge of a phrase, or by a combination of the two [1]. Head-prominence languages can have lexically distinctive tones (Mandarin, Swedish) or lexical stress (German, English). Edge-prominence languages, on the other hand, lack word- or phrase-level heads and use phrase boundary signals to mark prominence (Korean, certain Japanese and Mongolian dialects). In head/edge-prominence languages, prominence is marked by both the head and the edge (French, Japanese). These languages have postlexical (i.e. lexically non-distinctive) stress and accentual phrases (AP) with a uniform tonal pattern (rising, falling or rising-falling) [2].

Hungarian is a language with left-headed prosody both on the word and the sentence level. Word-level stress is fixed to the word-initial syllable and is thus fully predictable. Sentences that contain a narrow or contrastive focus are structured according to their information structure, and the strongest pitch accent falls to the left edge of the logical predicate of the sentence [3]. However, prosodic headedness is difficult to apply to broad focus sentences, since accent strength is described to be roughly equivalent throughout the intonation phrase [4, p. 131].

Recent studies on Hungarian have shown that the left-headed structure is also present in accentual phrases of Hungarian. Their tonal pattern is falling and can be described as H* La [5]. Another characteristics of languages with APs is that two adjacent content words forming a single syntactic unit tend to form one AP, whereas longer or more complex syntactic phrases contain more APs [2]. Evidence for this was found by [6]. Regressive voicing assimilation that applies over word boundaries in Hungarian was found to be weaker over AP boundaries by [7]. Unlike in Korean, pitch accents in Hungarian cannot be predicted based on the tonal pattern of the AP they initiate, thus the language shows a head/edge-prominence marking pattern.

In an experiment in which contrastive emphasis was elicited in read dialogues, it was observed that Hungarian speakers often inserted pauses before emphasised words [8]. This strategy might be used by speakers to enhance the prominence of a word by inserting a prosodic boundary before it, since the initial word of an AP automatically receives prominence due to the language's left-headed structure. In this study,

In the present study, the realisation of emphasis in Hungarian is compared to German that is prosodically right-headed on the IP level, but has no APs [9]. It is hypothesised that Hungarian speakers use edge-marking cues *before* an emphasised target to enhance its prominence, whereas Germans primarily make use of head prominence and potentially of the right-headed structure by inserting boundaries *after* a large fruit.

Hungarian is a so-called discourse-configurational language in which word order is highly dependent on pragmatic factors such as information structure [3]. This means that prominence is primarily expressed by syntax. It has been shown that prosodic cues do not play a crucial role of prominence production and perception in syntactically well-formed sentences [10]. Thus, in order to investigate prosodic cues of prominence marking, it was essential to create material that does not contain syntactic information.

2. Materials and methods

Participants saw two baskets containing altogether five fruits on a computer screen. The fruits and their order were identical throughout the experiment, whereas the size of the baskets (3+2 vs. 2+3) and the size of the fruits (small vs. large) varied. Participants were asked to name fruits and to indicate (1) whether the fruits are small or large and (2) whether the basket includes two or three fruits. The experiment was preceded by a training phase. Participants were shown a small and a large fruit with their names written in lower case and in capital letters, and they were asked to indicate the difference. They were not provided by any auditory material to avoid a bias due to priming. The training phase included a familiarisation session with the sequence of the fruits in order to make the naming task as fluent as possible.

Stimuli in the Hungarian material contained the fruits

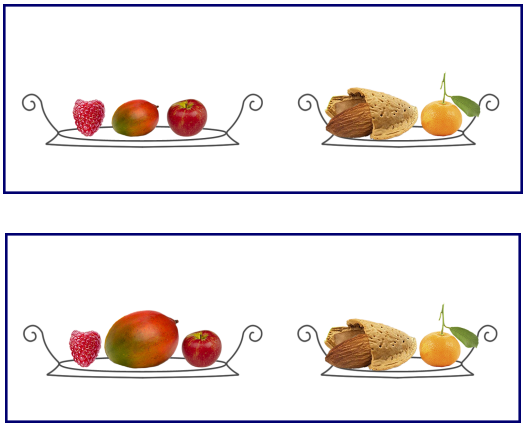


Figure 1: *Examples of fruit sequences in Hungarian. Analysis was based on the first unit, i.e. the first basket.*

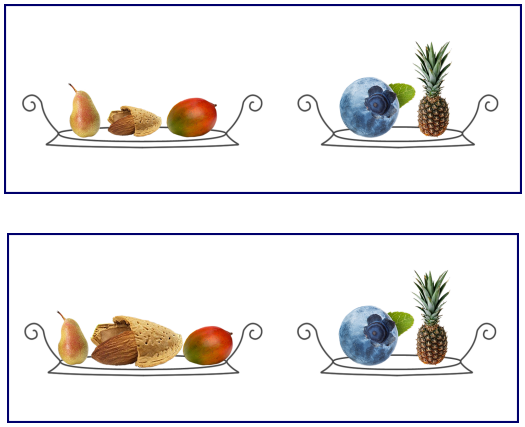


Figure 2: *Examples of fruit sequences in German. Analysis was based on the first unit, i.e. the first basket.*

málna mangó alma mandula mandarin 'raspberry mango apple almond mandarin'. German stimuli were *Birne Mandel Mango Blaubeere Ananas* 'pear almond mango blueberry ananas'. Stress was initial in all words.

Analysis was based on sequences in which the first basket contained three fruits which were either [small small small] or [small LARGE small] (the size of fruits in the second basket being [small LARGE] or [LARGE small], but the latter did not undergo further analysis), see Figure 2. The following parameters were used for analysis:

1. Pause **occurrence** before and after the second fruit.
2. Pause **duration** before and after the second fruit (if present).
3. Lengthening: duration of final syllable **before** the second fruit.
4. Lengthening: duration of the final syllable **in** the second fruit.
5. Accent: duration of the **initial** stressed syllable of the second fruit.

Speech samples were recorded in a sound-proof room with 10 Hungarian native speakers in Budapest and with 8 German native speakers in Munich, all being female students. The data set contained 432 realisations (4 sequences $\times$ 6 repetitions $\times$ 18 speakers). Linear mixed effect models were computed for each language separately with the size of the second fruit in the first basket as fixed effect, participant, repetition and the size of the fourth and fifth fruit in the second basket as random effects. χ^2 tests were used for the distribution of pauses if applicable.

The following hypotheses were tested:

- Hypothesis 1: Boundaries are likely to occur before each accented word in Hungarian.
- Hypothesis 2: Hungarian speakers mark the prominence of a large second fruit by a **preceding** prosodic boundary.
- Hypothesis 3: German speakers will mark a large second fruit either by a boundary **following** it, or by no boundary.

3. Results

3.1. Occurrence and duration of pauses

All fruits carried a pitch accent in both languages irrespectably of their size. Thus, the target of the subsequent analysis is not to compare accentuation with deaccentuation, but higher emphasis with lower prominence.

Table 1 presents the number of pauses before and after the second fruit in the two languages.

Table 1: *This is an example of a table.*

	before 2nd fruit		**after 2nd fruit**	
	small	**big**	**small**	**big**
Hungarian	59	92	83	103
German	0	7	5	34

Hungarian speakers produced a substantially higher number of pauses in all positions which is interpreted as a consequence of the presence of lower-level, i.e. AP boundaries before each pitch-accented fruit. According to χ^2 tests, the occurrence of pauses in the production of Hungarian speakers was significantly more frequent before a large second fruit ($\chi^2 = 7.21, df = 1, p = 0.007$), but not after it (($\chi^2 = 2.15, df = 1, p = 0.14$). At the same time, German speakers produced significantly more pauses after a large second fruit (($\chi^2 = 21.56, df = 1, p < 0.001$). (Due to the overall low number of pauses produced by German speakers before the second fruit, the test was not applicable.)

Pause durations in Hungarian showed the expected tendency before the second fruit being significantly longer if it was large ($t = 9.2, p < 0.001$), but the same tendency was found for the right boundary following a large fruit ($t = 10.7, p < 0.001$). Pause durations did not differ for German speakers in either position.

3.2. Pre-boundary lengthening

The last syllable of the first fruit (that was small throughout the experiment) showed substantial lengthening preceding a large fruit in Hungarian, ($t = 4.16, p < 0.001$), but not in German ($t = 1.4, p < 0.16$), see Figure 3. However, the final syllable of the second fruit was lengthened in both languages due to a larger

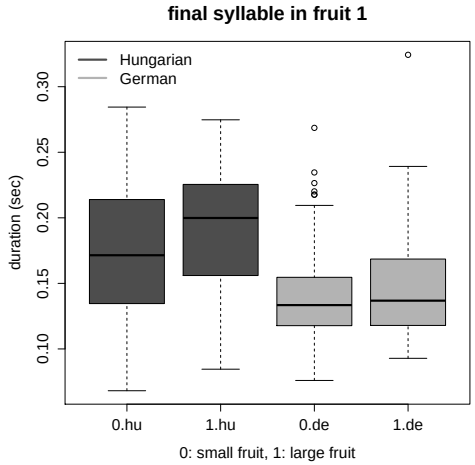


Figure 3: *Duration of the final syllable preceding the emphasised word and a potential pause.*

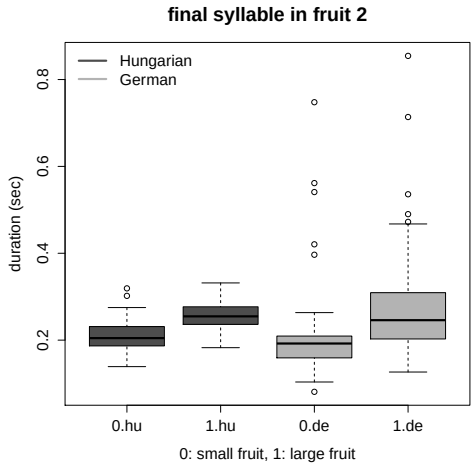


Figure 4: *Duration of the final syllable preceding the emphasised word and a potential pause.*

emphasis on the fruit as a unit ($t = 12.5$ for Hungarian, $t = 6.0$ for German, both $p < 0.001$), see Figure 4. The interpretation of the latter findings is problematic, because final lengthening *within* an emphasised word can both signalise head prominence, i.e. a carryover effect of the previous stressed syllable [11] and a boundary effect due to the following pause.

The assumption of a carryover effect is based on [11, 12] who found evidence that accentual lengthening is not limited just to the syllable carrying the word stress but also effects adjacent syllables. Thus the locus of domain-head and -edge effects [13], the stretch of speech over which the effects are manifested, can span more than one syllable.

A domain-head effect in terms of longer duration on the stressed syllable of the large second fruit was found in both languages ($t = 11.2$ for Hungarian, $t = 12.5$ for German, both $p < 0.001$), see Figure 5.

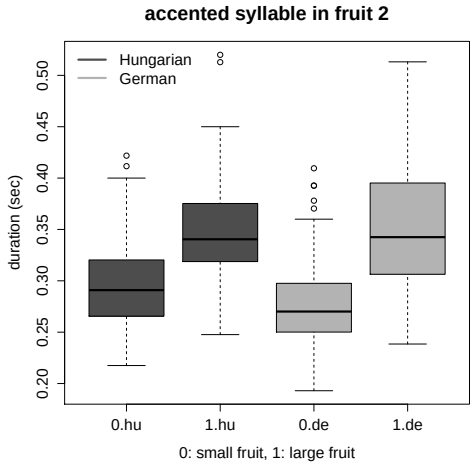


Figure 5: *Duration of the stressed syllable of the emphasised word.*

4. Discussion and conclusions

The results show that Hungarian and German speakers utilise different prosodic means to mark emphasis if syntax cannot be used. The frequent occurrence of pauses between any two fruits in the production of Hungarian speakers signalises that accented words are preceded by lower-level boundaries, which is not the case in German (evidence for Hypothesis 1). Both the more frequent occurrence of pauses and their longer durations at the left edge of the emphasised word provide evidence that higher prominence is connected with stronger boundaries in Hungarian (evidence for Hypothesis 2). German speakers did not make use of boundary strength on the left edge.

Although the utilisation of boundaries for prominence marking is characteristic of edge-prominence and head/edge-prominence languages that contain accentual phrases, it is not clear if the prominence-effect of the boundary is in fact cued by the parameters investigated here. Left-headed APs in Korean and Japanese were not found to be preceded by pauses at all, and no consistent lengthening of the preceding final syllable was found [14]. Thus it could be argued that boundaries between the first and the second fruit in Hungarian do not demarcate AP, but an IP boundary. However, the amount of lengthening in the final syllable of the third fruit (being the last one in the first basket) is larger, thus it cannot be excluded that Hungarian differs from Korean and utilises higher-level boundary markers to mark lower-level boundaries.

It is not clear how lengthening of the final syllable of large fruits can be interpreted in the two languages. Longer durations of non-stressed syllables of accented words have been shown before [15]. At the same time, they can alternatively or parallelly signalise pre-boundary lengthening following the second word. Since pauses were frequent after the second fruit in both languages, this possibility cannot be excluded at present.

As was mentioned in the Introduction, complex syntactic units tend to form several APs in Hungarian. In fact, adverb+adjective+noun sequences investigated in [6] tended to be divided in two APs rather than to bear a single pitch accent. Thus, it is possible that Hungarian speakers tend to split longer sequences into more accentual phrases – this would explain the overall higher number of pauses between the second and the

third fruit.

Based on the findings, it can be concluded that in Hungarian, boundaries at the left edge of a prominent word (here: the large fruit), both in terms of pauses and their durations and of pre-boundary lengthening are utilised for prominence strengthening. In German, lower-level boundaries do not seem to play a role in prominence marking.

5. Acknowledgements

This work was funded by the researchers' exchange grant *Form and function of prosodic structure in Hungarian and in German* between IPS LMU Munich and RIL HAS Budapest funded by DAAD (Germany) and MÖB (Hungary). Many thanks to the further participants of the projects: Susanne Beinrucker, Andrea Deme, Kristóf Galla, Anna Kohári, Nele Salveste, Balázs Surányi for data processing and discussions.

6. Bibliographie

[1] S.-A. Jun, "Prosodic typology: by prominence type, word prosody, and macro-rhythm," in *Prosodic Typology II: the new development in the phonology of intonation and phrasing*. Oxford: University Press, 2014, pp. 520–539.

[2] S.-A. Jun and J. Fletcher, "Methodology of studying intonation: From data collection to data analysis," in *Prosodic Typology II: the new development in the phonology of intonation and phrasing*. Oxford: University Press, 2014, pp. 493–519.

[3] K. É. Kiss, *The syntax of Hungarian*. Cambridge: Cambridge University Press, 2002.

[4] L. Varga, *Intonation and stress: evidence from Hungarian*. Basingstoke & New York: Palgrave Macmillan, 2002.

[5] Š. Beňuš, U. D. Reichel, and K. Mády, "Modeling accentual phrase intonation in Slovak and Hungarian," in *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014.*, L. Veselovská and M. Janebová, Eds. Olomouc: Palacký University, 2014, pp. 677–689.

[6] K. Mády, A. Szalontai, A. Deme, and B. Surányi, "On the interdepencende of prosodic phrasing and prosodic prominence in Hungarian," in *Proc. 11th International Conference on the Structure of Hungarian*, Piliscsaba, Hungary, 2013.

[7] K. Mády and Z. Bárkanyi, "Voicing assimilation at accentual phrase boundaries in Hungarian," in *Proc. ICPhS 2015*, Glasgow, 2015, p. ICPHS0796.

[8] K. Mády and F. Kleber, "Variation of pitch accent patterns in Hungarian," in *Proc. 5th Speech Prosody Conference, Chicago*, 2010, pp. 100 924:1–4.

[9] M. Grice, S. Baumann, and R. Benzmüller, "German intonation in autosegmental-metrical phonology," in *Prosodic typology*, S.-A. Jun, Ed. Oxford: Oxford University Press, 2005, pp. 55–83.

[10] K. Mády, "Prosodic (non-)realisation of broad, narrow and contrastive focus in Hungarian: a production and a perception study," in *Proc. Interspeech 2015*, Dresden, 2015, pp. 948–952.

[11] A. Turk and L. White, "Structural influences on accentual lengthening in English," *J. of Phonetics*, pp. 171–206, 1999.

[12] L. White and A. Turk, "English words on the Procrustean bed: polysyllabic shortening reconsidered," *J. of Phonetics*, vol. 38, pp. 459–471, 2010.

[13] L. White, "Communicative function and prosodic form in speech timing," *Speech Communication*, vol. 63–64, pp. 38–54, 2014.

[14] S.-A. Jun and C. Fougeron, "A phonological model of French intonation," in *Intonation: analysis, modeling and technology*, A. Botinis, Ed. Dordrecht: Kluwer, 2000, pp. 209–242.

[15] M. Beckman and J. Edwards, "Articulatory evidence for differentiating stress categories," in *Papers in Laboratory Phonology 3*, P. A. Keating, Ed. Cambridge: Cambridge U. P., 1994, pp. 7–33.

How to distinguish between self- and other-directed *wh*-questions?

Katalin Mády, Uwe D. Reichel

Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
{mady|uwe.reichel}@nytud.mta.hu

Abstract

The most general aim of *wh*-questions is to seek for information, but they can have a wide range of other pragmatic functions. In this paper we investigate self-directed questions in dialogues that are lexicalised forms of vacillation ("how should I explain?") and do not directly address the interlocutor. Their prosodic properties are compared with real *wh*-questions that seek for information.

Index Terms: *wh*-questions, self-directed speech, prosody, stylisation, Adaboost

1. Introduction

According to [1] most languages have three basic sentence types: declarative, interrogative, and imperative. For the interrogative type [1, p 160] point out as a first approximation that it "elicits a verbal response from the addressee. It is used principally to gain information". In accordance to Searle's question analysis [2] a question is an attempt to elicit information from the addressee the speaker wants to gain. [1] and many other researchers (e.g. [3, 4]) give numerous counter-examples suggesting a more fine-grained subdivision of interrogatives. Among these counter-examples are self-directed ("self-addressed" [3]) questions by which the speaker does not expect information from an addressee but is rather thinking aloud. Self-directed questions can be marked syntactically, e.g. in German by verb-last word order [4] (*"ob das wohl stimmt"?* – 'whether it is true?').

According to [3], self-directed questions do not request an answer, instead, they express the status of the speaker. In the example *"Now why did I say that?"* the speaker verbalises her surprise about her own utterance. A different view is provided by [5] who investigate self-directed queries in connection with disfluency signals. They claim that in this case, the speaker is the "addressee" of the query, since he/she can straightforwardly answer their own question.

In this paper *wh*-questions that act as verbalised vacillation are investigated. In these cases, speakers use self-directed questions to gain time to collect their thoughts and find a better way to explain something to their partner. As opposed to real questions, these questions do not aim at encouraging the interlocutor to be cooperative, instead, they can be characterised as offtalk. The goal of the paper is to compare prosodic features of real *wh*-questions that seek for information and require cooperativeness from the partner to self-directed questions that primarily signalise vacillation on the speaker's side. Based on Ohala's frequency code concept [6] we expect higher energy, higher f0 level and range values as well as more pronounced local f0 shapes for the interlocutor-directed than for the self-directed questions.

Figure 1: Image of the objects that appear on the screens of the two players. Left: screen of the describer with the raspberry blinking, right: screen of the follower who is supposed to place the raspberry into the position explained by the first player. Instruction left: Describe the position of the blinking object*, right:* Drag the object into the correct position.

2. Data

2.1. Corpus

Data are taken from the Hungarian version of the object game of the Columbia Game Corpus [7]. It is a computer-aided game with two participants. Participants use separate laptops, and they do not have visual contact with each other. The players see objects on their screen that are identical except for one object that is blinking on the screen of one player, while it is located in the lower part of the screen of the other player. The first player describes the position of the blinking object in relation to the other objects that are placed on the screen of the second player in the same position. The second player is supposed to place the object in exactly the same position. Participants get a score after each turn on a 0 to 100 scale. Their roles alternate in the course of the game, so that both speakers are describers in half of the altogether 14 turns. Figure 1 shows the objects from a turn as were shown on the the two screens.

In the Hungarian version of the game, players formed 4 triplet groups, and they played two games with partly different, partly identical objects with both other members of the group (A with B, B with C, C with A). They were payed for their participation. Additionally, the group that scored highest was promised additional payment, in order to enhance the accuracy of the descriptions. Participants within a group were familiar with each other (relatives or close friends), which lead to a high degree of naturalness during the task.

The corpus is currently being annotated among others for dialog acts. The current version of the paper presents first results on self- and other-directed *wh*-questions that were manually segmented and labelled. All interrogatives began with a *wh*-word that carries an accent in Hungarian as a default. Self-directed questions did not differ from other-directed questions

in their syntax. One self-directed question contained a lexical unit that would be improbable in a real question: "And this is located between the traffic light and the standard lamp. In addition, how is it located?". Another self-directed question expressed that the describer has difficulties to express himself: "Ow, how should I tell you?" The lexical form of the remaining self-directed questions was identical with potential string-identical other-directed questions.

2.2. F0 extraction and preprocessing

Fundamental frequency (f0) was extracted by autocorrelation (Praat 5.3, sample rate 100 Hz, [8]). Voiceless utterance parts and f0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky-Golay filtering [9] using third order polynomials in 5 sample windows and transformed to semitones relative to a base value. This base value was set to the f0 median below the 5th percentile of the speaker's f0 within the entire dialog and served to normalize f0 with respect to its overall level.

3. Prosody stylisation

3.1. Parameterisation

Within the utterance chunks three types of features were extracted: (1) f0 register features, (2) local f0 movements on the *wh*-word, and (3) energy. The features are listed in table 1. As register features we measure the f0 level and range starting points, trends and mean values. For this purpose a base- mid- and topline were fitted to the chunk as illustrated in the left half of Figure 2. As described in greater detail in [10] this fitting method does not depend on fuzzy f0 peak and valley detection but consists of three linear regressions through local f0 median values in the lower, mid and upper f0 range. As shown in [10] this method therefore is less error prone and more robust against the influence of local pitch events. The level trend within the chunk is defined as the midline slope. The range trend is defined as the slope of the regression through the pointwise distances between top- and baseline. These linear level and range stylisations are shown in the right half of Figure 2.

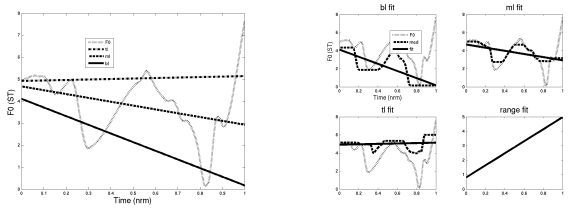


Figure 2: **A (left):** *Stylisation of base-, mid- and topline based on F0 median sequences below the 10th percentile for the baseline, above the 90th percentile for the topline and for all values for the midline. The F0 range is represented by a regression line fitted through the pointwise distances between the base- and topline.* **B (right):** *Base-, mid-, topline and linear range stylisation results.*

Next to the global f0 register variables we parameterized the local f0 movement of the stressed first syllable on the always chunk-initial wh-word by a third-order polynomial. The 30 ms window was placed on the vowel midpoint, the left half limited by the chunk onset. Within that window time was normalized from -1 to 1 with 0 placed on the vowel midpoint. Figure 3 shows the decomposition of a local f0 movement by a third order polynomial.

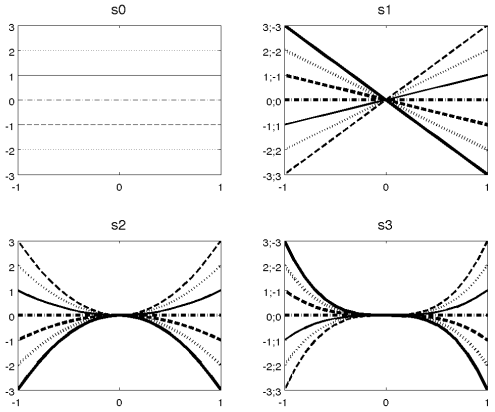


Figure 3: *Influence of each coefficient of the third order polynomial $t = \sum_i s_i \cdot t^i$ on the contour shape. All other coefficients set to 0. For the purpose of compactness both function and coefficient values are shown on the y-axis if they differ.*

Finally, energy was measured by RMS over the entire speech chunk.

3.2. Feature weights

Table 1 summarizes the examined features and their discriminatory power to hold apart self- and other-directed questions. These weights w for features i are derived from the Silhouette measure usually used for cluster validation as follows:

$$w(i) = \frac{\frac{\sum_{j=1}^{n} S(j)}{n} + 1}{2},$$

where the silhouette $S(j)$ measures for each of the n data points – i.e. for a feature vector – j how well it can be assigned to one of the classes *self-* and *other-directed*. More precisely

$$S(j) = \frac{d_B(j) - d_A(j)}{\max(d_A(j), d_B(j))}.$$

$d_A(j)$ stands for the mean squared Euclidean distance between vector j and other vectors of the same class. $d_B(j)$ stands for the mean distance between vector j and vectors of the other class. Adding 1 and dividing by 2 transposes the weight range to the interval $[-1\ 1]$.

4. Results

Figure 4 shows the values of the examined parameters for other- and self-directed questions. A visual inspection reveals that the difference between these question types is primarily quantitatively but not qualitatively expressed, i.e., there is a difference in the absolute values but not in the algebraic sign. As an example, for both other- and self-directed questions there is a falling local f0 movement on the accented syllable (negative c_1) which is more pronounced in other-directed speech. Generally absolute values are higher in other-directed speech indicating a more pronounced usage of intonation. 2-sided Wilcoxon tests on the absolute values reveal significant differences for *ml_slope, ml_icpt,*

Table 1: *Prosodic features and their weights in terms of mean Silhouette normalized to sum 1. Weights were calculated for absolute feature values.*

Feature	Description	Weight
Register		
ml_slope	f0 midline slope	0.5905
ml_icpt	f0 midline intercept	0.5781
ml_mean	f0 midline mean	0.5650
rng_slope	f0 range slope	0.3981
rng_icpt	f0 range intercept	0.5895
rng_rms	f0 range RMS	0.5718
Pitch accent		
c_3	cubic polynomial coefficient	0.4182
c_2	quadratic polynomial coef	0.6074
c_1	linear polynomial coef	0.4783
c_0	offset from midline	0.4511
Energy		
en	signal RMS over chunk	0.6437

rng_rms, c_3, c_2 ($p < 0.05$), and *en*, and tendencies for *rng_icpt*, c_1, and c_0 ($p < 0.1$). For *ml_mean* and *rng_slope* no significant differences were found. However, the boxplots suggest that additional data will move the differences towards significance.

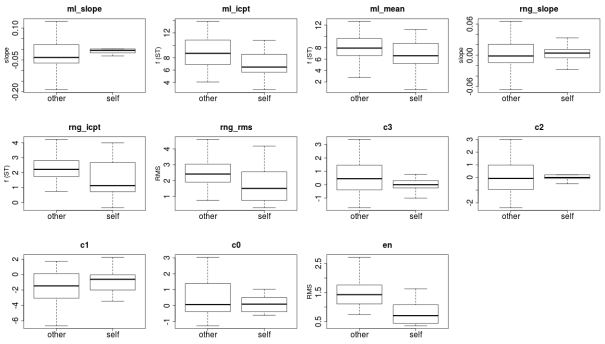


Figure 4: *Prosodic parameter values in other- and self-directed* wh*-questions. ml: f0 level, rng: f0 range, c: polynomial coefficients, en: energy*

5. Detection

We used the tree ensemble classifier AdaBoost M1 [11] designed for two-class problems. By brute force optimisation on a small development set the ensemble learner parameters were set as follows: number of learners: 100, maximum number of decision splits: 5, minimum number of observations at a leaf: 5, minimum number of observations at a non-terminal node: 10. The preliminary results of a tenfold cross-validation are presented in table 2. At the current state accuracy amounts to 87% and is expected to rise with additional training data.

6. Discussion

We found clear evidence for prosodic differences in self- and other-directed questions. Overall prosody is more expressive in other-directed than in self-directed speech: f0 level and range as well as energy are higher, and local f0 movements are more pronounced. This is reflected in overall higher absolute values of all examined features in other-directed questions.

Table 2: *10-fold cross-validation. Mean accuracy, weighted recall, precision, and F1 score (and yes: the F1 score can indeed be below precision and recall).*

Accuracy	0.87
Weighted Recall	0.87
Weighted Precision	0.94
Weighted F1 score	0.86
Kappa	0.65

The weights in Table 1 show that energy (*en*) and the sharpness of the local f0 movement (c_2) deviating from the midline on the question word are most influential in marking self- and other-directedness.

All differences are gradual and quantitative, not qualitative. To give examples, for both conditions there is f0 declination (negative *ml_slope*), which is flatter in self-directed speech. In both conditions there are concave as well as convex local f0 shapes on the question word (negative and positive c_2), but again the shape is less pronounced in self-directed questions (much less variation around 0).

These differences in expressiveness are in line with the findings of [12] who compared linguistic and prosodic features in on- and offtalk. Since [12] examined human-machine communication, they partly attributed this difference in expressiveness to the artefact that humans tend to hyperarticulate when talking to machines which therefore enlarges the differences between other- (here: computer) and self-directed speech. However, our data suggests that these differences also hold for human–human communication. They might be actively used by the speaker to signal whether or not a question is information-seeking and requires a reaction by the interlocutor.

Finally, automatic question type prediction based on the extracted features yields high accuracies which will be beneficial for more general offtalk detection for dialog systems.

7. Acknowledgments

The work of the first author was funded by OTKA K 115922 and by an institute partnership programme of the Alexander von Humboldt Foundation. The second author is financed by a fellowship of the Alexander von Humboldt Foundation.

8. References

[1] J. Sadock and A. Zwicky, "Speech act distinctions in syntax," in *Language Typology and Syntactic Description I: Clause Structure*. Cambridge: CUP, 1985, pp. 155–96.

[2] J. Searle, *Speech Acts*. CUP, 1969.

[3] D. Wilson and D. Sperber, "Mood and the analysis of non-declarative sentences," in *Human Agency*, J. Dancy, J. Moravcsik, and T. C., Eds. Stanford, CA: Stanford University Press, 1988, pp. 77–101.

[4] H. Truckenbrodt, "Zur Strukturbedeutung von Interrogativsätzen," in *Linguistische Berichte*. Hamburg: Helmut Buske Verlag, 2004, vol. 199, pp. 313–350.

[5] J. Ginzburg, R. Fernández, and D. Schlangen, "Self-addressed questions in disfluencies," in *Proc. 6th Workshop on Disfluency in Spontaneous Speech*, Stockholm, 2013, pp. 33–36.

[6] J. Ohala, "The frequency code underlies the sound symbolic use of voice pitch," in *Sound Symbolism*. Cambridge: Cambridge University Press, 1994.

[7] A. Gravano, v. Beňuš, H. Chávez, J. Hirschberg, and L. Wilcox, "On the role of context and prosody in the interpretation of 'okay'," in *Proc. 45th Annual Meeting of Association of Computacional Linguistics*, Prague, 2007, pp. 800–807.

[8] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Tech. Rep., 1999, 132–182.

[9] A. Savitzky and M. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[10] U. Reichel and K. Mády, "Comparing parameterizations of pitch register and its disconti nuities at prosodic boundaries for Hungarian," in *Proc. Interspeech 2014*, Singapore, 2014, pp. 111–115.

[11] Y. Freund and R. Schapire, "A short introduction to boosting," *J. Japanese Society for Artificial Intelligence*, no. 5, pp. 771–780, 1999.

[12] A. Batliner, C. Hacker, and E. Nöth, "To talk or not to talk with a computer – Taking into account the users focus of attention," *J. Multimodal User Interfaces*, vol. 2, pp. 171–186, 2008.

Perception of Pitch Scaling in Rising Intonation
On the Relevance of f0 Median and Speaking Rate in German

Jan Michalsky

Institute of German Studies, University of Oldenburg, Germany

j.michalsky@uni-oldenburg.de

Abstract

Although phonetic in nature, pitch scaling is described to assume certain linguistic functions such as marking of focus or sentence mode [1, 2, 3, 4, 5]. Accordingly, it is linguistically relevant to understand how precisely pitch scaling is perceived. Recent evidence suggests that the excursion size of f0 in semitones might not be a stable cue across speakers and registers [5]. As an alternative, De Looze and Hirst [6] propose to relate f0 measurements to the speaker's median. Two perception experiments on the evaluation of scaling of final rises in German were conducted to investigate which measurement is best suited to represent the perception of f0 movements. The results indicate that the distance between a specific f0 measurement and the speaker's median in semitones is the most stable parameter in describing pitch perception within speakers. On the other hand, no measurement solely based on f0 succeeded in explaining the observed differences in perception across speakers. A post-hoc analysis showed that speaking rate might be a non-intonational feature, which influences expectations about a given speaker's natural pitch range. Accordingly, we propose a measurement that incorporates speaking rate to adequately describe f0 movements with respect to pitch perception.

Index Terms: pitch perception, pitch scaling, f0 measurement, octave median scale, speaking rate, question intonation

1. Introduction

Traditionally, pitch scaling has been categorized as part of the paralinguistic component of intonation and thus as being primarily restricted to the signalling of attitudinal meanings [7]. Several studies have shown that this distinction is not a categorical one. A study by Ladd and Morton [1] suggests that languages might use gradual features like pitch scaling of accent peaks to distinguish broad and contrastive focus. Among other languages, this has been found for German as well [8, 9, 10]. Furthermore, preliminary results by Kügler [11] suggest that this marking of focus through pitch scaling is not restricted to high tones but affects f0 excursion sizes in general for example by lowering the onset of final rises under contrastive focus in German. This observation is supported by Gussenhoven and Rietveld [12] for Dutch. Additionally, Michalsky [2, 3, 4] showed that speakers of German use scaling of final rising contours to distinguish incomplete statements from questions in read speech as well as in spontaneous speech. A more recent study even suggests that speakers scale two separate rising parts of twice rising contours to signal focus and questioning [13]. This evidence suggests that the meaning of pitch scaling is not restricted to attitudinal features but may include linguistic meaning as well (cf. [14]), making it all the more important to understand the mechanisms of fine grained pitch perception in intonation.

While literature on the perception of categorical distinctions in intonation, intonational meaning and contour choice grows steadily, research on the precise perception of continuous acoustic features of intonation, such as f0 scaling or tonal alignment, is surprisingly scarce. Vaissière [15] argues that general psychoacoustic findings on perception of single acoustic events cannot easily be generalized and mapped on the perception of intonation. Since intonation involves f0 movements, its perception involves not only psychoacoustic mechanisms but higher cognitive and linguistic processes as well. This peculiarity of intonation has been captured in the two traditional schools of intonational modelling, the British School and the Autosegmental-Metrical (AM) approach. While the British School describes the meaningful parts of intonation via configurations, which resemble movement patterns [16, 17], the AM theory represents intonation via tonal targets on two distinct levels (high and low), which are characterized in relation to one another and thus also represent a direction of movement rather than absolute levels [18]. The latter is an improvement of older models working with more than two pitch levels [19, 20]. It excludes virtually all aspects of gradual pitch scaling and thereby avoids the necessity for a reliable definition for the classification of a tone as extra high.

In fact, there are traces of a more fine-grained distinction of pitch levels in both models. While the British School distinguishes between *low fall* and *high fall* or *low rise* and *high rise* [16, 17] the AM approach uses phonological representations for *downstep* and even *upstep* of tonal targets in some models [21, 22]. Although *downstep* and *upstep* are defined in immediate relationship to preceding tones, an open question remains nonetheless at which point a high tone is merely subject to declination or sufficiently deep to be objectively classified as *downstep*. As shown above this problem cannot be solved by excluding it from the model since aspects of pitch scaling might be categorical and/or linguistic and thus need to be incorporated into a phonological model.

Since fundamental frequency as the primary acoustic cue to pitch is not perceived in a linear fashion, f0 is generally transformed to a psychoacoustic scale such as Mel, Bark or ERB or to a logarithmic semitone scale. Additionally, speakers show a wide variety of different register levels primarily but not exclusively due to individual and sex differences. Accordingly, we would expect pitch scaling not to be evaluated on an absolute level such as the distance between an f0 target and a fixed reference line but rather as the distance between f0 targets within the same utterance. Focusing on the perception of nuclear final rises, this assumption seems to be contradicted by the findings of Michalsky [5]. Investigating differences in the perception of continuation intonation and question intonation

in German, it is shown that listeners apparently rely to a higher degree on the absolute value of the rise offset than on the relative size of the excursion measured as the distance between rise onset and rise offset.

As an alternative for representing f0 movements with respect to pitch perception De Looze and Hirst [6] propose the *Octave Median Scale*, in which the speaker's f0 median is assumed to be the center of his/her natural pitch range with boundaries half an octave above and below. Accordingly, the distance from a certain f0 point to the median of the utterance might prove as a stable measurement for pitch perception. Although Nolan [23] provides evidence that the semitone scale might be the most reliable of the scales proposed so far, De Looze and Hirst [6] criticize the semitone unit for being an artificial construct without any immediate connection to human perception. They reject the segmentation into semitones and measure the distance between an f0 point and the median relative to an octave.

Based on the work of Michalsky [2, 3, 4, 5] it can be assumed that the boundary between continuation rise and question rise coincides with a relatively constant value of perceived height or size in the final pitch excursion for a specific listener. Accordingly, a scale that represents pitch perception adequately should give a consistent value for rise excursions that are perceived equally across speaker and register variation. From this assumption we derived three hypotheses for the paper at hand. Firstly, final rises of different speakers with different registers should be perceived as equally high when the relative excursion size is kept constant. Secondly, speaker evaluation should remain the same when the register level is varied but the relative excursion size kept constant. Thirdly, different speakers speaking with the same register level and excursion sizes should be evaluated equally.

Conclusively, three potential measurements for measuring f0 scaling with respect to an adequate auditory representation are investigated: 1) the absolute value of the rise offset to a fixed reference line, 2) the distance between rise offset and rise onset in semitones, and 3) the distance between rise offset and speaker median. Additionally, the last scale could be measured in semitones [23] or in *ome* [6].

Two perception experiments were conducted to investigate which of the proposed scales suits best to represent f0 movement with respect to pitch perception if any.

2. Method

2.1. Material

For both experiments three different female speakers read a single sentence with declarative syntax and exclusively sonorous segments in the nuclear region. All sentences were realized with a low-rising nuclear contour L*H H% (ToBI: L*H-^H% [21]). Realizations with a final rise of roughly seven semitones and thus in the center of the desired target stimuli were selected for each speaker. The stimuli were resynthesized using PSOLA via the Praat package for audio editing [24]. For the first experiment the onsets of the final rises were fixed to the natural baseline of the three speakers at 165 Hz, 178.5 Hz, and 197 Hz. The utterance onset was fixed by creating the same declination slope for all three speakers as a straight declining line with all prenuclear accents deleted. The high nuclear accent tone was fixed at a distance of four semitones from the rise onset. Lastly, the rise offset was varied in steps of one semitone and set to distances from five to ten semitones relative to the rise onset (s. Figure 1).

For the second experiment the speakers with the lower and the higher register were synthetically shifted to the f0 values of the medial speaker differing only in the height of the utterance onset by keeping the declination constant across different utterance lengths. For each experiment and speaker ten repetitions were included yielding a total of 180 stimuli per experiment and 360 on the whole. All stimuli were randomized and concatenated with an inter-stimulus interval of four seconds.

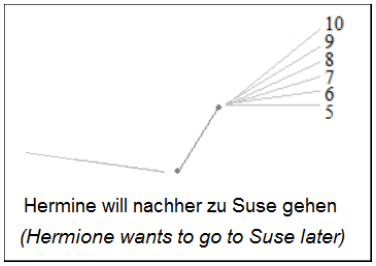


Figure 1: Stylized test sentence for one speaker with tonal annotation and the 6 levels for the final rise offset.

2.2. Subjects

Both experiments were conducted with the same group of 20 female students from the University of Oldenburg. All subjects were between 18 and 30 years old and monolingual speakers of German.

2.3. Procedure

Following the example of an identification task [25, 5] subjects received a questionnaire with a written presentation of the target sentence. On the questionnaire, the two tested categories were primed by displaying each sentence as a question, orthographically marked by the use of a question mark, and an incomplete statement, conveyed via a possible continuation of the sentence. The subjects listened to the stimuli via headphones (Sure SR2) in a sound-booth at the University of Oldenburg and were instructed to decide via two-way forced choice if the perceived stimulus was a question or an incomplete statement.

2.4. Statistical analysis

For the statistical analysis we conducted linear mixed effects models using R [26], the lme4-package [27] as well as the lmerTest-package [28]. Model fit was determined using maximum likelihood ratio tests. P-values were calculated using Satterthwaite-approximation. As fixed effects we used *speaker* and *experiment* and as a random factor we used *subject*. As the dependent variable we chose *crossing point* representing the relative excursion size, where question judgements first reached the 50 % mark measured on each of the three scales. This point was extrapolated via linear regression for every individual subject when reached between two stimuli.

3. Results

3.1. The Semitone Onset Scale

The Semitone Onset Scale measures the relative rise excursion as the distance between the offset of the final rising movement and its onset in semitones. The results show that the 50 % crossing point in statement/question-evaluation of all three

speakers differed significantly in both experiments (Experiment 1: speaker1=8.13, speaker2=7.69, speaker3=7.08, *b*=-0.53, *SE*=0.12, *df*=40.00, *t*=-4.24, ***p*<.001**; Experiment 2: speaker1=7.12, speaker2=7.39, speaker3=7.95, *b*=0.41, *SE*=0.09, *df*=40.00, *t*=4.35, ***p*<.001**) (s. Figure 2). Additionally, the comparison of experiment 1 and 2 shows that listeners evaluated the first speaker as having a later crossing point and thus relatively smaller excursions than speaker 2 in experiment 1 but having greater excursions in experiment 2 thus suggesting an unexplainable shift in pitch perception.

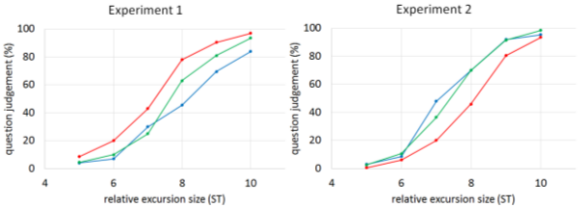


Figure 2: Evaluation of final rises measured on Semitone Onset Scale (blue=speaker 1, green=speaker 2, red=speaker 3)

This suggests that the same relative f0 excursion is evaluated differently not only across speakers but also within speakers with varying register levels. Accordingly, as can be seen in Figure 3, speaker 1 and 3 both differed in their *crossing point*, which again reached statistical significance (Speaker1: exp1=8.13, exp2=7.12, *b*=-1.01, *SE*=0.19, *df*=20.00, *t*=-5.20, ***p*<.001**, Speaker3: exp1=7.08, exp2=7.95, *b*=0.87, *SE*=0.20, *df*=20.00, *t*=4.43, ***p*<.001**).

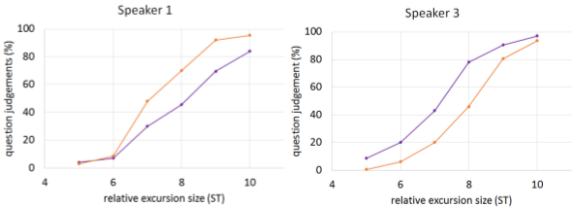


Figure 3: Evaluation of final rises measured on Semitone Onset Scale (purple=experiment 1, orange=experiment 2)

In conclusion, the distance from rise onset to rise offset in semitones does not serve as a reliable measurement for perceived pitch. Furthermore, since absolute rise offset values were identical for all three speakers, the differences between all three speakers found for experiment 2 rejects the possibility of an absolute reference line.

3.2. The Octave Median Scale

To investigate the perceptual relevance of the Octave Median Scale we transformed the original stimulus categories defined by the relative distance from rise offset to rise onset to the distance from the rise offset to the speaker's median. Additionally, comparable to the procedure used by De Looze and Hirst [6] we did not measure this distance in semitones but in *ome* using the octave as a unit. The same problems as presented above for the Semitone Onset Scale still occur. All three speakers were evaluated significantly different in both experiments but only scarcely so in the second experiment (Experiment 1: speaker1=0.46, speaker2=0.40, speaker3=0.34, *b*=-0.06, *SE*=0.01, *df*=40.00, *t*=-5.90, ***p*<.001**; Experiment 2: speaker1=0.38, speaker2=0.37, speaker3=0.42, *b*=0.02, *SE*=0.01, *df*=40.00, *t*=2.04, ***p*<.05**) (s. Figure 4). In addition, the direction of perceived excursion height between speakers changes again without meaningful explanation across experiments. Furthermore, the distance of half an octave to the median did not coincide with the *crossing point* for any speaker in either experiment, making it presumably irrelevant for distinguishing questions and statements.

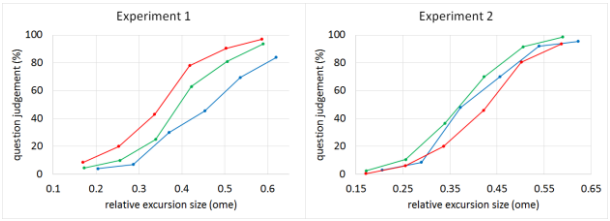


Figure 4: Evaluation of final rises measured on Octave Median Scale (blue=speaker 1, green=speaker 2, red=speaker 3)

In line with what has been shown for the previous scale, the OMe Scale fails to represent within speaker variation across register levels. Rise excursions for speaker 1 and 3 were again evaluated significantly different across experiments (Speaker1: exp1=0.46, exp2=0.38, *b*=-0.08, *SE*=0.02, *df*=20.00, *t*=-5.08, ***p*<.001**, Speaker3: exp1=0.34, exp2=0.42, *b*=0.07, *SE*=0.02, *df*=20.00, *t*=4.51, ***p*<.001**) (s. Figure 5). This pattern is roughly the same regardless whether the median is calculated from the whole utterance or from the part preceding the final rise only.

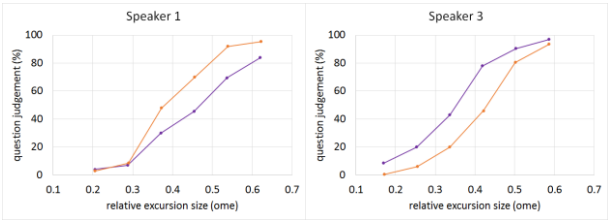


Figure 5: Evaluation of final rises measured on Octave Median Scale (purple=experiment 1, orange=experiment 2)

3.3. The Semitone Median Scale

Since the semitone scale did not prove to be reliable for representing the perception of rise excursions as a distance from offset to onset and the OMe Scale failed to represent the same by using the median as a reference point and the octave as a unit, we combined both scales into a new approach. The Semitone Median Scale measures the excursion size as the distance from the rise offset to the speaker median of the part of the utterance preceding the rise in semitones. At a first glance, Figure 6 does not suggest that the Semitone Median Scale improves the representation of pitch perception since the three speakers again show significant differences in both experiments (Experiment 1: speaker1=6.32, speaker2=5.70, speaker3=4.77, *b*=-0.77, *SE*=0.12, *df*=40.00, *t*=-6.19, ***p*<.001**; Experiment 2: speaker1=5.92, speaker2=5.41, speaker3=4.88, *b*=-0.52, *SE*=0.09, *df*=40.00, *t*=-5.58, ***p*<.001**). Nonetheless, it should be noted that the inconsistencies in the relationship between the three speakers across experiments have vanished.

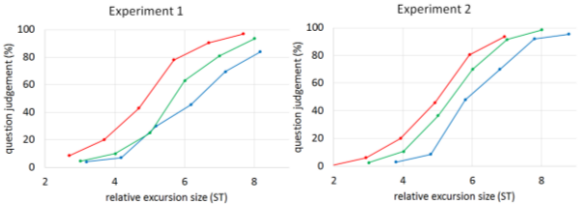


Figure 6: Evaluation of final rises measured on Semitone Median Scale (blue=speaker 1, green=speaker 2, red=speaker 3)

Figure 7 again shows the evaluations of speaker 1 and 3 across the two experiments. The results show that compared to the

Semitone Onset Scale and OMe Scale on the Semitone Median Scale there are no perceptual differences in pitch excursion within speakers across experiments (Speaker1: exp1=6.32, exp2=5.92, *b*=-0.39, *SE*=0.19, *df*=20.00, *t*=-2.03, **n.s.**, Speaker3: exp1=4.77, exp2=4.88, *b*=0.10, *SE*=0.20, *df*=20.00, *t*=0.53, **n.s.**). This suggests that the distance between a specific tonal target and the speaker's median in semitones is a stable cue to represent f0 excursions within speakers. It shall be noted that this is only the case when relating the rise offset to the median preceding the final rise and thus excluding the rise itself from the calculation of the median. Although De Looze and Hirst [6] provide a different approach, it seems plausible that the excursion that needs to be evaluated does not at the same time serve to calculate the reference value for its evaluation.

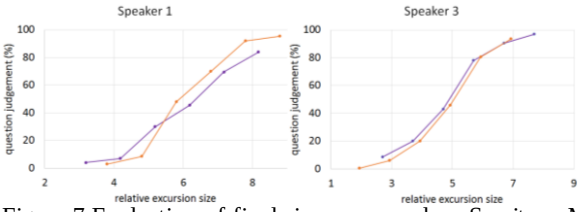


Figure 7 Evaluation of final rises measured on Semitone Median Scale (purple=experiment 1, orange=experiment 2)

Nonetheless, the interspeaker differences in both experiments remain, although the Semitone Median Scale succeeds in representing pitch perception based on f0. This may lead to the conclusion that there is no way to represent pitch perception across speakers based solely on f0 cues.

3.4. The Semitone Median Scale and Speaking Rate

To explain the interspeaker variation in the perception of pitch height we conducted a post-hoc analysis of non-intonational features on the three speakers. This yielded noticeable differences in speaking rate between all three speakers with 3.9, 4.2 and 4.7 syllables per second. Accordingly, we calculated a new measurement of relative excursion size by combining the distance between the rise offset and the median in semitones with the speaking rate in syllables per second according to the formula presented in (1).

$$12 * \log_2\left(\frac{Hz}{Median}\right) * \left(\frac{Syllables}{Utterancelength}\right) \quad (1)$$

Figure 8 shows that the crossing points of all three speakers more or less align in both experiments. While the differences between the crossing points still scarcely reach statistical significance in experiment 1, those differences completely vanish in experiment 2 (Experiment 1: speaker1=24.84, speaker2=23.75, speaker3=22.27, *b*=-1.29, *SE*=0.53, *df*=40.00, *t*=-2.43, ***p*<.05**; Experiment 2: speaker1=23.29, speaker2=22.53, speaker3=22.75, *b*=-0.27, *SE*=0.41, *df*=40.00, *t*=-0.65, **n.s.**).

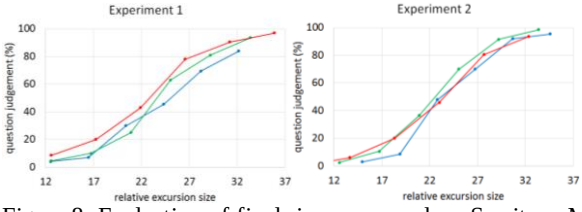


Figure 8: Evaluation of final rises measured on Semitone Median Scale incorporating speaking rate (blue=speaker 1, green=speaker 2, red=speaker 3)

4. Discussion

Neither the distance between rise offset and onset in semitones [2, 3, 4], the distance between rise offset and a fixed reference line in semitones [5] nor the distance between rise offset and the speaker's median in *ome* [6] succeeded in providing a stable cue to pitch perception. Nonetheless, the for this paper most important assumption that the speaker median serves as a reference value for f0 movements as suggested by De Looze and Hirst [6], was supported by the results of this study.

A finding that occurred through all of the scales was the differences across speakers within both experiments and thus different and equal register levels. Since all aspects of the f0 contour were controlled, we concluded that the cause for these differences in perception could not be attributed to fundamental frequency. This peculiar finding requires a measurement that takes features other than f0 into account for the description of pitch perception.

A preliminary non-intonational but prosodic feature that succeeded to better explain the across speaker differences in this paper was speaking rate measured in syllables per second. It has been shown that the differences in pitch perception nearly disappeared for different speakers across different registers and completely vanished for different speakers with the same register when speaking rate was included into the calculation. This suggests the existence of some sort of natural correlation between a higher speaking rate and greater f0 excursions. Accordingly, further production studies are necessary to investigate whether such a correlation exists. If this is the case, a higher speaking rate might invoke the expectation of higher f0 excursions, which causes f0 excursions failing to reach the projected topline to be evaluated as smaller than in speakers with a slower speaking rate. This is in accordance with the findings presented in this paper. Additionally, it is important to note that this correlation involves speaking rate and the distance of f0 excursions to the median and not the height of the median itself. We expect the correlation between lower register and higher speaking rate found in the three speakers to be merely coincidental.

In conclusion, the results suggest that measuring excursion sizes as the distance between an f0 point and the speaker's median combined with a measurement for speaking rate provides the most stable cue to represent f0 excursion with respect to pitch perception within and across speakers. Accordingly, we suggest incorporating this prosodic feature into the calculation of an auditory more adequate scale for pitch perception. As a basis for further discussion and experimental investigation both in production and perception we propose the formula presented in (1).

5. References

[1] D. R. Ladd and R. Morton, "The perception of intonational emphasis: continuous or categorical?," *Journal of Phonetics*, vol. 25, pp. 313-342, 1997.

[2] J. Michalsky, "Scaling of final rises in German questions and statements," *Proceedings of Speech Prosody 7, Dublin, Ireland, 2014*.

[3] J. Michalsky, *Frageintonation im Deutschen. Zur intonatorischen Markierung von Interrogativität*. Oldenburg. PhD thesis, 2015.

[4] J. Michalsky, "Phonetic effects of speaking style on final rises in German questions and statements," *Proceedings of ICPhS 18, Glasgow, Scotland, 2015*.

[5] J. Michalsky, "Pitch scaling as a perceptual cue for questions in German," *Proceedings of INTERSPEECH 2015, Dresden, Germany*, 2015.
[6] C. De Looze and D. Hirst, "The OMe (Octave-Median) scale: a natural scale for speech melody," *Proceedings of Speech Prosody 7, Dublin, Ireland, 2014*.
[7] D. R. Ladd, *Intonational Phonology*. Cambridge University Press, 2008.
[8] S. Baumann, M. Grice and S. Steindamm, "Prosodic marking of focus domains - categorical or gradient?," *Proceedings of Speech Prosody 5, Dresden, Germany, 2006*.
[9] S. Baumann, J. Becker, M. Grice and D. Mücke „Tonal and articulatory marking of focus in German," *Proceedings of ICPhS 16, Saarbrücken, Germany, 2007*.
[10] C. Féry and F. Kügler, "Pitch accent scaling on given, new and focused constituents in German," *Journal of Phonetics*, vol. 36, pp 680-703, 2008.
[11] F. Kügler "Focal lowering in German interrogatives," *Presentation at the DGfS 2015 conference, Leipzig, Germany, 2015*.
[12] C. Gussenhoven and T. Rietveld „Empirical evidence for the contrast between L* and H* in Dutch rising contours," in A. Botinis, G. Kouroupetroglou, and G. Caryannis (eds.), *Proceedings of the ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*. Athen: ESCA & University of Athens, pp. 169-172.
[13] J. Michalsky "Preliminary Results on Pitch Scaling for Signalling Focus and Sentence Mode in German - Competition, Merging or Splitting the Rise?," [Draft]
[14] C. Gussenhoven, *The phonology of tone and intonation*. Cambridge University Press, 2004.
[15] J. Vaissière, "Perception of intonation," in D. B. Pisoni and R. E. Remez (eds.), *The Handbook of Speech Perception*. Malden, MA: Blackwell Publishing, 2005, pp. 236-263.
[16] M. A. Halliday, *Intonation and grammar in British English*. The Hague: Mouton, 1967.
[17] J. D. O'Connor and G. F. Arnold, G. F., *Intonation of colloquial English*. London: Longmans, 1973.
[18] J. Pierrehumbert, *The Phonology and Phonetics of English Intonation*. MIT, PhD thesis, 1980.
[19] K. Pike, *The intonation of American English*. Ann Arbor: University of Michigan Press, 1945.
[20] M. Liberman, *The intonational system of English*. MIT, PhD thesis, 1975.
[21] M. Grice, S. Baumann and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology", in S.-A. Jun (eds.) *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 2005, pp. 55-83.
[22] F. Kügler, B. Smolibocki, D. Arnold, S. Baumann, B. Braun, M. Grice, S. Jannedy, J. Michalsky, O. Niebuhr, J. Peters, S. Ritter, C. T. Röhr, A. Schweitzer, K. Schweitzer and P. Wagner, "DIMA – Annotation guidelines for German intonation," *Proceedings of ICPhS 18, Glasgow, Scotland, 2015*.
[23] F. Nolan, "Intonational equivalence: an experimental evaluation of pitch scales," *Proceedings of ICPhS 15, Barcelona, Spain, 2003*.
[24] P. Boersma and D. Weening, *Praat: Doing phonetics by computer*. http://www.fon.hum.uva.nl/praat/, 2015.
[25] A. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, pp. 358-368, 1957.
[26] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
[27] D. Bates, M. Maechler, B. Bolker and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software* vol. 67, pp. 1–48, 2015.
[28] A. Kuznetsova, P. B. Brockhoff and R. H. B. Christensen, *lmerTest: Tests in linear mixed effects models*. R package version 2.0-30, 2016.

Effects of perceived attractiveness and likability on global aspects of fundamental frequency

Jan Michalsky[1] *& Heike Schoormann*[1]

[1]Institute of German Studies, University of Oldenburg, Germany
j.michalsky@uol.de, heike.schoormann@uol.de

Abstract

Acoustic parameters of the speech signal such as overall mean pitch, pitch range or variability of pitch can influence how listeners evaluate attractiveness as well as likability. This study asks two questions: Firstly, does the perception of the interlocutor's attractiveness or likability in turn influence the speaker's pitch behavior as well? Secondly, are those potential effects influenced by the interlocutor's pitch behavior in terms of entrainment? We conducted a speed dating experiment with 20 speakers in 100 mixed-sex pairs and analyzed acoustic correlates of the speakers' pitch as well as their evaluations of the interlocutors' attractiveness and likability. For both sexes, the results show a positive correlation of the speakers' pitch range with perceived attractiveness of the interlocutor and a positive correlation of the overall mean with the degree of perceived likability. Additionally, speakers showed a relative adaptation to the interlocutors' pitch which strengthened the effects of likability but diminished the effects of attractiveness. Conclusively, we suggest that speakers' pitch features are influenced by their perception of the interlocutor. However, adaptation to the interlocutors' pitch strongly interferes with these effects making it imperative to control for entrainment when investigating pitch effects of social variables.

Index Terms: pitch range, f0 mean, likability, attractiveness, perception, entrainment

1. Introduction

Pitch expresses not only linguistic but also paralinguistic functions. It is dependent on social aspects of the conversation and conveys information about the speaker's emotions, attitudes and personality traits. Additionally, pitch has been shown to contribute to a speaker's perception of an interlocutor regarding the overall impression of his/her personality in terms of likability or his/her appearance in terms of attractiveness.

By finding that female participants evaluated male voices with a lower mean pitch as more attractive, recent perception studies support the assumption that low pitch is associated with masculinity [1], [2], [3], and [4]. Ohala's [5], [6] frequency code describes how low pitch is linked to signalling largeness and accordingly dominance as often employed in the context of animal mating rituals. In contrast, high pitch has been assumed to signal femininity and thus be perceived as more attractive in female voices. Although comparable perception studies support this assumption [7], [8], and [4], the results are not as clear cut as for male voices. [9] and [10] contradict the previous findings. They report that male participants evaluate lower pitch as more attractive. One possible explanation is that attractiveness in men has been relatively unambiguously linked to masculinity while for female speakers there might be two conflicting strategies namely signalling sexiness or seductiveness through low pitch and signalling femininity through high pitch [11].

Likability on the other hand constitutes a cover term that has been studied under a variety of different names and a wide range of concepts. A common finding has been the association of lower pitch with warmth [12], pleasantness [13], [14], or likability in general [15], [16]. However, it has been suggested that this association is restricted to male voices, while in female voices higher pitch is perceived as friendlier [17]. Besides the general register or mean, greater pitch variety and range have been found to be perceived as positive, friendly, or pleasant [12], [13]. Again, contradicting evidence suggests that a shallower pitch range might be evaluated as more likable as well [15].

Albeit the contradictions in the results of previous studies, there is evidence that variable features of pitch can change the perception of attractiveness or likability. Consequently, the question arises whether the perception of attractiveness and likability also changes the pitch behavior, i.e. whether speakers try to sound more attractive or likable when attracted to or liking an interlocutor.

A handful of studies investigated the effects of attraction or perceived attractiveness on pitch. Their findings are largely compatible with the results on attractive voices reported above. Male speakers lowered their pitch, [18] whereas female speakers raised their pitch when speaking to a more attractive opposite-sex target [19]. However, female participants also lowered their pitch under comparable conditions [18]. On the one hand, the differing results could again hint at the previously mentioned opposition between sounding seductive and sounding feminine [11]. On the other hand, they are likely to result from the context with one strategy specifically restricted to mating situations [19].

So far, no studies explicitly investigated the effects of liking an interlocutor on pitch behavior. However, a majority of recent studies on entrainment include social variables pointing in the same direction. Accordingly, it has been found that speakers who were trying to be liked or giving encouragement entrained to their interlocutor, i.e. became closer in mean pitch [20]. A higher degree of pitch entrainment has also been found to correlate with stronger collaboration or rapport in learning dialogues [21]. Since entrainment of pitch is likely to increase with a stronger bond between participants, this should be considered when investigating the effect of perceived likability.

All in all, the results regarding the effects of attraction and liking on the pitch behavior remain inconclusive. Furthermore, the reported studies rarely separate the two aspects which leads to attractiveness being influenced by likability and vice versa. Another common drawback of the reported studies is that the production data were rarely matched with the perception data of the actual participants. Instead, the interlocutor's attractiveness or likability was often judged by external observers, completely detached from the respective conversation. The perception of a voice's likability may greatly vary between the active participants of the conversation and external judges.

The study at hand investigates the effects of attraction/attractiveness and liking/likability in natural spontaneous speech and explicitly separates the evaluations of both parameters. Production and perception data were collected from the actual participants engaged in the conversation, i.e. from the same experiment. Lastly, because research on entrainment suggests that speakers adapt to their interlocutor in pitch depending on the interpersonal relationship [22], [20] as well as regardless of social factors as an automatism [23], [24], the influence of the interlocutor's pitch behavior is considered in the current investigation. Through the experimental analysis we seek to answer the following two research questions:

1) Does the perception of attractiveness/likability correlate with global features of the speakers' pitch?

2) Do perceived attractiveness/likability and the interlocutors' pitch features show interactions in affecting the speakers' pitch?

2. Method

2.1. Speakers

Ten female and 10 male students from the University of Oldenburg participated in the study as paid volunteers. All subjects were monolingual speakers of High German aged between 19 and 28 years. They all grew up in (northern) Germany, i.e. share a common cultural background. Only heterosexual singles were included in the study. All subjects were unacquainted and had no interactions prior to the experiment.

2.2. Procedure

The subjects participated in a speed dating setting, which was altered to meet the research objective. To this end, each participant was paired with each of the 10 participants of the opposite sex resulting in a total of 100 opposite-sex combinations from which 98 could be included in the acoustic analysis. The subjects were placed in a quiet room and instructed to freely engage in a conversation with no restrictions to the topics. A note with sample topics was placed on the table in case participants had difficulties starting the conversation. Each conversation lasted between 15 and 20 minutes. Prior to the first verbal interaction as well as immediately after each conversation, participants received a questionnaire and were asked to evaluate their interlocutor in terms of purely visual attractiveness as well as perceived likability on a 10-point Likert scale. The participants were given the necessary privacy and their ratings were not revealed to the respective interlocutor. Recordings were made in stereo using a portable digital recorder (Tascam HD P2) at a sampling rate of 48 kHz and 24 bit resolution with one head-mounted microphone (DPA 4065 FR) per speaker.

2.3. Acoustic analysis

The acoustic analysis was carried out using Praat [25]. Audio tracks were separated for each speaker. Subsequently, all filled pauses, laughter, overlaps as well as the interlocutors' speech parts were manually silenced to preserve the time structure of the recordings. F0 features were extracted from the f0 track of the conversation as a whole and all measurements calculated in semitones. Table 1 shows the measured variables. All measurements were taken from each speaker as well as their respective interlocutor for the measurement of entrainment. In the scope of this paper entrainment is restricted to relative global synchrony describing the correlation between a speaker's pitch features and the corresponding feature of their interlocutor [26], [24], and [20]. Convergence and similarity, local entrainment, as well as exact entrainment will not be included.

Table 1: *Acoustic measurements of f0 features*.

Feature	Description
mean	overall f0 mean
median	overall f0 median
max	95th percentile (to exclude outliers)
range1	difference between upper quartile and lower quartile
range2	difference between f0 max and f0 mean
variation	standard deviation

2.4. Statistical analysis

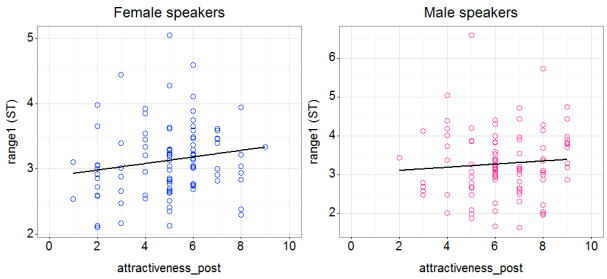
For the statistical analysis we conducted linear mixed effects models using R [27], the lme4-package [28] as well as the lmerTest-package [29]. Model fit was determined by maximum likelihood ratio tests. *P*-values were calculated using Satterthwaite-approximation. As fixed effects we used the perceived attractiveness (*attractiveness_post*) and perceived likability (*likability_post*) after the conversations as well as *speaker sex*. As random effects we included random intercepts for *speaker*. The dependent variables were the six f0 measurements listed in table 1. We included the corresponding pitch feature of the interlocutor as a fixed effect in the analysis of entrainment.

3. Results

3.1. Perceived attractiveness without entrainment

Significant effects of perceived attractiveness were found for the speakers' pitch range (*range1*) from upper quartile to lower quartile (b=.05, SE=.02, df=182.19, t=2.58, $\mathbf{p<.05}$). Pitch range was positively correlated and increased with the degree of perceived attractiveness. No effect of *speaker sex* or its interaction with *attractiveness_post* was found for *range1*. Additionally, contradicting expectations from previous research, no significant effects were found for *f0 mean* or *f0 median*. Furthermore, none of the other f0 measurements, including *range2* and *variation*, reached significance. Figure 1 illustrates the main results for perceived attractiveness.

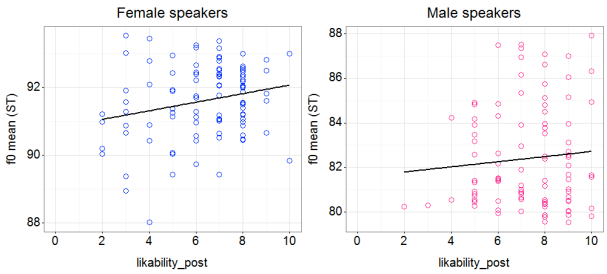
Figure 1: Effects of perceived attractiveness on the speakers' pitch range.



3.2. Perceived likability without entrainment

We found significant effects of perceived likability on the speakers' *f0 mean* (*b*=.06, *SE*=.03, *df*=176.51, *t*=1.98, ***p*<.05**). Again, these were positively correlated for both sexes without significant interaction between *likability_post* and *speaker sex*. Male speakers as well as female speakers raised their *f0 mean* with an increasing degree of perceived likability of the interlocutor. All other pitch features, including *f0 median*, failed to reach significance. Figure 2 illustrates the main results for perceived likability.

Figure 2: Effects of perceived likability on the speakers' f0 mean.



3.3. Perceived attractiveness including entrainment

We found a significant correlation between the speakers' pitch range (*range1*) and the interlocutors' pitch range (*b*=.11, *SE*=.04, *df*=175.23, *t*=2.32, ***p*<.05**). Pitch ranges were positively correlated for both sexes without interaction with *speaker sex*. Figure 3 illustrates the results of pitch range entrainment. Furthermore, when including the interlocutor's pitch range as a fixed effect, perceived attractiveness failed to reach significance. This suggests that perceived attractiveness somehow correlates with the interlocutor's pitch range and does not serve as an independent factor for the speaker's pitch range. Furthermore, perceived attractiveness and the interlocutors' pitch range showed no significant effects of interaction. We conducted post hoc linear mixed effects models on the correlation of the interlocutor's pitch range and perceived attractiveness and found significant effects (*b*=.39, *SE*=.18, *df*=183.13, *t*=2.25, ***p*<.05**). Perceived attractiveness and the interlocutors' pitch range were positively correlated for both sexes and showed no interaction with *speaker sex*. Figure 4 illustrates the results for the correlation of perceived attractiveness and the interlocutors' pitch range.

Figure 3: Correlation of the speakers' pitch range and the interlocutors' pitch range.

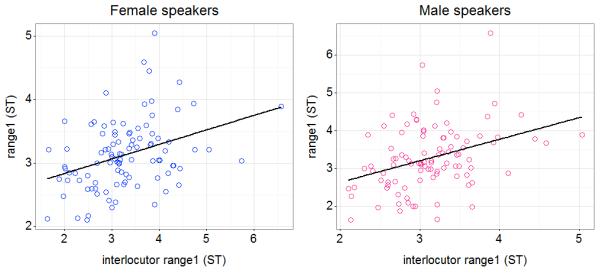
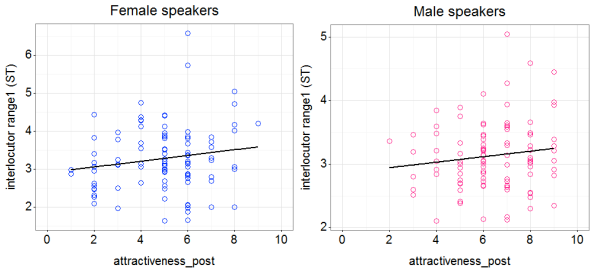


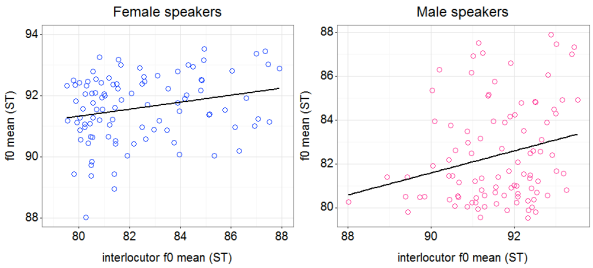
Figure 4: Correlation of perceived attractiveness and the interlocutor's pitch range.



3.4. Perceived likability including entrainment

A significant correlation between the speakers' *f0 mean* and the interlocutors' *f0 mean* was found was well (*b*=.12, *SE*=.02, *df*=174.16, *t*=4.73, ***p*<.001**). The *f0 means* were again positively correlated for both sexes. The interlocutors' *f0 mean* did not interact with *speaker sex*. Figure 5 illustrates the results for *f0 mean* entrainment. In contrast to perceived attractiveness, the effects of perceived likability still reached significance when entrainment was included in the model (*b*=.06, *SE*=.03, *df*=175.38, *t*=2.28, ***p*<.05**) and even showed a slight increase in significance (***p*=.048** to ***p*=.024**). Lastly, the interlocutors' *f0 mean* did not interact with perceived likability.

Figure 5: Correlation of the speaker's f0 mean and the interlocutor's f0 mean.



4. Discussion

While pitch behavior has been found to influence the perception of social attributes such as attractiveness [1], [7], [2], [8], [3], [4], [10], and [9] or likability [12], [13], [15], [14], and [16], the results suggest that perceived attractiveness and likability in return affect the pitch behavior of a speaker. Since attractiveness was found to correlate with lower overall pitch in male speakers [23], [2], [3], and [4], as well as both lower and higher pitch in female speakers [7], [8], [4], [10], and [9], it was assumed that speakers show comparable strategies of sounding more attractive when speaking to a

more attractive opposite sex target, which has been supported by previous studies [19], [18]. The results of the paper at hand contradict this assumption since no effects of perceived attractiveness on the f0 mean or the f0 median were found for either sex in any direction. However, we found perceived attractiveness to affect the speakers' pitch range, which might be explained by a higher degree of involvement or interest correlating with a higher degree of perceived attractiveness. The results further suggest that the pitch range is not only dependent on the interlocutors' attractiveness but also on his/her pitch range. Furthermore, when including the speakers' entrainment to the interlocutors' pitch range, the effects of perceived attractiveness disappeared. This suggests that the speakers did not adjust to the interlocutors' attractiveness but to his/her respective pitch range in general which coincided with his/her perceived attractiveness. To investigate this correlation, we conducted a post hoc analysis and found that perceived attractiveness indeed correlated significantly with the interlocutor's pitch range. The nature of this correlation could not be investigated at this point but assuming causality it suggests two possible explanations. Either interlocutors who are perceived as more attractive based on purely visual criteria employ a greater pitch range, or a greater pitch range significantly influences the perceived attractiveness of an interlocutor. Both possibilities remain open to further investigation.

However, it needs to be discussed why our results greatly contradict the assumption based on the research overview. Previous research elicited attracted speech under very strict experimental conditions without actual interaction between the speaker and the supposed addressee by presenting participants only with a picture of the addressee and instructing them to leave a voice mail [19]. Another important factor might be that attractiveness of the presented addressees was assessed beforehand based on scientific criteria and not on the actual perception of the participants. Either the nature of the interaction or the variability of subjective perception of attractiveness or both might have influenced the results. Accordingly, the combination of perception and production can be considered an advantage.

In addition, the experiment presented in this paper might have failed to elicit the intended mating context assumed by a speed dating scenario. Prior to the experiment, all speakers filled out separate questionnaires including information about their motivation to participate in multiple choice form. Possible answers included an interest in getting to meet new people, in conversations in general, in flirting, and in meeting a partner. Although explicitly advertised as a speed dating experiment, none of the female speakers and only three of the male speakers stated that they were interested in flirting or even meeting a partner. A preliminary analysis of the content of the conversations supports the assumption that the experiment succeeded in eliciting fluent conversations but failed to elicit a recognizable mating context.

In case of likability, the research overview has drawn an inconsistent picture regarding the attributes of likable voices as well as the influence of likability. Nonetheless, the results suggest one consistent main effect with speakers of both sexes increasing their f0 mean continuously with a greater degree of perceived likability. The assumption that higher overall pitch corresponds with liking or likability is supported by some studies [15], [17] but also contradicted by others [13], [14], and [16]. Additionally, also the finding that perceived likability did not interact with *speaker sex* and female and male speakers thus employ the same strategies, is supported by some studies [17], [15] and contradicted by others [13], [14]. As mentioned before, these contradictions can be explained through the general inconsistency of the likability concept throughout the literature. Again, the combination of perception and production data of the same speakers can be interpreted as an advantage since we found significant effects albeit this inconsistency and the high degree of noise in natural speech.

Furthermore, we found a significant correlation between the speaker's and the interlocutor's f0 mean. Both means were positively correlated for both sexes which means that all speakers regardless of sex adjusted their relative f0 mean to the relative position of the interlocutor's f0 mean. In contrast to the effects of perceived attractiveness, the effects of perceived likability prevailed when including entrainment and additionally showed no interaction with the interlocutors' f0 mean. This suggests that speakers adjust to their interlocutors' f0 mean regardless of their likability and additionally to their perceived likability independently. The phonetic effects of likability draw a very complex picture since previous studies suggest that there might also be an interaction between pitch entrainment and social behavior strongly correlating with mutual liking, which was not incorporated in the present study [15], [20].

Conclusively, we found that aspects of a speaker's pitch might, among other linguistic and paralinguistic functions, be influenced by the speaker's perception of his/her interlocutor in terms of attractiveness and likability. Furthermore, speakers entrain to their respective interlocutor's pitch behavior, thus strengthening or weakening the sheer effects of said social variables. Accordingly, it is imperative for future studies on social pitch behavior to take both the speakers' as well as their interlocutors' acoustic measurements into account.

5. References

[1] S. A. Collins, "Men's voices and women's choices," *Animal Behaviour*, vol. 60, pp. 773–780, 2000.

[2] D. R. Feinberg, L. M. Debruine, B. C. Jones, and D. I. Perrett, "Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices," *Animal Behaviour*, vol. 69, pp. 561–568, 2005.

[3] C. R. Hodges-Simeon, S. J. C. Gaulin, and D. A. Puts, "Different vocal parameters predict perceptions of dominance and attractiveness," *Human Nature*, vol. 21, pp. 406–427, 2010.

[4] B. C. Jones, D. R. Feinberg, L. M. Debruine, A. C. Little, and J. Vukovic, "A domain-specific opposite-sex bias in human preferences for manipulated voice pitch," *Animal Behaviour*, vol. 79, no. 57-62, 2010.

[5] J. Ohala, "Cross-language use of pitch. An ethological view," *Phonetica*, vol. 40, pp. 1–18, 1983.

[6] J. Ohala, "An ethological perspective on common cross-language utilization of f0 in voice," *Phonetica*, vol. 41, pp. 1–16, 1984.

[7] S. A. Collins and C. Missing, "Vocal and visual attractiveness are related in women," *Animal Behaviour*, vol. 65, pp. 997–1004, 2003.

[8] D. R. Feinberg, L. M. Debruine, B. C. Jones, and D. I. Perrett, "The role of femininity and averageness of voice pitch in aesthetic judgements of women's voices," *Perception*, vol. 37, pp. 615–623, 2008.

[9] T. Oguchi and H. Kikuchi, "Voice and interpersonal attraction," *Japanese Psychological Research*, vol. 39, pp. 56–61, 1997.

[10] K. Leaderbrand, J. Dekam, A. Morey, and L. Tuma, "The effects of voice pitch on perceptions of attractiveness: Do you sound hot

or not," *Winona State University Psychology Student Journal*, 2008.

[11] A. Karpf, *The human voice*. New York, NY: Bloomsbury Publishing, 2006.

[12] B. L. Brown, W. J. Strong, and A. C. Rencher, "Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech," *Journal of the Acoustical Society of America*, vol. 55, no. 2, pp. 313–318, 1974.

[13] L. Bruckert, J. Lienard, A. Lacroix, M. Kreutzer, and G. Leboucher, "Women use voice parameter to assess men's characteristics," *Proceedings in Biological Sciences*, vol. 237, pp. 83–89, 2006.

[14] B. Ketzmerick,"Zur auditiven und apparativen Charakterisierung von Stimmen," *Studientexte zur Sprachkommunikation. TUD Press, Dresden*, 2007.

[15] A. Gravano, R. Levitan, L. Willson, S. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," *Proceedings of Interspeech*, 2011.

[16] B. Weiss, „Prosodische Elemente vokaler Sympathie," *Studientexte zur Sprachkommunikation. TUD Press, Dresden*, vol. 65, pp. 212–217, 2013.

[17] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics*, pp. 638–646, 2009.

[18] S. M. Hughes, S. D. Farley, and B. C. Rhodes, "Vocal and physiological changes in response to the physical attractiveness of conversational partners," *Journal of Nonverbal Behavior*, vol. 34, pp. 1–13, 2010.

[19] P. J. Fraccaro, B. C. Jones, J. Vukovic, F. G. Smith, C. D. Watkins, D. R. Feinberg, A. C. Little, and L. M. Debruine, "Experimental eivdence that women speak in higher voice pitch to men they find attractive," *Journal of Evolutionary Psychology*, vol. 9, no. 1, pp. 57–67, 2011.

[20] R. Levitan, A. Gravano, L. Willson, S. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 11–19, 2012.

[21] N. Lubold and H. Pon-Barry, "Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues," *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge, November 12-12, 2014, Istanbul, Turkey*, 2014.

[22] H. Giles, N. Coupland, and J. Coupland, "Accomodation theory: Communication, context, and consequence. Contexts of accomodation," *Developments in applied sociolinguistics*, vol. 1, 1991.

[23] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[24] R. Levitan, *Acoustic-prosodic entrainment in human-human and human-computer dialogue. Columbia University*. PhD thesis, 2014.

[25] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*, 2016.

[26] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," *Proceedings of Interspeech*, 2009.

[27] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[28] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[29] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, *lmerTest: Tests in linear mixed effects models*. R package version 2.0-30, 2016.

Opa vs Oper: Neutralization of /ɐ/ and unstressed /a/ contrast in a perception and production study.

Bokelmann, J.[1]; Ellert, M.-T.[1]; Friesenhan, N.[1]; Generlich, G.[1]; Hiebert, Y.[1]; Malon, T.[1]; Naumann, F.[1]; Sachs, R.[1]; Swiech, A.[1]; Walak, M.[1]; Yovcheva, M.[1]; Zander, I.[1]; Rathcke, T.[2]; Mooshammer, C.[1]

[1] Institut für deutsche Sprache und Linguistik, Humboldt Universität zu Berlin, Germany
[2] English Language and Linguistics, University of Kent, UK

T.V.Rathcke@kent.ac.uk, christine.mooshammer@hu-berlin.de (corresponding authors)

Abstract

The present study examined differences in production and perception of the German vowels /a/ and /ɐ/ in word-final, unstressed position. In the first experiment, 3 male and 3 female speakers produced minimal pairs embedded in meaningful sentences and varied in prosodic environment. In the second experiment, the minimal pairs were extracted from the context and presented to 44 listeners for a forced-choice identification task. Results showed a better-than-chance performance that was, however, mainly driven by one male speaker. Temporal and spectral measures confirmed that only this speaker produced an acoustic difference between /a/ and /ɐ/.

Index Terms: reduced vowels, German, vowel production, vowel perception

1. Introduction

Traditionally, the final Standard German vowels in *Opa* and *Oper* are transcribed with two different symbols, /a/ and /ɐ/, respectively. However, it is not clear if this contrast is indeed produced and perceived by native speakers of Standard German. In descriptions of German phonology the /ɐ/-schwa following consonants is derived from /əʁ/ (see e.g. Hall 1993 [7]). In this view, [ɐ] and [ʁ] are not two phonemes, but two allophones of one rhotic phoneme in complementary distribution: [ʁ] appears in the onset of a syllable, [ɐ] in any other position.

Meinhold (1989) [11] listed the following contexts for the occurrence of the /ɐ/-schwa: as a postvocalic /ʁ/-allophone, as in *hört* [hø:ɐt], as a realization of the <-er> suffix, as in *weiter* ['vaɪ̯tɐ] and as a realization of the prefixes <er-, ver-, zer->. He further showed that both schwas, [ɐ] and [ə], could be correctly identified, provided there is sufficient context to follow the vowel.

Barry (1995) [1] investigated the relation between /ə/ and /ɐ/ and especially addressed the issue of how the context influenced the realisation of the two phonemes. He found less variability in the production of /ɐ/, but the results of the study were limited to just one speaker. The study by Dittrich and Reibisch (2006) [6] provided evidence against postvocalic diphthongisation following a long vowel /a:/, i.e. in words like *Paar* 'pair', unlike in words like *hört* [hø:ɐt], mehrt [me::ɐt], the stressed vowel showed formant trajectories of a monophthong. Therefore, the authors argued against transcribing [a:ɐ] in this context.

The current perception and production experiment focuses on the difference between word-final unstressed /a/ and /ɐ/. Both vowels are supposed to be central vowels but the two IPA symbols imply a perceivable difference in vowel height with /a/ being lower than /ɐ/ (see e.g. Kohler 1990) [8]. Vowels in unstressed position usually undergo target undershoot (see Lindblom 1990) [10] which results in a more closed tongue configuration and a lower F1 frequency for German unstressed /a/ than for stressed /a/ (see Mooshammer & Geng 2008) [12]. Target undershoot could potentially lead to a neutralization of the contrast between these two vowels in unstressed position by raising the unstressed /a/.

According to Vennemann (1991)[14] the two vowels /a/ and /ɐ/ also belong to different vowel sets distinguished by a prosodic characteristic: /a/ is a full vowel of German that can occur in stressed position whereas /ɐ/ is a reduction vowel that cannot be stressed. Within Articulatory Phonology, the schwa vowel is assumed to be targetless, i.e. articulators that are not involved in an active gesture move towards a neutral position. Since German has two schwas /ə/ and /ɐ/, they cannot be completely targetless but they might still be more variable than other vowels (see Barry 1995 [1] for a discussion). Assuming a schwa vowel does not have a target it should be more prone to coarticulatory influences of the neighbouring segments.

The aim of this study is to investigate the perception and production of word-final unstressed /a/ and /ɐ/ in the following conditions: phrase-medial vs. -final position, accented and unaccented. When accented, the difference in production and therefore perception is expected to be more salient than in the unaccented condition due to hyperarticulation (de Jong et al. 1993) [5]. When in phrase-final position, the distinction is also expected to be less salient, due to lesser coarticulation. In the phrase-medial position, the coarticulation effects should support the distinction of the two phonemes and show higher recognition of /ɐ/.

2. Method

2.1. Speakers and Material

We recorded six speakers (3 male, 3 female) originating from the area of Kiel.

The minimal pairs of the investigation were:

- *Dina – Diener* ([diːna] - [diːnɐ]; female first name – butler, servant)
- *Opa – Oper* ([ʔoːpʰa] - [ʔoːpʰɐ]; grandfather – opera)
- *Feta – Väter* ([feːtʰa] - [feːtʰɐ]; feta cheese – fathers),
- *Clara – klarer* ([klaːʁa]- [klaːʁɐ]; female first name – clearer).

Each word was embedded in meaningful sentences that were designed to vary the prosodic environment - the phrasal position (phrase-final vs. phrase-medial), level of prominence (accented vs. unaccented) – as well as the following segmental context - next word starting with /z/ vs. /l/. Here are 4 examples for our test sentences, varying the condition phrasal position for the minimal pair *Opa – Oper* (all accented and with following /l/ context):

(1) *Meine Großeltern sind toll. Vor allem mein **Opa** lässt sich immer so spannende Geschichten einfallen. Es ist toll ihm zu zuhören*. (medial position)
My grandparents are amazing. Especially, my grandpa is always coming up with exciting stories. It is great listening to him.

(2) *Ich liebe meine Oma, aber nicht so sehr wie meinen **Opa**. Lass es sich vielleicht gemein anhören, aber es ist nun mal so.* (final position)
I love my grandma, but not as much as my grandpa. It might sound mean but that's how it is.

(3) *Auf den Kieler Bühnen läuft am Wochenende nicht viel. Ich habe die **Oper** letzte Woche schon gesehen. Aber wir können ins Schauspielhaus gehen.* (medial position)
On Kiel stages there is not much happening this weekend. I have already seen the opera last week. But we could go to the theatre.

(4) *Ich gehe nicht oft aus. Nur ab und zu in die **Oper**. Letzten Monat war ich dreimal dort.* (final position)
I don't go out very often. Only sometimes to the opera. Last month I was there three times.

Each sentence was produced 3 times in randomized order. Thus each speaker produced a total of 192 target tokens (4 minimal pairs x 2 accent conditions x 2 contexts x 2 phrase positions x 3 repetitions).

2.2. Perception Test

2.2.1. Stimuli

In order to test the target words separately from their context, we selected one instance of the three repetitions for which the word was produced clearly, without errors, stutters or hesitation and could be extracted easily Only words in /z/ context were used here. Silences of 100 ms were added before and after each word. These 384 stimuli (32 per speaker) were used to create a perceptual experiment script, using the speech analysis software *Praat* (Boersma & Weenink 2016)[3].

2.2.2. Listeners

44 native German speakers between 18 to 40 years old participated in the experiment. None of them reported any hearing impairment.

2.2.3. Procedure

For the experiment procedure, we used laptops running the experiment script in *Praat* [3] and a pair of headphones. A blank screen was shown for 100 ms before and after each stimulus. Then a choice of two buttons labelled with the respective minimal pair appeared on the screen alongside with the written question "*Which word did you hear?"*. Participants were instructed to click on the button of their perceptual choice to give their answer as appropriate and to proceed to the next stimulus. Each stimulus was randomly presented twice during the experiment to counterbalance the position of the correct target button on the screen. Overall, the experiment took approximately 20 to 30 minutes.

2.3. Acoustic Annotation and Measurements

Praat [3] was used to annotate and analyse the characteristics of unstressed [a] and [ɐ]. Measurements included the durations of the initial stressed syllable, of the second unstressed syllable and of the vowel. Accent, boundaries, context, type of syllable and segments were annotated to compare the results. For acoustic analysis standard procedures in EmuR were used (Bombien et al. 2006) [4].

2.4. Acoustic Annotation and Measurements

All statistics were carried out using R 3.3.0 (see R Core Team 2016) [13] with the packages lme4 (Bates et al. 2015) [2] and lmeTest (Kuznetsov et al. 2016) [9].

3. Results

3.1. Perception Test

Figure 1 shows the proportion of correct responses broken down by speaker and accentuation (pooled for all target words). For most speakers the number of correct responses was close to chance. The proportion of correctly perceived stimuli was slightly higher in accented than unaccented words. The results for the speaker M2 are remarkable: His scores were much higher compared to all other speakers.

Logistic linear mixed effects models indicate that there is a significant effect of accent ($p<0.01$), position ($p<0.05$), word ($p<0.001$) and speaker ($p<0.001$). The speaker effect is based on speaker M2 whose stimuli were recognised correctly more often than other speakers' items (see also Fig. 1). Excluding the perceptual results for this speaker, the effect of speaker and the effect of position are not significant. For this speaker words in final position were identified better than in medial position. For all speakers, the presence of accentuation improved recognition significantly.

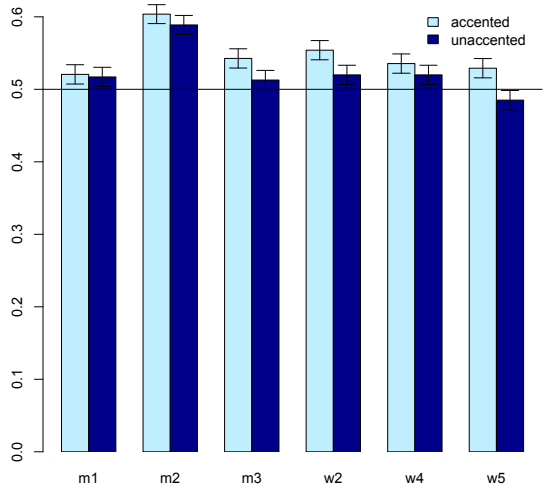


Figure 1: *Means and standard deviation of correct answers for 6 speakers and two accentuation levels (light blue: accented; dark blue: unaccented).*

3.2. Acoustic Analysis

So far, preliminary acoustic analyses have included exclusively three male speakers. The first parameter addressed here is the ratio of the duration of the stressed syllable to the unstressed syllable, shown in Figure 2 for the analysed speakers. The hypothesis is that reduced /ɐ/ syllables are shorter than unstressed syllables at the full vowel (yielding larger ratios for reduced /ɐ/ syllables). This was tested by calculating linear mixed effects models. There was no significant difference between reduced /ɐ/ syllables and unstressed /a/ syllables but as can be seen in the Figure 2, speaker M2 tends to have lower ratios for words with /ɐ/ syllables than for words with unstressed /a/ syllables.

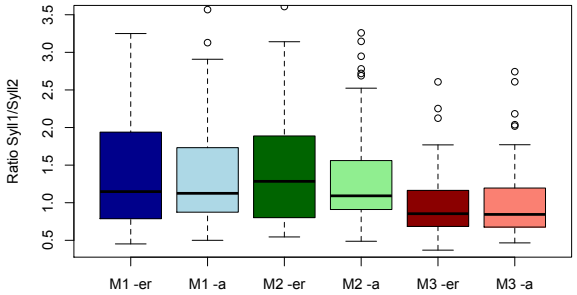


Figure 2: *Boxplots of the duration ratios of syllable 1 to syllable 2 for speakers M1 (blue), M2 (green) and M3 (red). Darker colours represent items with reduced /ɐ/ and lighter colours items with unstressed /a/.*

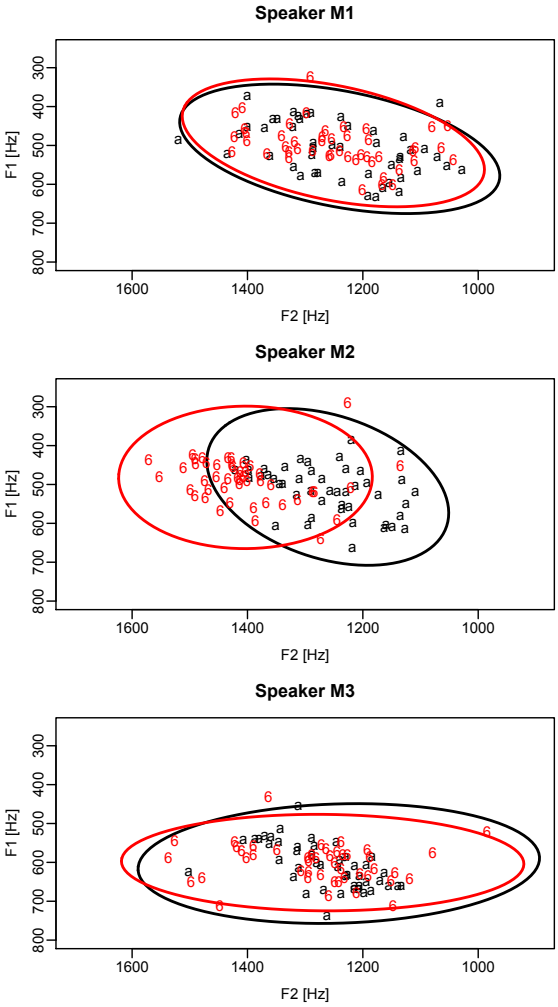


Figure 3: *Scatterplots of the formant frequencies for F1 and F2 with 2 SD dispersion ellipses for the vowels /ɐ/ (denoted as 6) in red and /a/ in black.*

Secondly, the formant frequencies of F1 und F2 were considered in order to quantify a potential quality difference between the two vowels. Figure 3 shows dispersion ellipses of the formant frequencies measured in mid vowel for reduced /ɐ/ (in red, denoted as "6") and unstressed /a/ (in black) for the three speakers. Speakers M1 and M3 display complete overlap of the ellipses for /a/ and /ɐ/, implying that there is no quality difference between the two vowels. Speaker M2, however, does distinguish the two variants with the unstressed /a/ having a lower F2, suggesting a more retracted position compared to /ɐ/.

4. Discussion and Conclusion

Results from the perception test with around 53.5 % of correct recognition suggest that the difference between /a/ and /ɐ/ is subtle and therefore difficult to detect. Identification scores implied that accent enhanced the participant's ability to distinguish the given stimuli slightly but significantly to 54.8 %, supporting the associated hypothesis. Contrary to our hypothesis identification scores were improved in final

position. We assumed that more coarticulation in medial position might enhance the contrast because the reduced vowel /ɐ/ should be affected by the context to a greater degree than /a/. The opposite was the case. However, both effects, accent and position, vanished when one speaker, M1, was excluded from the data set.

This was corroborated by the analysis of the corresponding production data with almost no differences between temporal measures and formant values. This would lead to the conclusion that there is only a very slight contrast, most of the time not perceptible, between unstressed /a/ and /ɐ/. However, one speaker had improved recognition rates and at the same time showed a contrast for the acoustic measures, speaking for a clear connection between perception and production. For further investigating individual differences, we will provide acoustic data from the remaining speakers.

In conclusion, unstressed /a/ and /ɐ/ are almost undistinguishable because the difference is only rarely produced by speakers of German. Different IPA symbols should only be used for detectable differences.

5. References

[1] Barry, W.J. (1995): Schwa vs. schwa + /r/ in German. In: *Phonetica, 52,* pp. 228-235.

[2] Bates, D., Maechler, M., Bolker, B., Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

[3] Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.20, retrieved 3 September 2016 from http://www.praat.org/

[4] Bombien, L., Cassidy, S., Harrington, J., John, T., & Palethorpe, S. (2006, December). Recent developments in the Emu speech database system. In: *Proc. 11th SST Conference Auckland*, pp. 313-316.

[5] De Jong, K., Beckman, M. E., & Edwards, J. (1993). The interplay between prosodic structure and coarticulation. *Language and speech*, *36*(2-3), 197-212.

[6] Dittrich, R.& Reibisch, G. (2006): An acoustic study of /r/-vocalization in word-finalposition. Arbeitsberichte des Instituts für Phonetik und digitaleSprachverarbeitung der Universität Kiel, pp. 19-26.

[7] Hall, T. A. (1993): The Phonology of German /R/. In: Phonology, Vol. 10, No. 1 (1993), pp.83-105.

[8] Kohler, K. (1990) Illustrations of the IPA: German. Journal of the International Phonetic Association, 20(1), 48-60.

[9] Kuznetsova, A., Bruun Brockhoff. P. and Haubo Bojesen Christensen, R. (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-32. https://CRAN.R-project.org/package=lmerTest.

[10] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403-439). Springer Netherlands.

[11] Meinhold, G. (1989): Das problematische [ɐ]. In: Slembek, Edith (Hg.): Von Lauten und Leuten: Zeitschrift für Peter Martens zum 70.Geburtstag. Frankfurt am Main

[12] Mooshammer, C., & Geng, C. (2008). Acoustic and articulatory manifestations of vowel reduction in German. *Journal of the International Phonetic Association*,*38*(02), 117-136.

[13] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[14] Vennemann, T. (1991). Skizze der deutschen Wortprosodie. *Zeitschrift für Sprachwissenschaft*, *10*(1), 86-111.

Universal phonetics revisited: Eine cross-linguistische Untersuchung zum Einfluss der Stimmhaftigkeit des Folgekonsonanten auf die Vokallänge im Polnischen und Deutschen

Katharina Nimz[1], *Judith Baumann*[1], *Arkadiusz Rojczyk*[2]

[1]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
[2]Institute of English, University of Silesia

katharina.nimz@uni-bielefeld.de, judith.baumann@web.de, arkadiusz.rojczyk@us.edu.pl

Abstract

Die vorliegende Studie umfasst zwei Produktionsexperimente, die sich mit dem Einfluss der Stimmhaftigkeit des Folgekonsonanten auf die Vokallänge im Deutschen und Polnischen befassen. Beide Sprachen sind prominente Beispiele für die phonologische Regel der Auslautverhärtung, weshalb die Mehrzahl der vorherigen Studien Vokallänge im Rahmen von *Incomplete Neutralization* untersucht hat. Bisher existieren kaum Daten zu Vokallängen im Deutschen und Polnischen in nicht-neutralisierenden Kontexten; unsere Studie greift dieses Forschungsdesiderat auf.

Schlüsselbegriffe: Vokalproduktion, Vokallänge, Einfluss von Stimmhaftigkeit, Deutsch, Polnisch

1. Einleitung

Sowohl im Polnischen als auch im Deutschen wird der Kontrast zwischen stimmhaften und stimmlosen Konsonanten am Wort- bzw. Silbenende zugunsten des stimmlosen Konsonanten neutralisiert (z. B. *Kinder* [kʰɪn.dɐ], aber *Kind* [kʰɪnt]). Eine Reihe von experimentellen Untersuchungen hat in beiden Sprachen nachgewiesen, dass die Neutralisierung zum Teil nicht vollständig ist und sich vorangehende Vokale beispielsweise in ihrer Vokallänge vor (zugrundeliegend) stimmhaften versus stimmlosen Konsonanten unterscheiden (für Polnisch z. B. [1, 2]; für Deutsch z. B. [3, 4]).

Die Relevanz der Vokallängenunterschiede in diesen Studien setzt voraus, dass Vokallänge in nicht-neutralisierenden Kontexten ebenfalls eine Rolle bei der Unterscheidung von stimmhaften versus stimmlosen Konsonanten spielt. Dies wäre im Deutschen und Polnischen der Fall für intervokalische Konsonanten in deutschen Minimalpaaren wie *Boden* versus *Boten* oder *nuta* („Note“) versus *nuda* („Langeweile“) im Polnischen. In seiner vielzitierten cross-linguistischen Untersuchung zu Vokallängenunterschieden in verschiedenen Sprachen postuliert bereits Chen [5], dass Vokale vor stimmlosen Konsonanten physiologisch bedingt kürzer seien als vor stimmhaften Konsonanten. Die von ihm zitierten Daten für das Deutsche stammen allerdings aus einer Studie von 1903, was unter heutigen Standards zur akustischen Analyse von Lautdauern kaum eine valide Quelle sein kann. Des Weiteren zweifelt Keating [6] an, dass Vokallängenunterschiede in Abhängigkeit von der Stimmhaftigkeit des Folgekonsonanten wirklich sprachübergreifend gelten und präsentiert Daten zum Polnischen, die keine signifikanten Vokallängenunterschiede aufweisen. Ihre Analyse basiert lediglich auf einem Minimalpaar.

Aufgrund der spärlichen Datenlage zu Vokallängenunterschieden in diesem spezifischen Kontext wurden zwei Produktionsexperimente mit Polnischen und Deutschen Probandinnen und Probanden durchgeführt.

2. Produktionsexperimente

Beide Experimente wurden mit jeweils 6 Versuchspersonen durchgeführt. Alle Probandinnen und Probanden sprachen Deutsch bzw. Polnisch als Muttersprache.

2.1. Deutsch

2.1.1. Experimentdesign

Die Produktionen der deutschen Items wurden mittels Definitionen und Lückensätzen der Form „Tor zum Sauerland und männlicher Vorname (für Item *Hagen*) oder „Das kann nicht stimmen! Wo ist der ...?“ (für Item *Haken*) erhoben, die den Versuchspersonen auf einem Bildschirm präsentiert wurden. Nachdem die Versuchspersonen zunächst das gesuchte Wort isoliert produzieren sollten, wurden sie aufgefordert, das Wort in den Trägersatz „Ich habe ... gesagt“ einzubetten.

2.1.2. Items

Insgesamt wurden 24 Wörtern (12 Minimalpaare) der Form *Boden-Boten* oder *Hagen-Haken* elizitiert, von denen zwei von der akustischen Analyse ausgeschlossen wurden, da die Vokalisierung von /r/ in *Härte/Herde* nicht vergleichbar war mit den anderen Items.

2.1.3. Akustische Analyse

Die Analyse der Vokallänge des betonten Vokals erfolgte mittels PRAAT [7]. Grundlage waren dabei ausschließlich jene Items, die im Kontext des Trägersatzes gesprochen wurden, um eine möglichst natürliche Aussprache zu gewährleisten. Als Anfangs- und Endpunkt eines Vokals wurde stets der Beginn beziehungsweise das Ende der periodischen Schwingung gewählt. Die Vokallängen der einzelnen Items wurden mit Hilfe eines PRAAT-Skripts extrahiert.

2.1.4. Ergebnis

Abbildung 1 zeigt die durchschnittliche deutsche Vokallänge vor stimmhaften und stimmlosen Konsonanten.

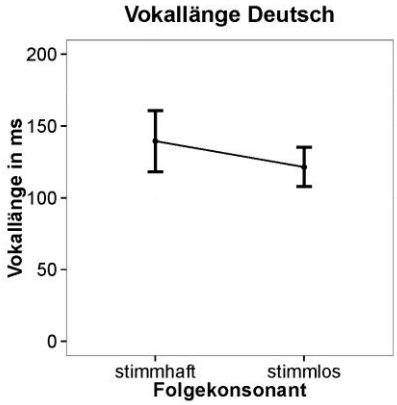


Abbildung 1: Vokallänge (in ms) vor stimmhaften und stimmlosen Konsonanten im Deutschen. Fehlerbalken zeigen zwei Standardfehler.

Im Durchschnitt waren deutsche Vokale vor stimmlosen Konsonanten 121.4 ms lang, Vokale vor stimmhaften Konsonanten um 14.9 % länger. Pro Versuchsperson wurden die Vokallängen vor stimmhaften versus stimmlosen Konsonanten gemittelt und mit Hilfe gepaarter t-test geprüft, ob die Differenz signifikant ist. Dies war der Fall bei einer Wahrscheinlichkeit von $p=0.01$ ($t(5)=3.9$).

2.2. Polnisch

2.2.1. Experimentdesign

Die Produktionen der polnischen Items wurden mittels Wortliste elizitiert. Die Probandinnen und Probanden wurden dabei aufgefordert, so schnell wie möglich zu sprechen, um eine starke Explizitlautung zu vermeiden.

2.2.2. Items

Während im deutschen Experiment seltene Wörter aufgrund des Designs ausgeschlossen werden mussten, konnte im polnischen Experiment eine größere Anzahl von Minimalpaaren elizitiert werden. Insgesamt wurden 58 Wörter (29 Minimalpaare) der Form *nuta-nuda* randomisiert als Einzelwörter produziert. Zwei Minimalpaare wurden von der akustischen Analyse ausgeschlossen, da sie einen Nasalvokal an der kritischen Stelle beinhalteten, welche von einigen Versuchspersonen entweder als nasalierter Monophthong oder als /Vn/ produziert wurden. Die akustische Analyse erfolge wie bei den deutschen Items.

2.2.3. Ergebnis

Abbildung 2 zeigt die durchschnittliche polnische Vokallänge vor stimmhaften und stimmlosen Konsonanten. Im Durchschnitt waren polnische Vokale vor stimmlosen Konsonanten 98.7 ms, Vokale vor stimmhaften Konsonanten 122.1 ms lang. Wie im deutschen Experiment war dieser Gruppenunterschied signifikant ($t(5)=7.49$, $p<0.001$).

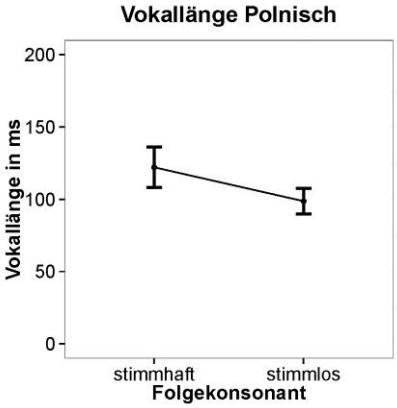


Abbildung 2: Vokallänge (in ms) vor stimmhaften und stimmlosen Konsonanten im Polnischen. Fehlerbalken zeigen zwei Standardfehler.

3. Diskussion

Unsere Daten untermauern Chens [5] Hypothese, dass Vokale – sprachübergreifend – vor stimmhaften Konsonanten länger sind als vor stimmlosen und widerlegen Keatings Befund zum Polnischen [6]. Der Effekt ist in beiden Sprachen zudem auffällig ähnlich, obgleich sich das Deutsche und das Polnische phonetisch/ phonologisch nicht sehr ähneln. Chen erklärt dieses sprachuniverselle Phänomen mit höherem intraoralen Druck bei stimmlosen Konsonanten und den entsprechend kürzeren Lautverschlussübergangzeiten zwischen Vokal und Konsonant. Andere Erklärungsansätze [6, 8] fokussieren auf einen möglichen Längenausgleich innerhalb der Silbe bzw. eines Wortes, da stimmlose Konsonanten längere Verschlussdauern aufweisen als stimmhafte. Allerdings wurde diese Hypothese von Chen zum Englischen bereits getestet und verworfen. Da Keating diesbezüglich von einem Unterschied zwischen akzent- und silbenzählenden Sprachen ausgeht, könnte es interessant sein, letztere Hypothese mit den vorliegenden Daten mittels Sekundäranalyse zu überprüfen.

4. Bibliographie

[1] Slowiaczek, L. M.; Dinnsen, D. A. (1985): On the neutralizing status of Polish word-final devoicing. In *Journal of Phonetics* (13), 325–341.

[2] Slowiaczek, L. M.; Szymanska, H. J. (1989): Perception of word-final devoicing in Polish. *Journal of Phonetics* (17), 205–212.

[3] Port, R. F.; O'Dell, M. L. (1985): Neutralization of syllable-final voicing in German. In Journal of Phonetics (13), 455–471.

[4] Port, R. F.; Crawford, P. (1989): Incomplete neutralization and pragmatics in German. *Journal of Phonetics* (17), 257–282.

[5] Chen, M. (1970): Vowel length variation as a function of the voicing of the consonant environment. In *Phonetica* (22), 129–159.

[6] Keating, P. (1984): Universal phonetics and the organization of grammar. In *Working Papers in Phonetics* (59), 35–49.

[7] Boersma, P.; Weenink, D. (2014): *Praat: doing phonetics by computer* [Computer program]. Version 5.4.04.

[8] Kohler, K,; Krützmann, U.; Reetz, H.; Timmermann, G. (1982): Sprachliche Determinanten der signalphonetischen Dauer. In: William Barry und Klaus Kohler (Hrsg.): *Experimentelle Untersuchungen zur Lautdauer im Hoch- und Niederdeutschen*. Kiel: Institut f. Phonetik (Arbeitsberichte 17), S. 3–48.

Wer die Qual hat, hat keinen Wal: Orthographische Effekte bei der Produktion deutscher Vokale

Katharina Nimz, Katharina Immel, Kai Ole Koop

Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld

{katharina.nimz,katharina.immel,kkoop}@uni-bielefeld.de

Abstract

In dem vorliegenden Pilotprojekt wurde untersucht, ob und inwiefern sich die explizite orthographische Markierung von Vokallänge (z. B. Dehnungs-*h*) auf die Produktion deutscher Langvokale ausübt. In einem Produktionsexperiment mit neun deutschen MuttersprachlerInnen wurden mittels Definitionen und Lückensätzen (ohne orthographischen Input) die Produktionen von 16 heterographen Minimalpaare erhoben (z. B. *Wahl* und *Wal*), deren Vokallänge mittels PRAAT ermittelt wurde. Obgleich diese Paare als homophon gelten, zeigte sich ein signifikanter Längenunterschied zwischen explizit markierten und unmarkierten Testitems: Vokale werden ca. 8% länger artikuliert, wenn sie orthographisch markiert sind.

Schlüsselbegriffe: Vokalproduktion, orthographische Effekte, Vokallänge

1. Einleitung

An der Schnittstelle zwischen Phonetik und Zweitspracherwerb zeichnet sich seit einigen Jahren ein wachsendes Interesse an dem Einfluss von Orthographie auf die Produktion und Perzeption von Lauten in der Zweitsprache ab [1-3]. Was den Einfluss der Orthographie auf die (Laut-)Produktion in der Muttersprache betrifft, liegen bisher wenig Daten vor [4]. Insbesondere für das Deutsche ist die Datenlage spärlich. Zwar untersuchte [5] den Einfluss des Dehnungs-*h* bei der Vokalproduktion polnischer DeutschlernerInnen und einer deutschen Kontrollgruppe, den MuttersprachlerInnen galt aber nicht das Hauptinteresse der Untersuchung. In der deutschen Kontrollgruppe zeigten sich dennoch unerwartete orthographische Effekte: Durch das Dehnungs-*h* markierte Vokale (z. B. in *Sahne*) waren länger als jene in nicht markierten Wörtern (z. B. in *Gabel*). Allerdings wurde in diesem Experiment die konsonantische Lautumgebung nicht kontrolliert. Beispielsweise waren in der Gruppe der orthographisch markierten Wörter – bedingt durch das Auftreten des Dehnungs-*h* ausschließlich vor Sonoranten – alle Folgekonsonanten stimmhaft, während in der Gruppe der unmarkierten Wörter wesentlich weniger Konsonanten stimmhaft waren. Dies ist problematisch, da man weiß, dass in den meisten Sprachen der Welt Vokale vor stimmhaften Konsonanten länger sind als vor stimmlosen [6]. Das vorliegende Pilotprojekt setzt hier an und untersucht Vokallängenunterschiede in Wörtern mit vergleichbaren konsonantischen Umgebungen, wobei die Vokale entweder explizit orthographisch markiert oder unmarkiert sind. Mit „explizit“ markiert referieren wir auf Dehnungsgraphien wie <h>, <ie> oder Doppelvokale.

2. Produktionsexperiment

Zur Untersuchung des Einflusses von orthographischer Markierung wurde ein Produktionsexperiment durchgeführt, in dessen Verlauf Probandinnen und Probanden heterographe Minimalpaare produzieren sollten.

2.1. Experimentdesign

In dem Experiment wurde bewusst darauf verzichtet, orthographischen Input zu geben, damit mögliche Unterschiede in der Vokallängenproduktion nicht allein auf den Leseprozess reduziert werden können. Die Produktionen der Items wurden daher mittels Definitionen und Lückensätzen der Form „Großes Meeressäugetier“ (für Item *Wal*) oder „Wer die Qual hat, hat die …“ (für Item *Wahl*) erhoben, die den Versuchspersonen auf einem Bildschirm präsentiert wurden. Nachdem die Versuchspersonen zunächst das gesuchte Wort isoliert produzieren sollten, wurden sie aufgefordert, das Wort in den Trägersatz „Ich habe … gesagt“ einzubetten.

2.2. Versuchspersonen

Insgesamt wurden 9 Versuchspersonen analysiert. Alle sprachen Deutsch als Muttersprache und hatten einen Universitätsabschluss. Letzteres wurde als Kriterium festgelegt, da man sicherstellen wollte, dass die Versuchspersonen der deutschen Orthographie mächtig sind und ihnen auch orthographisch schwierige Wörter wie beispielsweise *Waagen* oder *Mine* bekannt sind.

2.3. Items

Insgesamt wurden 32 Wörter (16 Minimalpaare) elizitiert. Die ursprüngliche Wortliste umfasste vier zusätzliche Paare, die allerdings nicht erhoben wurden, da der Folgekonsonant ein lateraler Approximant war (z. B. *Stil* versus *Stiel*) [7]. Von den übrigens 16 Minimalpaaren wurden zwei durch Doppelvokale unterschieden (z. B. *Wagen* versus *Waagen*), vier durch das Vorhandensein von <ie> (z. B. *Mine* versus *Miene*) und die restlichen durch Dehnungs- (*Wal* versus *Wahl*) beziehungsweise silbentrennendes *h* (*Rute* versus *ruhte*). Obgleich das silbentrennende *h* primär eine andere Funktion hat als das Dehnungs-*h*, wurden die Items dennoch aufgenommen, da diese orthographische Markierung ebenfalls ausschließlich nach Langvokalen auftritt. Auch [8] fasst das silbentrennende *h* zu den Dehnungsgraphien.

2.4. Akustische Analyse

Die Analyse der Vokallänge erfolgte mittels der Software PRAAT [9]. Grundlage waren dabei ausschließlich jene Items, die im Kontext des Trägersatzes gesprochen wurden, um eine möglichst natürliche Aussprache zu gewährleisten. Als

Anfangs- und Endpunkt eines Vokals wurde stets der Beginn beziehungsweise das Ende der periodischen Schwingung gewählt. Die Vokallängen der einzelnen Items wurden mit Hilfe eines PRAAT-Skripts extrahiert.

2.5. Ergebnis

Abbildung 1 zeigt die durchschnittlichen Längen für markierte und unmarkierte Vokale.

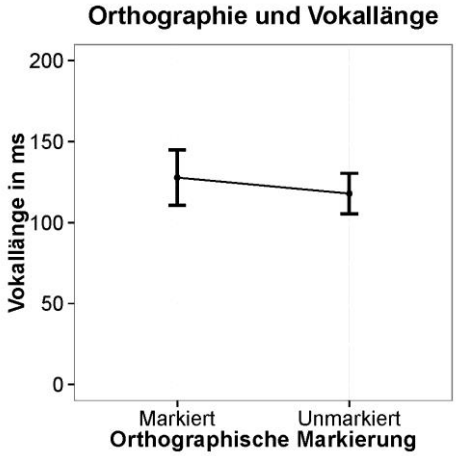


Abbildung 1: Vokallänge (in ms) bei orthographisch markierten und unmarkierten Wörtern. Fehlerbalken zeigen zwei Standardfehler.

Im Durchschnitt waren orthographisch unmarkierte Vokale 117.9 ms lang, markierte Vokale um 8.4 % länger (127.8 ms). Pro Versuchsperson wurde über die markierten bzw. unmarkierten Items gemittelt und mittels gepaartem t-test geprüft, ob diese Differenz signifikant ist. Da der Effekt bei jeder Versuchsperson aufgetreten ist, d. h. jede Versuchsperson markierte Vokale länger artikuliert hat als unmarkierte Vokale, ist der relativ kleine Unterschied von 9.9 ms signifikant ($t(8)=3.48$, $p=0.004$).

3. Diskussion

Da bei dem Experiment darauf verzichtet wurde, die orthographischen Formen der Testitems zu präsentieren, kann dieser Effekt nicht allein auf den Leseprozess zurückgeführt werden, d. h. eine Art hyperkorrekte Aussprache [10, S. 189] der sprachlichen Formen ist hier unwahrscheinlich. Psycholinguistische Studien haben gezeigt, dass die geschriebene Form eines Wortes einen Einfluss auf seine auditive Wahrnehmung haben kann („orthographic consistency effect“, z. B. [11]). Dass die orthographische Form eines Wortes ebenfalls Einfluss auf seine Produktion haben kann, ist vor diesem Hintergrund möglicherweise weniger überraschend, wenn auch in Bezug auf das deutsche Dehnungs-*h* und <ie> schwieriger erklärbar. Ein Erklärungsansatz könnte in der Schwere der Schreibsilbe liegen: Mehr Grapheme (in den explizit markierten Wörtern) könnten suggerieren, dass das Wort – in Analogie zu anderen Wörtern mit ähnlich langen Schreibsilben – länger ist als ein Wort, das durch weniger Grapheme wiedergegeben wird. Für die weitere Analyse könnte es daher von Interesse sein, nicht nur die Vokallänge zu untersuchen, sondern auch die Länge der umgebenden Konsonanten.

Des Weiteren wäre es interessant zu überprüfen, ob der signifikante, aber relativ kleine Längenunterschied wahrnehmbar ist. Zum Beispiel untersuchten [12] *incomplete neutralization* in deutschen Nonsens-Wörtern mit und ohne Auslautverhärtung und stellten fest, dass selbst ein relativ kleiner Längenunterschied von durchschnittlich 8 ms überdurchschnittlich häufig perzipiert werden kann. Sollte dies auch der Fall für die vorliegenden Daten, wäre eine andere Erklärung der Daten möglich und es müsste diskutiert werden, ob den Längenunterschieden eine kommunikative Funktion zugrunde liegt.

4. Dank

Wir danken Matthias Schrumpf für seine Mitarbeit bei der Datenerhebung und Annotation.

5. Bibliographie

[1] Escudero, P.; Hayes-Harb, R.; Mitterer, H. (2008): Novel second-language words and asymmetric lexical access. In *Journal of Phonetics* (36), 345–360.

[2] Hayes-Harb, R.; Nicol, J.; Barker, J. (2010): Learning the phonological forms of new words: Effects of orthographic and auditory input. In *Language and Speech* (53), 367–381.

[3] Bassetti, B.; Escudero, P.; Hayes-Harb, R. (2015): Second language phonology at the interface between acoustic and orthographic input. In *Applied Psycholinguistics* (36), 1–6.

[4] Alario, F.-X.; Perre, L.; Castel, C.; Ziegler, J. C. (2007): The role of orthography in speech production revisited. In Cognition (102), 464–475.

[5] Nimz, K. (2016): Die Rolle der Orthographie in der fremdsprachlichen Lautperzeption und -produktion. In *Potsdam Cognitive Science Series* (9). Potsdam: Universitätsverlag.

[6] Chen, M. (1970): Vowel length variation as a function of the voicing of the consonant environment. In *Phonetica* (22), 129–159.

[7] Turk, A. E.; Nakai, S.; Sugahara, M. (2006): Acoustic segment durations in prosodic research: A practical guide. In Sudhoff, S.; Lenertová, D.; Meyer, R.; Pappert, S.; Augurzky, P.; Mleinek, I.; Richter, N.; Schliesser, J. (Hrsg.): *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, 1–26.

[8] Eisenberg, P. (2013): *Das Wort. Grundriss der deutschen Grammatik*. 4. Auflage. Stuttgart: Metzler.

[9] Boersma, P.; Weenink, D. (2014): *Praat: doing phonetics by computer* [Computer program]. Version 5.4.04.

[10] Ternes, E. (2012): *Einführung in die Phonologie*. 3. Auflage. Darmstadt: Wissenschaftliche Buchgesellschaft.

[11] Petrova, A.; Gaskell, M. G.; Ferrand, L. (2011): Orthographic consistency and word-frequency effects in auditory word recognition: New evidence from lexical decision and rime detection. In *Frontiers in Psychology* (2), 1–11.

[12] Röttger, T. B.; Winter, B.; Grawunder, S. (2011): The robustness of incomplete neutralization in German. In Lee, W.-S.; Zee, E. (Hrsg.): *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong: Department of Chinese, Translation and Linguistics, University of Hong Kong, 1722–1725.

Vocal Fry – A Marker of Sophistication or Stupidity?

Amra Odobasic

Ruprecht-Karls-Universität Heidelberg

amra.odobasic@alumni.uni-heidelberg.de

Abstract

The purpose of this paper is to give an overview of vocal fry and discuss the ambiguity regarding its evaluation. Formerly classified as part of a clinical voice disorder, vocal fry (or creaky voice, pulse register, glottalization, etc.) appears to be omnipresent in American English today.

Its acoustic characteristics include a combination of rapid and short glottal pulses, as well as a very low fundamental frequency. The sporadic use of vocal fry has been found to have structural and pragmatic purposes. It is commonly used as a marker of phrase and sentence boundaries and, in RP, indicates the end of a speaker's turn in conversation.

Sociolinguistically, it might also be a cue to speaker identification and a gender marker. Studies show that even though young men have begun speaking in vocal fry more frequently, a vast quantity of young women already use it on a regular basis. This fact has been cause to rising polemics in newspaper articles and online videos: Vocal fry has been referred to as an "epidemic", making the female speaker appear uneducated, annoying, and shallow.

While the mainstream opinion seems to be rather negative, there are studies that have shown that vocal fry is also evaluated as a sophisticated trait, being associated with high social status and authority and making the female speaker seem "professional", "urban" and "upwardly mobile".

Index Terms: vocal fry, gender, social evaluation

1. Introduction

Vocal fry, also known as glottal fry, glottalization, pulse phonation, or creaky voice is a type of phonation, i. e. movement of the vocal folds, with an irregular laryngeal vibratory pattern. According to Hollien (1974) it is one of three vocal registers (or modes of laryngeal vibrations), which are distributed along a frequency-pitch continuum. Falsetto is considered the highest, modal register the normal register for speaking and singing, and vocal fry register is the lowest. [1]

Blomgren et al. (1998) have characterized vocal fry in terms of acoustics, aerodynamics, physiology and perceptual properties. Acoustically, they have found that the average fundamental frequency (F_0) of vocal fry ranges from 20 to 70 Hz, the mean being 50 Hz. In comparison, modal register F_0 ranges from 100-140 Hz for men and 175-240 Hz for women (2649). Surprisingly, the F_0 for vocal fry is very similar for both genders, other than in modal and falsetto register. The average airflow for vocal fry ranges from 2.0 to 71.9ml/s, with there being no proportional increase of airflow rate and vocal fry frequency, which has been experienced with modal register. Physiologically, the chronological use of slow motion films, ultra high-speed motion pictures and electroglottography (EGG) has shown that vocal fry possesses a specific dicrotic vibratory pattern, which suggests that the vocal folds rapidly open and close twice per cycle, causing brief glottal pulses, while finally remaining completely adducted for a relatively long period of time. With regard to perception, vocal fry is said to sound like the popping of corn [2] or, as Dr. Nicole Hardman claims in a Youtube video, bacon "sizzling in a pan". [3]

2. History

Abdelli-Beruh et al. point out that vocal fry was examined particularly in speech-language pathology, where it was typically considered a vocal disorder due to its often co-occurrence with other signs of abnormal laryngeal outputs. [4] Still, even as early as 50 years ago Hollien et al. (1966) were of the opinion that the sporadic use of vocal fry was not a sign of a vocal pathology, but rather a normal physiological mode of laryngeal vibrations, i. e. a register that speakers without a vocal pathology are able to deliberately switch from and to. [5] Even though vocal fry is said to be a common feature of English and is particularly found at the very end of sentences when pitch and vocal intensity are beginning to decay [6], there has been a noticeable increase in the use of young adult females and men for the last decades.

3. Gender distribution

Opinions differ with regard to the gender distribution of the use of vocal fry. While the studies of Abdelli-Beruh et al.and Yuasa suggest that it is more common for women, others scholars remain in opposition. The researchers around Abdelli-Beruh have tested 34 females and 34 male college students and found that for the reading task the women tended to use vocal fry four times more often than the men. Both genders used vocal fry predominantly at the end of a phrase, and particularly at the end of a sentence [4], which could imply that it is also used to mark syntactic boundaries. Although the discussions regarding the rationale behind this gender difference have not come arrived to a consensus, scholars seem to be in agreement that it is not due to anatomic difference, as could be deduced from the fact that women's vocal folds tend to be shorter than men's. [7]

Vocal fry has not always been associated with women, on the contrary: it connoted masculinity and authority. Henton & Bladon called vocal fry "hyper-masculine" and a "robust marker of male speech." [8] Yuasa suggests that "creaky voice may provide a growing number of American women with a way to project an image of accomplishment […] while retaining female desirability", as Yuasa points out that vocal fry is associated with a sexual desirability of American women. For her study, Yuasa recorded conversations among male and

females. The analysis showed that 12.4% of the words uttered by American females contained vocal fry, 6.9% of the words spoken by Japanese females and 5.6% of the words uttered by American males. [9] Henton & Bladon have found that in Received Pronunciation, men were to glottalize final syllables more often than women [8] so that one could come to the conclusion that vocal fry might be used as a means of speaker identification. [10]

4. Ambiguity in attitude

Many newspaper articles and Youtube videos have described vocal fry an "epidemic" [11] "a beast of [a] virus" [12], and "a female fad". [13] Anderson et al. conclude from their study that young women using vocal fry appear "less competent, less educated, less trustworthy, less attractive, and less hirable". [14] A 1986 report stated that in Sydney, vocal fry is used more than twice as often by young generations than by older ones, and particularly by women, which may be a possible explanation to the negative evaluation of vocal fry, particularly in job interviews. [15]
Yet, the study of Yuasa suggests exactly the opposite: female vocal fry is associated with being "educated", "professional", "urban" and "upwardly mobile". [9]

5. Discussion

Regarding why vocal fry seems sophisticated to some, there is a theory that it imitates the lower male voice and speech, which, according to Kramer, is associated with "demanding, dominating, militant, [and] authoritarian" traits. [16] Another possibility is that it counteracts the stereotype of high and shrill women's voices that could be related to that of the "hysterical woman". With vocal fry, women may attempt to audibly distance themselves from this by rather emulating the vocal attributes socially recognized as calm and sociable.

Why is vocal fry frequently described in a negative manner? One possible answer could be that the negative attitude toward vocal fry may be due to generational divides. Penelope Eckert, a linguistics professor at Stanford, has stated that while the older generation may disapprove of vocal fry, millennials may perceive it as natural. [17] A second hypothesis is that the negative perception derives from an excessive use of vocal fry, manifesting itself in a high word-to-vocal fry ratio. A third potential explanation may be a combination of vocal fry and high rising terminal ("uptalk"), leading to the conclusion that vocal fry per se is not the cause of negative attributions.

6. Method

My dissertation concentrates on the combination of vocal fry and high rising terminal, which is regularly linked to the stereotype of the shallow "valley girl" and presented as the main reason for the negative perception of vocal fry. Research is conducted by means of spoken corpora, radio, talk and reality shows to trace out under what circumstances, to what extent and by who vocal fry is used, and how it is perceived.
In a first study, participants will be recorded while reading text passages and talking to both peers and non-peers, as Wolk et al. suggest. [18] It is important to include both types of tasks as the reading task provides insight into the linguistic use of vocal fry while the conversations offer information about its social component.

The second study aims at comprehending the attitudes towards vocal fry. Listeners are therefore asked to fill out questionnaires after listening to samples from a first group of participants who speak in modal register, a second group that uses vocal fry and a third group of participants speaking with vocal fry in combination with high rising terminal.

The focus of further research could lie on the investigation of the use of vocal fry with children and middle-school teenagers. To receive a diachronic perspective, these results could then be compared to corpus data of groups of similar age from the last decades.

7. Conclusion

This paper sought to enhance present knowledge of phonation and provide an overview regarding the state of the art of vocal fry, while also focusing on its social parameters. The discussion about gender illustrated that the increased use of vocal fry by young women is loosening its former association with masculinity in favor of a stronger female component. Furthermore, the potential origin of the ambiguous perception and evaluation of vocal fry was elaborated upon, indicating that negative connotations derive from a combinatory use of vocal fry and high rising terminal.

8. Acknowledgements

I would like to thank Hannah Schachter, who provided assistance and valuable comments on earlier versions of this paper.

9. References

[1] H. Hollien (1974), "On vocal registers", J. Phon. 2. 25-43.
[2] Michael Blomgreen, Yang Chen, Manwa L. Ng, and Harvey R. Gilbert (1998), "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers", J. Acoust. Soc. Am. 103 (5), Pt. 1, May issue. 2649-2658.
[3] MightyFalcon2011, "Kardashians' 'vocal fry' fad's effect on the voice, <https://www.youtube.com/watch?v=txSSO4VglMI>, last checked September 15, 2016.
[4] Nassima B. Abdelli-Beruh, Lesly Wolk, and Dianne Slavin (2014), "Prevalence of vocal fry in young adult male American English speakers", Journal of Voice, 28 (2). 185-190.
[5] H. Hollien, P. Moore, and RW Wendahl (1966), "On the nature of vocal fry", J Speech Hear res. 9. 245-247.
[6] W. R. Zemlin (1988), *Speech and Hearing Science: Anatomy and Physiology.* Prentice-Hall, Englewood Cliffs, NJ.
[7] D. Byrd (1994), "Relation of sex and dialect to reduction", Speech Commun. 15. 39-54.
[8] C. G. Henton and A. Bladon (1988), "Creak as a sociophonetic marker", *Language, speech and mind: studies in honour of Victoria A. Fromkin* (L. Hyman and C. Li, editors), London: Routledge. 33-29.
[9] Ikuko Patricia Yuasa (2010), "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?", American Speech 85 (3), fall issue. 315-337.
[10] T. Bohm and S. Shattuck-Hufnagel (2007), "Utterance final glottalization as cue for familiar speaking recognition", *Interspeech-2007*. 2657-2660.
[11] Abby Normal, "The vocal fry epidemic",

<https://www.youtube.com/watch?v=UsE5mysfZsY>, last checked: September 15, 2016.

[12] Ann Hornaday, "Lake Bell talks about 'In a world…' and the politics of dialect", <https://www.washingtonpost.com/goingoutguide/movies/lake-bell-talks-about-in-a-world--and-the-politics-of-dialect/2013/08/08/71eb5ed0-ff76-11e2-96a8-d3b921c0924a_story.html>, last checked: September 15, 2016.

[13] Katy Steinmetz, "Get your crack on: Is vocal fry a female fad?", <http://healthland.time.com/2011/12/15/get-your-creak-on-is-vocal-fry-a-female-fad/>, last checked: September 15, 2016.

[14] Rindy C. Anderson, Casey A. Klofstad, William J. Mayew, and Mohan Venkatachalam (2014), "Vocal fry may undermine the success of young women in the labor market", PLOS ONE 9 (5) May issue. 1-8.

[15] G. Guy, B. Horvarth, J. Vonwiller, E. Dasiley, I. Rogers (1986), "An intonation change progress in Australian English." Language in Society 15. 23-52.

[16] C. Kramer (1975), "Excessive loquacity: women speech as represented in American etiquette books", paper presented at the Annual Meeting of Speech Communication Association, Austin, TX. 3-15.

[17] Jan Hoffman, "Overturning the myth of valley girl speak", <http://well.blogs.nytimes.com/2013/12/23/overturning-the-myth-of-valley-girl-speak/?_r=0>, last checked: September 15, 2016.

[18] Lesly Wolk, Nassima B. Abdelli-Beruh, and Dianne Slavin (2012), "Habitual use of vocal fry in young adult female speakers", Journal of Voice 26 (3). e111-e116.

Sprachdatenerhebung und Kontextvariation

Wie die Intonation in Fragen vom kommunikativen Kontext gesteuert wird

Benno Peters [1], *Matthias Hoffmann* [2], *Laura-Marie Andresen* [3]

ISFAS -Abteilung für Allgemeine Sprachwissenschaft, CAU Kiel [1,2,3]

peters@ipds.uni-kiel.de, stu107929@mail.uni-kiel.de, laura.andresen32@web.de

Abstract

Phonetische Strukturen von Äußerungen unterscheiden sich in Abhängigkeit von Gesprächssituationen und interpersonellen Beziehungen. So können sich Gesprächspartner ihrem Gegenüber unterordnen oder sich in die dominante Position begeben. Eine solche Positionierung zeigt sich unter anderem in der Intonation von Fragen.

Die systematische empirische Untersuchung solcher Phänomene braucht eine geeignete Datengrundlage. Wir benötigen Sprachdaten aus Situationen, in denen sich Gesprächspartner interaktional klar positionieren. Die Methode gut analysierbare Zielsätze in interaktional steuernde Kontexte einzubetten ist geeignet, um systematisch Daten in gelesener Sprache zu erheben, die ebenfalls Aufschlüsse über Alltagskommunikation geben können.

Folgendes Beispiel zeigt wie der Zielsatz *Was sollen wir machen?* (a) in einem sich unterordnenden Kontext eingebettet ist: *Ich weiß nicht weiter. Was sollen wir machen?* und (b) auf einen initialen Kontext folgt, der Dominanz signalisiert: *Na toll. Du hast das kaputt gekriegt. Was sollen wir machen?*

Eine solche Einbettung resultiert in sich klar unterscheidenden Verteilungen phrasenfinaler Intonationsmuster in den verschiedenen Kontexten. Die Satzintonation fällt typischerweise in dominanten Fragen und steigt in sich unterordnenden Fragen. Dieser Beitrag stellt die Funktionsweise und das Potential der Methode der Kontextvariation am Beispiel der Untersuchung der phrasenfinalen Intonation in W-Fragen im Deutschen dar.

Schlüsselbegriffe: Datenerhebung, Sprachkorpora, Prosodie, Frageintonation, Pragmatik

1. Einleitung

Dieser Beitrag zeigt, wie eng die Verbindung zwischen dem phonetischen Signal, dem kommunikativen Kontext und der Übermittlung attitudinaler sowie affektiver Bedeutung ist. Es wird der Einfluss von Dominanz und Unterordnung auf die Intonation in strukturell und inhaltlich vergleichbaren W-Fragen untersucht.

Für die Untersuchung von phonetischen Strukturen in unterschiedlichen interaktionalen Konstellationen benötigen wir vergleichbare Äußerungen aus Situationen, in denen sich Gesprächspartner interaktional klar positionieren, sich also z.B. deutlich in eine dominante oder sich unterordnende Position zum Hörer begeben. Solche Sprachdaten verschiedener Sprecherinnen und Sprecher zu erheben bedarf einer hohen Kontrolle über die Aufnahmesituation. In den bekannten spontansprachlichen Korpora lassen sich nicht genügend vergleichbare Äußerungen finden, um quantitative Analysen durchführen zu können. Mit der Methode der Kontextvariation stellen wir eine Möglichkeit vor, mit der solche Fragen kontrolliert elizitiert werden können, um ein systematisches Korpus aufzubauen.

Im Verfahren der Kontextvariation nutzen wir gelesene Texte, um ein hohes Maß an Kontrolle sowohl über situative Kontexte als auch über Zielsätze zu gewährleisten. Das bedeutet, dass ein konstanter Zielsatz in verschiedenen Gesprächssituationen geäußert wird, denen gänzlich unterschiedliche Einstellungen der Gesprächspartner zugrunde liegen. Unterschiedliche Dimensionen der Einstellung können sein: *sich unterordnend vs. dominant*, *höflich vs. unhöflich*, *gelassen vs. genervt*, *fragend vs. befehlend*, usw.

Anhand der konkreten Untersuchung der Intonation in W-Fragen im Deutschen soll die Funktionsweise und der starke Effekt von Kontextvariation dargestellt werden. Wir zeigen welchen Einfluss der Kontext und die Einstellung des Sprechers zum Hörer auf die Intonationskontur von Fragen besitzt.

Fragen sind als sprachliche Handlungen funktional in kommunikative Kontexte eingebettet. Die Einstellung und Intention [1] sind eng verbunden mit der Positionierung von Sprecher zu Hörer: Neben einer Kommunikation auf Augenhöhe kann sich der Sprecher - unabhängig von seiner Persönlichkeit und Disposition - in eine dominante oder sich dem Hörer unterordnende Position begeben [2]. Fraglich ist, was die intonatorischen Korrelate dieser sozialen Konzepte von Dominanz und Unterordnung sind.

Ohala [3] assoziiert mit seinem ethologisch geprägten Konzept des *Frequency Code* Intonation und Tonhöhe in lautsprachlichen Äußerungen mit der Übermittlung sozialer Funktionen wie Dominanz und Unterordnung. Nach Ohala sind eine global hohe bzw. steigende Grundfrequenz mit dem Konzept von Unterordnung verknüpft, während Dominanz durch eine global tiefe oder fallende Grundfrequenz ausgedrückt wird. Die sprachübergreifende Bevorzugung steigender Konturen in Fragen erklärt Ohala aus einem informationsorientierten Standpunkt. Der Fragende erhofft sich eine positive Antwort und ist auf die Kooperationsbereitschaft des Gegenübers angewiesen. Mit Hilfe eines semantischen Differentials zeigt Uldall [4], dass Probanden eine hohe Grundfrequenz perzeptiv als unterwürfig und freundlich bewerten.

Nach Kohler [5] trägt das phrasenfinale Intonationsmuster in Fragen attitudinale und interaktionale Funktion und ist kein formales Kennzeichen des Satzmodus. Kohler unterscheidet zwischen einer hörerbezogenen final steigenden Kontur und einer sach- bzw. sprecherbezogenen fallenden Kontur. Mit

einer satzfinal steigenden Intonation ordnet sich der Sprecher dem Hörer unter und klingt höflich und bittend. Eine mit fallender Intonation versehene Frage deutet darauf hin, dass sich der Sprecher dem Hörer gegenüber in dominanter Position sieht und kann forsch und befehlend klingen.

Für unsere laufende Untersuchung *Attitude and Intonation in German Questions - Effects of dominance and submission on pitch*, in der Versuchspersonen in verschiedene Kontexte eingebettete Fragen sprechen, zeigt sich, dass die Signalisierung der intentionalen Funktion bzw. der Einstellung zum Gesprächspartner durch die Intonationskontur kodiert ist. Versuchspersonen passen die verwendeten melodischen Konturen der Einstellung an, die der gegebene Kontext auslöst. Das heißt W-Fragen weisen bevorzugt eine steigende phrasenfinale Intonation auf, wenn sich der Sprecher in einer sich unterordnenden Position sieht. Konträr dazu ist zu erwarten, dass ein Sprecher, der sich in dominanter Position befindet, bevorzugt fallende Intonationsmuster nutzt.

2. Kontextvariation

Sobald Sprachdaten in Laborsituationen erhoben werden, muss ein Kompromiss zwischen möglichst natürlich klingender Sprache sowie möglichst hoher Kontrolle über die Zielsätze erreicht werden [6].

Wir benötigen Situationen, in denen sich der Sprecher interaktional klar positioniert, sich also deutlich in eine dominante oder sich unterordnende Position zum Hörer begibt. Weiterhin sollten diese Situationen in inhaltlich und strukturell ähnlichen W-Fragen auftauchen, damit die Intonationskonturen vergleichend analysiert werden können. In spontansprachlichen Aufnahmen kann weder die Positionierung des Sprechers zum Hörer noch eine vergleichbare Struktur der Zielsätze kontrolliert werden. Daher stellen wir mit der Methode der Kontextvariation eine Möglichkeit vor solche Fragen systematisch zu elizitieren.

Wir steuern die gesamte Situation, indem wir identische W-Fragen in vorangestellte Kontexte einbetten. Diese initialen Kontexte bestehen jeweils aus einem bis zwei kurzen Sätzen und signalisieren verschiedene Einstellungen des Sprechers zum Hörer. Ein befehlender Kontext wie *Sag's mir endlich.* bewirkt, dass sich der Sprecher in eine dominante Position begibt. Ein Kontext wie *Ich brauch' mal Deine Hilfe.* wirkt entschuldigend bzw. bittend und lässt den Sprecher sich dem Gegenüber unterordnen. Innerhalb der unterschiedlichen Kontextkategorien werden damit vergleichbare illokutionäre Akte evoziert [7]. Folgendes Beispiel zeigt den Zielsatz *Was sollen wir machen?* in den zwei interaktional unterschiedlichen Kontexten.

- sich unterordnend:
  Ich weiß nicht weiter. Was sollen wir machen?
- dominant:
  Na toll. Du hast das kaputt gekriegt. Was sollen wir machen?

Im sich unterordnenden Kontext wird für die W-Frage eine Situation konstruiert, in welcher der Sprecher sich hilflos bittend an den Hörer wendet, da er sich Unterstützung erhofft. Im dominanten Kontext hingegen wirkt die W-Frage anschuldigend, da der Sprecher den Hörer damit konfrontiert, dass sich durch dessen Fehler eine komplizierte bis ausweglose Situation ergeben hat. Der Sprecher fordert vom Hörer eine Stellungnahme dazu.

3. Kontextsensitive Sprache

Jede sprachliche Handlung ist abhängig vom kommunikativen Kontext, in der diese stattfindet [8]. Wir nutzen dies aus, indem wir die kommunikative Situation durch die Methode der Kontextvariation selbst steuern. Sowohl der dominante als auch der sich unterordnende Kontext soll dabei innerhalb der beiden Kategorien jeweils vergleichbare Intentionen bei den Sprechern hervorrufen. Die initialen Kontexte wecken Konnotationen, die die Sprecher dazu bringen, sich zum Gegenüber in dominante oder sich unterordnende Position zu begeben. Beispielsweise wirkt der initiale Kontext *Na toll.* ironisch bis anklagend und verleitet den Sprecher dazu sich in die dominante Position zu begeben.

Intentionen gehen in natürlicher Kommunikation jeder sprachlichen Handlung voraus und werden u.a. durch prosodische Mittel zum Ausdruck gebracht. Intentionen sind dabei im Vergleich zu Emotionen eher gesellschaftlich verankert sowie kulturell konventionalisiert [9]. Auch kann der Ausdruck von Intentionen besser willentlich kontrolliert werden und ist folglich weniger abhängig von der Spontanität einer Äußerung [10]. Daher ist der Ausdruck von Intention weniger anfällig für Artefakte, die sich aus der Art der Datenerhebung und künstlichen Aufnahmesituationen ergeben, weil Sprecher unterbewusst und somit authentisch wissen, wie sie sich in einer Gesprächssituation kommunikativ angemessen zum Gegenüber positionieren können. Durch den beschriebenen Charakter von Intentionen können diese von Sprechern in künstlichen Situationen natürlicher präsentiert werden als Emotionen. Auch wurden bewusst keine Schauspieler aufgenommen, da diese dazu neigen, Intentionen mit antrainierten Mitteln unter Einbezug unterstützender geschauspielerter Emotion auszudrücken [11]. Unsere unvoreingenommenen Versuchspersonen hingegen vermitteln die Intentionen und Einstellungen gemäß eigener kommunikativer Strategien. Die genutzten Mittel zum Ausdruck der Intention und Positionierung zum Gesprächspartner ähneln damit denen in natürlichen Gesprächskontexten. Natürlich muss bei der Übertragung von Ergebnissen aus Lesesprache auf spontane, unbeobachtete Alltagssprache sehr vorsichtig vorgegangen werden.

4. Datenerhebung

Tabelle 1 zeigt die verwendeten Zielsätze und jeweils den dominanten sowie den sich unterordnenden Kontext, in den diese eingebettet werden. Die Zielsätze sind so konzipiert, dass sie sehr wahrscheinlich als einakzentige Äußerung gesprochen werden und jeweils auf ein zweisilbiges Wort enden. Übliche Reduktionsphänomene wie Schwa-Elisionen wurden in die orthographische Repräsentation integriert. Alle Äußerungen wurden zusammen mit Dummys sowie einem neutralen Kontext randomisiert. Die Dummy-Äußerungen bestehen aus einfachen Deklarativsätzen, die in Kontexte unterschiedlicher Konnotation eingebunden sind, um der Struktur der eigentlichen Stimuli zu entsprechen.

Es wurden insgesamt 30 Versuchspersonen - 15 Männer und 15 Frauen - aufgenommen. Alle Teilnehmer des Versuchs sprechen Standarddeutsch und waren zum Zeitpunkt der Aufnahme zwischen 20 und 32 Jahre alt (M = 26,3, SD = 3,06). Es handelt sich um Studenten und Angestellte der Christian-Albrechts-Universität zu Kiel. Dadurch, dass die vier Zielsätze in zwei Kontexten von 30 Sprechern realisiert wurden, ergeben sich insgesamt 240 Äußerungen.

Tabelle 1. *Sprachmaterial*

sich unterordnend	dominant	Zielsatz
Oh Mann, ich kann jetzt leider echt nicht mehr.	Das ist alles Deine Schuld. So kommen wir nie voran.	Wer macht jetzt weiter?
Tut mir leid, dass ich Dich gestern versetzt hab'. Dann brauchen wir ja jetzt wohl 'n neuen Termin.	Ich will das so schnell wie möglich vom Tisch haben.	Wie ist es denn mit Samstag?
Ich weiß nicht weiter.	Na toll. Du hast das kaputt gekriegt.	Was sollen wir machen?
Ich brauch' mal Deine Hilfe.	Sag's mir endlich.	Was ist hier anders?

Den Versuchspersonen wurde schriftlich die Instruktion gegeben, die Äußerungen so zu sprechen, wie sie in einer entsprechenden Situation im Freundeskreis vorkommen könnten. Nachdem einige Probesätze gesprochen wurden, um sich an die Aufgabe und Ausnahmesituation zu gewöhnen, wurde eine Liste mit den Stimuli ausgehändigt. Die Versuchspersonen wurden gebeten sich für jede Äußerung Zeit zu nehmen, sich dabei in die entsprechende Situation hineinzu-versetzen und die Äußerung dann situationsangemessen zu sprechen.

5. Auswertung und Ergebnisse

Für alle 240 Äußerungen wurde per Hand annotiert, ob der Zielsatz mit einer steigenden, fallenden oder ebenen Intonationskontur realisiert wurde. Es haben sich keine geschlechtsspezifischen und sprecherindividuellen Unterschiede gezeigt. Im sich unterordnenden Kontext finden sich zu ca. 70 % steigende und ca. 27 % fallende Konturen. Die Verteilung im dominanten Kontext ist etwa spiegelbildlich. Etwa 18 % der Äußerungen sind mit steigender Intonationskontur realisiert und ca. 73 % mit fallender Kontur (siehe Abbildung 1). Dieser Unterschied der Verhältnisse im dominanten und sich unterordnenden Kontext ist klar signifikant ($\chi 2$ (2, N = 240) = 69.781, p = <.001).

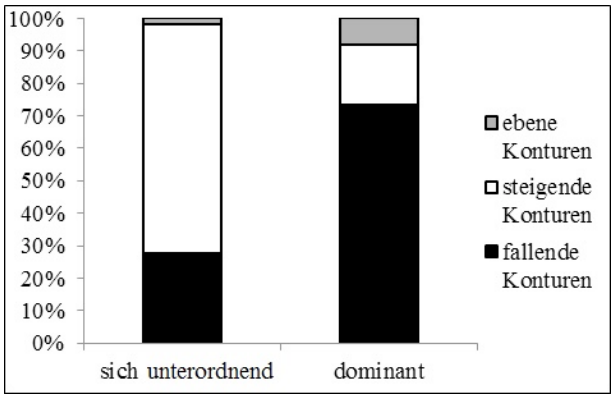


Abbildung 1: Prozentuale Anteile steigender, fallender und ebener Konturen im sich unterordnenden und dominanten Kontext

6. Fazit

Die Methode, konstante Zielsätze in initiale Kontexte einzubetten, ist geeignet, die Einstellung zum Gesprächspartner und die Signalisierung der intentionalen Funktion zu steuern. Die Versuchspersonen haben die kommunikativen Intentionen der Stimuli erkannt und produzieren im dominanten sowie sich unterordnenden Kontext klar voneinander zu differenzierende Intonationskonturen. Die durch die Methode der Kontextvariation elizitierten Sprachaufnahmen wirken dabei alltagsnah.

Die Ergebnisse gehen dabei konform mit dem Frequency Code [3] sowie Arbeiten, die sich mit der Funktion von Intonationsmustern in Fragen beschäftigen, wie Kohler [5] und Dombrowski und Niebuhr [12]. Auch Hellbernd und Sammler [10] zeigen, dass die Intention eines Sprechers durch prosodische Hinweise repräsentiert wird. Die Resultate stehen ebenfalls in Übereinstimmung mit verhaltensbiologischen Studien wie Puts et al. [13] und Hodges-Simeon et al. [14], die eine niedrige Grundfrequenz nicht nur mit physischer Kraft, sondern auch mit sozialer Dominanz assoziieren.

7. Diskussion und Ausblick

Lexikalisch und syntaktisch identische Äußerungen können allein durch die Veränderung des Grundfrequenzverlaufs gänzlich unterschiedliche kommunikative Funktion besitzen. Die Intonation von Fragen hängt substanziell von der kommunikativen Situation und damit von der Einstellung des Sprechers zum Hörer ab. Diese manifestiert sich in einer dominanten oder sich unterordnenden Position des Sprechers dem Gesprächspartner gegenüber und wird durch die Intonation einer Äußerung kodiert. Soziale Zeichen wie Dominanz und Unterordnung sind essentielle Bestandteile des Sprachsignals. Um ein phonologisch orientiertes System prosodischer Funktionen im Gesprächsprozess zu entwickeln, bedarf es demnach einer Verbindung des konkreten phonetischen Signals mit der kommunikativen Funktion unter Einbezug der durch die Intonationskontur kodierten intentionalen Funktion.

Die hier vorgestellte Methode der Datenerhebung findet ebenfalls Verwendung in unserer laufenden Untersuchung *Attitude and Intonation in German Questions - Effects of dominance and submission on pitch*. Zusätzlich zu den hier vorgestellten W-Fragen sind dort durch Verberststellung markierte VE-Fragen ebenfalls in initiale Kontexte eingebettet. Neben den hier vorgestellten dominanten und sich unterordnenden Kontexten, nutzen wir einen neutralen Kontext, der eine symmetrische Gesprächssituation zwischen Sprecher und Hörer evoziert.

8. Bibliographie

[1] H.P. Grice, "Meaning," *The Philosophical Review*, vol. 66, pp. 377–388, 1957.

[2] P. Watzlawick, J.H. Beavin and D.D. Jackson, *Menschliche Kommunikation – Formen, Störungen, Paradoxien*. Bern: Huber, 1969.

[3] J.J. Ohala, "An ethological perspective on common crosslanguage utilization of F0 of voice," *Phonetica*, vol. 41, pp. 1–16, 1984.

[4] E. T. Uldall, "Attitudinal meanings conveyed by intonation contours," *Language and Speech*, vol. 3, pp. 223–234, 1964.

[5] K. J. Kohler, "Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions," In G. Fant, H. Fujisaki, J. Cao and Y. Xu (eds), *From Traditional Phonology to Modern Speech Processing. Festschrift for Professor Wu Zongji's 95th Birthday*. Beijing: Foreign Language Teaching and Research Press, pp. 205–214, 2004.

[6] I. Lehiste, "Review of K. Hadding-Koch," *Language*, vol. 39, pp. 352–360, 1963.

[7] N. Himmelmann, "Prosody in Language Documentation," In J. Gippert, N.P. Himmelmann and U. Mosel (eds), *Essentials of Language Documentation*. Berlin: de Gruyter, pp. 163–181, 2006.

[8] O. Niebuhr and A. Michaud, "Speech data acquisition–: the underestimated challenge," Kieler Arbeiten in Linguistik und Phonetik (KALIPHO), vol. 3, pp 1–42, 2015.

[9] A. Rilliard, T. Shochi, J.-C. Martin, D. Erickson and V. Auberge, "Multimodal indices to Japanese and French prosodically expressed social affects," *Language and Speech*, vol. 52, pp. 223–243, 2009.

[10] N. Hellbernd and D. Sammler, "Prosody conveys speaker's intentions: Acoustic cues for speech act perception," *Journal of Memory and Language*, vol. 88, pp. 70–86, 2016.

[11] R. Jürgens, A. Grass, M. Drolet and J. Fischer, "Effect of acting experience on emotion expression and recognition in voice," *Journal of Nonverbal Behaviour*, vol. 39, pp. 195–214, 2015.

[12] E. Dombrowski and O. Niebuhr, "Acoustic Patterns and Communicative Functions of Phrase-Final F0 Rises in German: Activating and Restricting Contours," *Phonetica*, vol. 62, pp. 176–195, 2005.

[13] D.A. Puts, C.R. Hodges, R.A. Cárdenas and J.C. Gaulin, "Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men," *Evolution and Human Behaviour*, vol. 28, pp. 340–344, 2007.

[14] C.R. Hodges-Simeon, M. Gurven, D.A. Puts and J.C. Gaulin, "Vocal fundamental and formant frequencies are honest signals of threat potential in peripubertal males," *Behavioral Ecology*, vol. 25, no. 4, pp. 1–5, 2014.

Geflüsterte Angst und behauchte Trauer – Stimmqualität und Emotionen

Louise Probst, Angelika Braun

Universität Trier
s2lodirk@uni-trier.de, brauna@uni-trier.de

Abstract

Der Parameter Stimmqualität ist im Bereich der phonetischen Erforschung von Emotionen in der Stimme selten betrachtet worden. Die vorliegende Studie beschäftigt sich daher mit der Frage, welche Stimmqualitäten Sprecher zur Porträtierung von Emotionen einsetzen. Hierfür wurden emotionale und neutrale Sprachäußerungen von sechs professionellen Sprechern des Deutschen aufgezeichnet und hinsichtlich ihrer Stimmqualität beurteilt. Während die Vorhersagen für das Auftreten von breathy voice bei Freude zutreffen, lassen sich die Vorhersagen für harsh voice bei Wut nur teilweise und für creaky voice bei Trauer nicht bestätigen.

Index Terms: Emotionen, Stimmqualität, Perzeption

1. Einleitung

Obwohl sich viele Studien mit dem Thema Emotionen und Stimme/Sprechen beschäftigen und auch viele ein Augenmerk auf die akustische Untersuchung der Produktion von Emotionen legen, wird die Stimmqualität als Parameter nur bei wenigen betrachtet. Hier stellt sich die Frage, ob Sprecher nicht auch unterschiedliche Stimmqualitäten einsetzen, um Emotionen stimmlich zu markieren. Die wenigen Vorhersagen, die aus den Ergebnissen früherer Studien abgeleitet werden können, lassen bei Freude *breathy voice* [1-4], bei Trauer *creaky voice* [3] und bei Wut *harsh voice* [1, 3] sowie *tense voice* [4] erwarten, hinsichtlich anderer Emotionen liegen uneinheitliche Befunde vor. Nur wenige Studien differenzieren zwischen Heißer Wut und Kalter Wut, wofür Scherer [5] und Banse/Scherer [6] argumentieren. Diese Differenzierung wurde hier aufgegriffen, bestätigt sie sich doch in den akustischen Untersuchungen zur Produktion und auch in Ergebnissen von Perzeptionstests (vgl. z.B. [3]).

Eine durchgehende Schwäche vorausgegangener Studien besteht darin, dass nicht klar gesagt wird, welcher Grad an Emotionalität untersucht wurde. Die Zusammenstellung in Banse/Scherer [6, S. 617] macht jedoch deutlich, dass Unterschiede in Abhängigkeit vom Ausprägungsgrad zu erwarten sind. Auch dieser Gedanke wird in der vorliegenden Untersuchung aufgegriffen.

2. Material und Methode

Für die vorliegende Studie wurden emotional gefärbte sowie neutrale Äußerungen von sechs professionellen Sprechern des Deutschen aufgezeichnet. Sie wurden gebeten, sechs verschiedene Emotionen in drei Intensitätsstufen – gering, mittel und extrem – zu porträtieren. Als Sprachmaterial dienten fünf deutsche Nonsense-Sätze. Die vorgegebenen Emotionen waren: Angst, Trauer, Freude, Ekel, kalte Wut und heiße Wut. Die Sprachaufnahmen wurden in einem professionellen Studio des Westdeutschen Rundfunks (WDR) mittels eines Neumann U 87 und eines DHD-RM4200 Vorverstärkers aufgezeichnet. Die Abtastrate betrug 48kHz bei 16 bit Abtasttiefe. Die Sprecher produzierten die fragliche Emotion und den jeweiligen Ausprägungsgrad so lange, bis sie selbst damit zufrieden waren.

Die Stimmqualität wurde auditiv ermittelt. Als Grundlage dienten die gering und extrem ausgeprägten Stimuli, da auch nur diese für die Perzeptionsexperimente Verwendung fanden. Die Analyse erfolgte durch die Erstautorin, wobei Zweifelsfälle mit der Zweitautorin erörtert und konsensuell gelöst wurden. Die Systematik der Stimmqualitäten orientiert sich an Laver [7]. Da für die Perzeptionsexperimente nur die geringen und extremen Stimuli verwendet wurden, werden zur besseren Vergleichbarkeit hier auch nur die Stimmqualitäten dieser Stimuli betrachtet.

In zwei getrennten Perzeptionsexperimenten beurteilten zwei Gruppen naiver Hörer die wahrgenommen Emotionen. Um die Hörer nicht zu überfordern, wurden hierfür lediglich die Stimuli mit geringgradig und extrem ausgeprägten Emotionen berücksichtigt. Die Gesamtzahl der Hörer betrug 121, von denen 61 der Gruppe A und 60 der Gruppe B angehörten. Beide Gruppen erhielten einen Fragebogen im forced-choice-Design. Der Fragebogen der Gruppe A enthielt lediglich eine Liste mit den Emotionen, wie sie den Sprechern vorgegeben worden waren, während sich im Fragebogen der Gruppe B sechs weitere Antwortkategorien befanden (bei beiden Gruppen wurden Heiße Wut und Kalte Wut jeweils unter der Kategorie Wut zusammengefasst). Für den vorliegenden Überblick über die Ergebnisse der Perzeptionstests wurden die Daten beider Gruppen zusammengefasst.

3. Ergebnisse

3.1. Produktion

Die neutralen Stimuli wurden von keinem Sprecher durch eine von modaler Stimme abweichende Stimmqualität markiert. Modale Stimmqualität tritt aber auch bei allen Emotionen auf, vor allem bei Kalter Wut und Freude. Dabei findet sie sich wesentlich häufiger bei geringgradig emotionalen (52%) als bei extremen Stimuli (34,7%).

Tabelle 1 zeigt die Ergebnisse der Produktionsanalyse. Sie verdeutlicht die prozentuale Verteilung der unterschiedlichen Stimmqualitäten auf die intendierten Emotionen. Aufgrund von Mehrfachkodierungen können sich die Zahlen auf über 100% addieren.

Stimm-qualität	A	HW	KW	E	T	F	N
modal	48,3	48,3	60,0	41,7	40,0	86,7	100,0
breathy	33,3		20,0	16,7	68,3	41,7	
whispery	26,7	20,0	38,3	25,0	18,3	1,7	
whisper	5,0		13,3	1,7	3,3		
tense	33,3	51,7	8,3	41,7	15,0	6,7	
creaky			8,3	6,7	5,0	1,7	
harsh		31,7	20,0	23,3		1,7	
falsetto	1,7					13,3	
harsh whispery		3,3	15,0	3,3			
whispery creak			11,7	3,3			

Tabelle 1: *Auftreten der Stimmqualitäten per Emotion in Prozent*.

Trauer wird deutlich häufiger als andere Emotionen mit *breathy voice* gesprochen. *Breathy voice* findet sich auch bei Freude und Angst häufiger. *Whispery voice* tritt am häufigsten bei Kalter Wut auf. Diese Stimmqualität findet auch bei Heißer Wut und Ekel. Tatsächliches Flüstern (*whisper*)[1] wird von den Sprechern selten genutzt, am häufigsten noch bei Kalter Wut (13,3% der Stimuli).

Tense voice findet sich besonders bei Heißer Wut und Ekel, auch bei Angst, und selten bei Kalter Wut. Diese drei Emotionen porträtieren die Sprecher auch mit *harsh voice*. *Creaky voice* wird insgesamt sehr selten von den Sprechern genutzt, in lediglich 8,3% der Stimuli mit Kalter Wut. Einen besonderen Fall stellt hier auch die Falsettstimme (*falsetto voice*) dar, die fast ausschließlich bei der Porträtierung von Freude zum Einsatz kommt (die sich "vor Freude überschlagende Stimme") – und nur von den weiblichen Sprechern verwendet wird.

Harsh whispery voice und *whispery creak* stellen Kombinationen von je zwei Stimmqualitäten dar. Da sie von einigen Sprechern mehrfach (und durchgehend kombiniert) genutzt wurden, werden sie als eigenständige Kategorien gelistet. Beide treten gehäuft bei Kalter Wut auf und nur in wenigen Einzelfällen bei Heißer Wut oder Ekel.

3.2. Gendereffekte

Die weiblichen Sprecherinnen verwenden nicht-modale Stimmqualitäten etwas häufiger als ihre männlichen Kollegen. Vor allem aber variieren sie ihre Stimmqualität innerhalb eines Stimulus stärker, zeigen also nicht durchgehend eine, sondern verschiedene Stimmqualitäten im Verlauf eines Satzes. Mit welcher Häufigkeit eine bestimmte Stimmqualität bei den männlichen und weiblichen Sprechern auftritt, zeigt Tabelle 2.

[1]Es wird in dieser Untersuchung im Sinne von Laver (1980) zwischen *whispery voice* und *whisper* unterschieden.

Stimmqualität	männl.	weibl.
modal	57,9 %	57,4 %
breathy	23,6 %	32,2 %
whispery	18,5 %	21,5 %
whisper	4,1 %	3,1 %
tense	29,7 %	19,0 %
creaky	4,1 %	2,6 %
harsh	10,3 %	13,3 %
falsetto	0 %	4,6 %
harsh whispery	2,6 %	4,1 %
whispery creak	3,1 %	1,5 %

Tabelle 2: *Verwendung der verschiedenen der Stimmqualitäten bei den männlichen und weiblichen Sprechern in Prozent*

Breathy voice wird deutlich häufiger von den weiblichen als den männlichen Sprechern verwendet, *tense voice* hingegen wesentlich häufiger von den männlichen Sprechern. *Falsetto voice* wird ausschließlich von den weiblichen Sprechern genutzt. Überraschend ist, dass keiner der Sprecher durchgehend *creaky voice* zur Markierung einer Emotion verwendet, und dass *creaky voice* generell sehr selten genutzt wird. Aufgrund der Literatur (vgl. [1-3]) wäre anderes zu erwarten gewesen.

3.3. Perzeption

Bei einer allgemeinen Auswertung der Erkennungsraten ergibt sich folgendes "Ranking" für die verschiedenen Emotionen:

N (79,7%) > T (51,6%) > HW (50,4%) > E (42,2%) > F (40,2%) > A (36,0%) > KW (31,3%)

Stimm-qualität	Erkennungsrate gesamt
modal	41,0 %
breathy	54,2 %
whispery	46,7 %
whisper	44,9 %
tense	66,5 %
creaky	36,9 %
harsh	55,9 %
falsetto	53,7 %
harsh whispery	37,9 %
whispery creak	22,0 %

Tabelle 3: *Erkennungsraten für Stimuli mit modaler bzw. einer nicht-modalen Stimmqualität, gemittelt über alle Emotionen.*

Tabelle 3 verdeutlicht die über alle Emotionen gemittelten Erkennungsraten, reduziert den Blick also darauf, zu wieviel Prozent Stimuli korrekt erkannt werden, die mit einer bestimmten Stimmqualität produziert werden. Besonders gut erkannt werden Stimuli mit *tense voice*, ebenso auch *harsh voice*, *breathy voice* und *falsetto voice*. Schlecht erkannt werden hingegen Stimuli mit *whispery creak*. Letzteres mag dem Umstand geschuldet sein, dass *whispery creak* relativ selten verwendet wird und dies auch nur für zwei Emotionen.

Tabelle 4 bietet eine Übersicht über die korrekte Erkennung der einzelnen Emotionen, wenn sie mit einer bestimmten Stimmqualität produziert wurden. Dabei liegen den Prozentzahlen naturgemäß unterschiedlich große Anzahlen an Stimuli zugrunde.

	A	HW	KW	E	T	F	N
modal	20,9	35,3	31,1	24,0	54,2	41,9	79,7
breathy	31,0		18,2	50,0	61,7	39,7	
whispery	37,2	44,7	37,5	30,3	50,1	19,8	
whisper	77,7		32,2	57,0	66,1		
tense	55,1	74,9	56,8	49,7	78,0	20,1	
creaky			26,5	38,7	54,8	2,5	
harsh		60,4	51,4	62,9		34,7	
falsetto	88,4					49,3	
harsh whispery		53,7	40,3	23,6			
whispery creak			27,3	23,6			

Tabelle 4: *Erkennungsraten für die einzelnen Emotionen, die mit einer bestimmten Stimmqualität produziert wurden, in Prozent. Grau hinterlegt Werte, bei denen nur wenige Stimuli zugrundeliegen (n<8).*

Trauer und Heiße Wut werden jeweils in Kombination mit *tense voice* sehr gut erkannt. Ekel und Heiße Wut in Kombination mit *harsh voice* werden ebenfalls gut erkannt, außerdem auch Trauer in Kombination mit *breathy voice*.

4. Diskussion und Ausblick

Sprecher nutzen verschiedene Stimmqualitäten zur Porträtierung von Emotionen. Einige Emotionen werden von allen Sprechern mit einer bestimmten Stimmqualität gekennzeichnet, wie z.B. Trauer durch *breathy voice* oder Heiße Wut durch *tense voice*, bei anderen variiert dies sprecherabhängig (zwei Sprecher nutzen niemals *whispery voice* zur Kennzeichnung von Angst, andere sehr häufig) bzw. auch geschlechterspezifisch (*falsetto voice* wird nur von den weiblichen Sprechern genutzt). Bei der geringen Anzahl von Sprechern kann eine Gruppierung nach Männern und Frauen allerdings lediglich Hinweise auf eine geschlechterspezifisches Verhalten geben.

Stimuli, die extreme Emotion porträtieren, werden häufiger mit einer nicht-modalen Stimmqualität gesprochen als "geringgradige" Stimuli.

Entgegen den Erwartungen verwendet keiner der Sprecher *creaky voice* bei Trauer, wie dies z.B. von Braun/Heilmann [3] berichtet wird. Insgesamt tritt *creaky voice* sehr selten auf. Da diese Stimmqualität andererseits auch in nicht-emotionaler Sprache zur Signalisierung des Äußerungsendes genutzt wird (vgl. Henton/Bladon [8]), lässt sich hier nicht sauber differenzieren, ob satzfinale *creaky voice* auch zur Emotionsmarkierung genutzt wird oder nicht. Daher wurde satzfinales Auftreten dieser Stimmqualität nicht gewertet.

Harsh voice bei Wut – in beiden Ausprägungen, Heiße und Kalte Wut – bestätigt die Erwartungen hingegen. Bei Heißer Wut wird sie etwas häufiger genutzt als bei Kalter Wut (31,7% der Stimuli gegenüber 20,0%). Auch die Markierung von Freude mit *breathy voice* entspricht den Ergebnissen früherer Untersuchungen – mit 41,7% der Stimuli ist sie die am häufigsten für Freude gewählte Stimmqualität.

Whisper wird selten genutzt (deutlich häufiger *whispery voice*), dies mag der Tatsache geschuldet sein, dass tatsächliches Flüstern (*whisper*) aufgrund seiner akustischen Eigenschaften nicht weit trägt, für die Sprecher der vorliegenden Studie, die als professionelle Sprecher arbeiten, aber textliche Verständlichkeit eine sehr hohe Priorität hat. *Whisper* tritt vor allem bei Kalter Wut auf (25% der Stimuli) und wird häufig nur sehr kurz, etwa zur Hervorhebung eines einzelnen Wortes, genutzt.

Die vorliegenden Daten bestätigen die Distinktion zwischen Heißer Wut und Kalter Wut, wie sie Scherer [5] vorschlägt. Zwar nutzen die Sprecher für beide Emotionen *whispery voice* und die zu erwartende *harsh voice*, darüber hinaus aber aber auch weitere unterschiedliche Stimmqualitäten. Bei Heißer Wut kommt in 51,7% auch *tense voice* vor, während diese bei Kalter Wut nur in 8,3% der Stimuli auftritt. Kalte Wut hingegen wird auch mit *breathy voice* porträtiert (20%, besonders geringe Kalte Wut, wie es sowohl Abadjieva et al. [4] als auch Murray/Arnott [1] berichten), bei Heißer Wut tritt diese Stimmqualität nie in Erscheinung.

Betrachtet man die emotionalen Äußerungen genauer, so wird deutlich, dass hinsichtlich der Stimmqualität geringe und extreme Intensität der Emotion durchaus unterschiedlich markiert werden. Nur in einigen Fällen tritt bei extremer Intensität noch eine Stimmqualität hinzu, während geringe Intensität der Emotion mit modaler Stimme gesprochen wird. Wesentlich häufiger kennzeichnen die Sprecher aber beide Intensitätsstufen mit unterschiedlichen Stimmqualitäten, z.B. geringe Angst durch *breathy voice*, extreme Angst hingegen durch *whispery voice* oder *tense voice*. Auch bei der Analyse anderer akustischer Parameter zeigt sich, dass Sprecher bei der Produktion von geringen (zu mittleren und) zu extremen Emotionen durchaus nicht eine bloße Steigerung ihrer produktionseitigen "Einstellungen" vornehmen. Für weitere Untersuchungen empfiehlt es sich daher unbedingt, mindestens zwischen einer geringen und einer extremen Graduierung zu unterscheiden.

Die Erkenntnisse lassen Rückschlüsse darauf zu, welche Emotionen von Sprechern mit charakteristischen Stimmqualitäten porträtiert werden. Inwieweit diese Stimmqualitäten für die Erkennung der jeweiligen Emotion seitens der Hörer relevant sind, muss eine statistische Analyse zeigen. Die z.T. sehr guten Erkennungsraten von z.B. Trauer-Stimuli, die mit *breathy voice* produziert werden, und andererseits die auffallend schlechte Performanz der Hörer bei Kalter Wut mit *breathy voice* legen zumindest eine gewisse Einheitlichkeit hinsichtlich "Emotion – Stimmqualität" sowohl auf der Produktionsseite als auch bei der Erwartungshaltung der Hörer nahe. Dass Freude und auch Trauer, die beide häufig in modaler Stimmqualität kodiert werden, gleichermaßen gute Erkennungsraten erhalten, kann als Hinweis darauf gewertet werden, dass Hörer sich für ihr Urteil nicht ausschließlich auf die wahrgenommene Stimmqualität verlassen. In der vorliegenden Studie wurden auch weitere akustische Parameter wie Grundfrequenz und abgeleitete Parameter, Sprechtempo und Formanten analysiert. Eine Überprüfung der Korrelation dieser Werte mit den Erkennungsraten durch die Hörer steht noch aus.

5. Literatur

[1] M. Kienast, *Phonetische Veränderungen in emotionaler Sprechweise*, Shaker: Aachen, 2002.

[2] I. R. Murray und J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *Journal of the Acoustical Society of America*, vol. 93, no. 2, 1993, S. 1097-1108.

[3] A. Braun und C. Heilmann, *SynchronEmotion*, Frankfurt a. M.: Peter Lang, 2012. (Hallesche Schriften zur Sprechwissenschaft und Phonetik, Bd. 41).

[4] E. Abadjieva, I.R. Murray und J.L. Arnott, "Applying analysis of human emotional speech to enhance synthetic speech", *EUROSPEECH '93. Proceedings of the 3rd ISCA Conference on Speech Communications and Technology*, Berlin 1993. S. 909-912.

[5] K. R. Scherer, "Vocal Affect Expression: A Review and a Model for Future Research," *Psychological Bulletin*, vol. 99, no. 2, 1986, S. 143-165.

[6] R. Banse und K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, 1996, S. 614-636.

[7] J. Laver, "The phonetic description of voice qualiy", Cambridge University Press: Cambridge, 1980.

[8] C. Henton und A. Bladon, "Creak as a Sociophonetic Marker", in Language, Speech and Mind. Studies in Honour of Victoria A. Fromkin, L. M. Hyman und C. N. Li, Routledge: London/New York, 1988, S. 3-29.

An automatic chunk segmentation tool for long transcribed speech recordings

Nina Poerner[1], *Florian Schiel*[1]

[1]Bavarian Archive for Speech Signals,
Ludwig-Maximilians-Universität München
npoerner@phonetik.lmu.de, schiel@phonetik.lmu.de

Abstract

Forced alignment tools such as the Munich Automatic Segmentation System (MAUS) [1] do not scale well with input size. In this paper, we present a preprocessor chunk segmentation tool to combat this problem. It dramatically decreases MAUS's runtime on recordings of duration up to three hours, while also having a slightly positive effect on segmentation accuracy. We hope that this tool will advance the use of non-scientific transcribed recordings, such as audio books or broadcasts, in phonetic research. The chunker tool will be made available as a free web service at the Bavarian Archive for Speech Signals (BAS) [2].
Index Terms: segmentation, speech technology, web services

1. Introduction

The Munich Automatic Segmentation System is a phonetic segmentation tool for transcribed audio files. Apart from being able to perform forced alignment, it can also be used to model pronunciation variation, leading to a potentially more accurate segmentation. MAUS is currently available as a web service at the BAS, and is used by a growing number of academic users [3]. Unfortunately, MAUS's runtime does not scale well with input duration, making its use impractical for audio files beyond the 20 minute mark (see Figure 1). There are segmentation tools that are less susceptible to long input durations (e.g. [4], [5], [6]), however, none of them are to our knowledge able to model pronunciation variation in the way that MAUS is. Therefore, we aimed to find a way to make MAUS faster, while preserving this feature. One way to do so is by providing MAUS with a chunk segmentation, thereby breaking up the segmentation task into smaller subtasks. The chunker tool presented below is able to create such a chunk segmentation automatically in a relatively short period of time.

2. Previous works

The chunker builds heavily on a method introduced by [4]. Their main insight is that, after performing speech recognition on the signal, the alignment can be performed in the symbolic instead of the signal domain, which is generally less costly. While the resulting symbolic alignment is unlikely to be perfect, there may be stretches where the recognized string and the transcription match for a sufficient number of symbols, meaning that they can be considered aligned (so-called 'anchors'). Any non-aligned stretches can be recursively subjected to the same procedure, taking advantage of the fact that information about their content has become more specific since the last iteration. [5] take a similar approach, but also adapt acoustic models on already aligned stretches. [6] perform phoneme recognition instead of word-based recognition, which speeds up the process at the cost of accuracy.

3. The tool

The chunker presented here takes a similar approach to the tools mentioned above. However, instead of producing a phonetic segmentation, it segments the material into chunks, leaving the more fine-grained segmentation to MAUS. The chunker is interoperable with other tools developed at the BAS, as it complies with their data interchange strategies. It is implemented as a C++ module and requires a number of HTK [7] tools.

3.1. Three algorithms in one

The chunker is able to perform recognition and symbolic alignment at both the word and the phoneme level. Therefore, it can be used in three ways:

- **T chunker**: recognition on word-based lattice, followed by token-based symbolic alignment
- **P chunker**: recognition on phoneme-based lattice, followed by phone-based symbolic alignment
- **TP chunker**: recognition on word-based lattice, followed by phone-based symbolic alignment

3.2. Recognition

Recognition is performed by the HTK tool HVite. The underlying language or phonotactic model is trained on the transcription, with a customizable n-gram degree and smoothing factor. If multithreading is enabled, the audio file is split and parallelized, thereby further speeding up the process.

3.3. Symbolic alignment

Symbolic alignment is performed by an implementation of the Hirschberg algorithm [8], with naive Levenshtein [9] edit costs. The Hirschberg algorithm is well suited for long symbolic

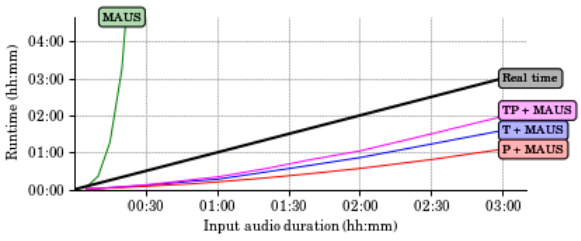


Figure 1: *Runtime T chunker + MAUS, TP chunker + MAUS, P chunker + MAUS and MAUS-only phonetic segmentation at different input audio durations on an HP ProLiant DL160 G6 server, using 6 threads*

alignment tasks as its memory requirements are linear, and its quadratic runtime requirements can be parallelized to a certain degree.

3.4. Chunk boundary selection

After symbolic alignment, the alignment path is traversed to find suitable anchor sequences for chunk boundaries. Anchor sequences must have a certain length, edit costs below a certain threshold, and contain a certain number of singleton tokens. These parameters, which govern how conservative the chunker is in its choices, can be set by the user. The discovered anchors are then screened for suitable chunk boundaries, making sure that any chunks created have at least a minimum duration, which is also set by the user. Whenever possible, the chunker tries to place chunk boundaries in inter-word pauses.

3.5. Recursion

For every discovered chunk that has at least twice the minimum chunk duration, a 'child chunker' is spawned, which is limited to the given area of the audio and transcription. This means that its language model is more strongly attuned to its transcription, making it better equipped to find anchors that its parent could not discover.

4. Evaluation

The chunker was evaluated on the German sessions of the Verbmobil corpora [10], using the acoustic models of the MAUS tool [1]. A set of long recordings was created by concatenating turns and their transcriptions from the Verbmobil material. Details and the exact configurations used for the tests can be found in [11].

4.1. Performance

See Figure 1 for a comparison of the time it takes to segment concatenated recordings of various durations with MAUS alone, and with the chunker and MAUS combined. It is obvious that all three chunker algorithms decrease segmentation time, with the phoneme-based chunker taking the lead.

4.2. Chunk boundary errors

Chunk boundary errors were evaluated on 27 one-hour concatenated recordings. All turns in these recordings had been MAUS-aligned prior to concatenation. The word boundaries from this alignment were taken as a ground truth. It appears that most boundaries placed by any of the three algorithms are within 100 ms of their ground truth locations (see Figure 2).

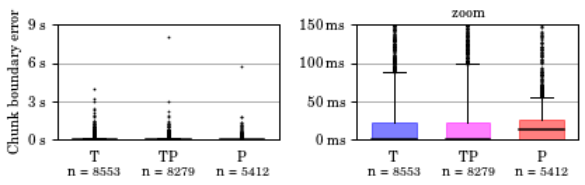


Figure 2: *Chunk boundary errors relative to ground truth segmentation (whiskers are 5% and 95% percentiles, zero error = correctly located inter-word pause)*

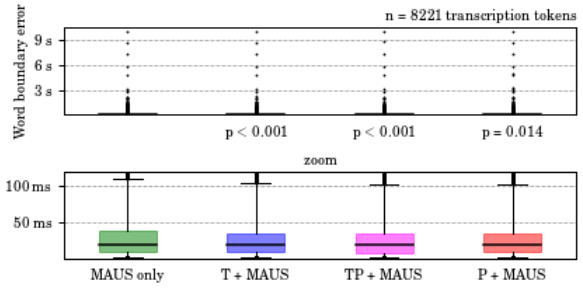


Figure 3: *MAUS word boundary errors with and without chunker preprocessing (whiskers are 5% and 95% percentiles, p values refer to paired t-tests Chunker + MAUS vs. MAUS only)*

4.3. Effects on MAUS's segmentation accuracy

A 55 minute subset of the Verbmobil material was used to test the effects of the chunker on MAUS's segmentation accuracy. This subset had the additional advantage of containing a manual ground truth segmentation. It appears that using the chunker as a preprocessor to MAUS does not decrease MAUS's segmentation accuracy (as measured by the distance between predicted word beginnings and their manually segmented beginnings). Quite the contrary, segmentation accuracy increased, although the effects were rather small (see Figure 3)[1].

4.4. Effects on MAUS under adverse conditions

The effects of the chunker on MAUS were also tested under adverse conditions such as missing transcription tokens, missing transcription turns and cross-talk in the audio. In most scenarios, MAUS's word segmentation appears to improve slightly when the chunker is used as a preprocessor, with the biggest positive effect in the missing turns scenario (see Figure 4).

4.5. Audio book segmentation

To test a more realistic application scenario, the T chunker and MAUS were used to phonetically segment a 14 hour public domain audio book of a German translation of 'Madame Bovary' [12], with chapter durations up to 56 minutes. The full process, from plain text to phonetic segmentation, took less than 7 hours on three threads. While there was no ground truth for formal tests of accuracy, manual checks showed promising results.

5. Conclusions

In this paper, we have presented an automatic chunk segmentation tool that can speed up MAUS segmentation on long transcribed recordings, while not appearing to decrease segmentation quality. We therefore hope that the chunker will prove a useful addition to the BAS web services. Of course, its chunk segmentation does not provide the semantic information of a manual turn or sentence segmentation, meaning that it cannot be used for analyses of sentence durations, speaker comparisons, and the like. Nonetheless, we hope that it will be a step towards a broader inclusion of non-scientific resources, such as audio books or broadcasts, into phonetic research.

[1] Due to the expected runtime of MAUS on the 55 minute recording, the input to the MAUS-only tests was manually presegmented into chunks of duration 10 minutes, effectively giving MAUS a head start over the chunker.

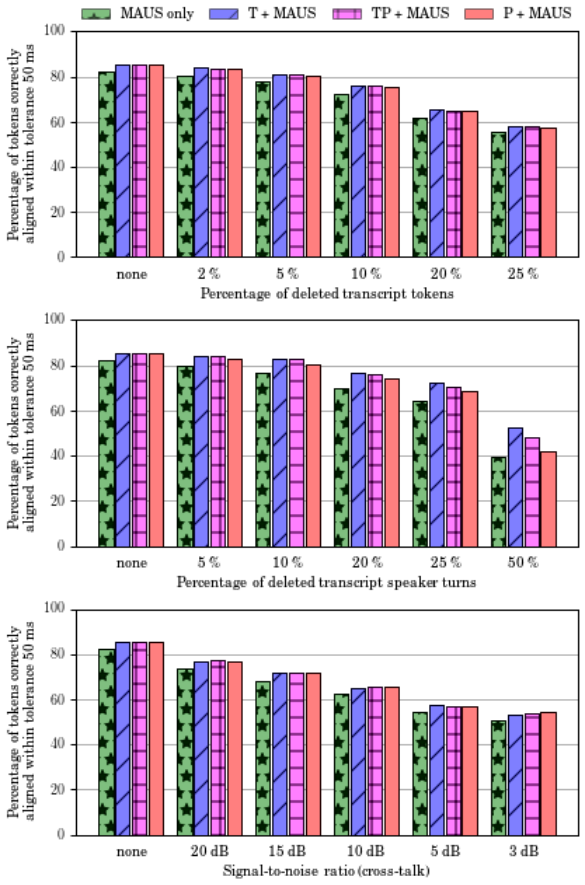


Figure 4: *Percentage of words correctly aligned within tolerance 50 ms with and without chunker preprocessing, under adverse conditions*

6. Bibliography

[1] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proc. of the ICPhS*, 1999, pp. 607–610.

[2] Bavarian Archive for Speech Signals. [Online]. Available: http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B

[3] T. Kisler, U. D. Reichel, F. Schiel, C. Draxler, B. Jackl, and N. Poerner, "BAS speech science web services – an update on current developments," in *LREC*, 2016.

[4] P. J. Moreno, C. F. Joerg, J.-M. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *ICSLP*, vol. 98, 1998, pp. 2711–2714.

[5] A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[6] G. Bordel, M. Peagarikano, L. J. Rodrguez-Fuentes, and A. Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in *INTERSPEECH*, 2012, pp. 1840–1843.

[7] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, and D. Povey, *The HTK book*. Cambridge University, 2006.

[8] D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences," *Communications of the ACM*, vol. 18, no. 6, pp. 341–343, 1975.

[9] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, 1966, pp. 707–710.

[10] S. Burger, K. Weilhammer, F. Schiel, and H. G. Tillmann, "Verbmobil data collection and annotation," in *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000, pp. 537–549. [Online]. Available: http://hdl.handle.net/11022/1009-0000-0000-EB30-1 and http://hdl.handle.net/11022/1009-0000-0000-FC54-6

[11] N. Poerner, "Development of an automatic chunk segmentation tool for long transcribed speech recordings," MA thesis submitted for approval.

[12] G. Flaubert, *Frau Bovary (translation by Arthur Schurig)*, 1857. [Online]. Available: http://www.gutenberg.org/ebooks/15711 and https://librivox.org/frau-bovary-by-gustave-flaubert

Open data for speech synthesis of Austrian German language varieties

Michael Pucher[1], *Michaela Rausch-Supola*[1], *Sylvia Moosmüller*[1],
Markus Toman[2], *Dietmar Schabus*[3], *Friedrich Neubarth*[3]

[1]Acoustics Research Institute (ARI), Austrian Academy of Sciences (OAW)
{michael.pucher,michaela.rausch-supola,sylvia.moosmueller}@oeaw.ac.at
[2]Vienna University of Technology (TUW), Austria
m.toman@neuratec.com
[3]Austrian Research Institute for Artificial Intelligence (OFAI)
{friedrich.neubarth,dietmar.schabus}@ofai.at

Abstract

In this paper we summarize open data sets and open source software that we have released for Austrian German language varieties as a result of several research projects. We also describe some data sets that are released for research purposes only, due to licensing limitations. From the development of these resources we draw conclusions concerning the collection and licensing of such data with a special focus on the problem of speech synthesis where the voice identity of the speaker plays an important role. Furthermore we discuss recordings that we plan to perform in the future, where we aim to cover most Austrian dialects.

Index Terms: speech synthesis, language varieties, dialect, sociolect, open data, open source

1. Introduction

In this paper we describe a set of tools and data sets that we have already released for Austrian German language varieties (dialects, sociolects) concerning audio as well as audio-visual speech synthesis. This includes the following:

- The **open source** SALB front-end for speech synthesis using Hidden Markov Model (HMM)-based Speech Synthesis System (HTS) voice models (available from http://m-toman.github.io/SALB/) [1].
- "Leo", an **open data** HTS based voice model for Austrian German (available from https://sourceforge.net/projects/at-festival/) [2].
- **Open data** for building unit selection voices for Viennese dialects with the Festival speech synthesis system (available from http://speech.kfs.oeaw.ac.at/vdvoices/) [3].
- **Open research data** for audio-visual dialect synthesis - Goisern and Innervillgraten Audiovisual Dialect Speech Corpus – GIDS (available from http://speech.kfs.oeaw.ac.at/gids/) [4].
- **Open research data** for triple modality speech synthesis – Multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech MMASCS (available from http://speech.kfs.oeaw.ac.at/mmascs) [5].

Since all synthetic voices in the current speech synthesis paradigms (unit selection, HMM, Deep Neural Network (DNN)) are built using data of a specific speaker, they will reproduce the speaker's identity to a certain extent. This brings about some specific licensing problems that we will discuss in Section 8.

In the future we also aim to record dialect data of 40 different dialect regions in Austria in the field using a mobile recording studio. These recordings shall also be released under an open data or open research data license.

2. SALB front-end for speech synthesis using HTS voice models

Hidden-Markov Model (HMM) based speech synthesis provides a methodology for flexible speech synthesis while keeping a low memory footprint [5]. It also enables speaker adaptation from average voice models, allowing the creation of new voice models from sparse voice data [6], as well as techniques like interpolation [7][8] and transformation [9][10] of voices. A well-known toolkit for creating HMM-based voice models is HTS [11, 12]. Separate software toolkits are available to actually synthesize speech waveforms from HTS models. A popular, freely available framework is hts_engine [13]. Speech synthesis front ends on the other hand provide means for analyzing and processing text, producing the necessary input for the synthesizer. In HTS this input is a set of labels where usually each label represents a single phone and contextual information, including surrounding phones, position in utterance, prosodic information etc. While not exclusively being front ends and not specifically targeted for HTS, popular choices are Festival [14] or Flite [15]. Festival is a complex software framework for building speech synthesis systems focusing Unix-based operating systems.

Our main goal when creating the SALB front-end framework was to easily allow HTS voices to be used with the Speech Application Programming Interface 5 (SAPI5). This allows the framework to be installed on different versions of the Microsoft Windows operation system as speech synthesis engine, making HTS voice models available as system voices to applications like screen readers, e-book creators etc. The second goal was

[1]Published in [1]
[2]Published in [1]
[3]Published in [2]
[4]Published in [3]
[5]Published in [4]

simple integration of new languages and phone sets. The third goal was portability to mobile devices.

Flite has been adapted for HTS in the Flite+hts_engine software [13] and due to its small and portable nature it seemed like a good fit to our requirements. The structure of Flite makes integrating new languages rather cumbersome. [6] Therefore our framework integrates Flite for text analysis of English while additionally providing a second text analysis module that can utilize Festival style pronunciation dictionaries and letter to sound trees. Text preprocessing tasks (e.g. number and date conversion) can be added to the module in C++. Adding a completely new text processing module is also possible. The framework includes hts_engine for speech waveform synthesis and can be extended by other synthesizers.

3. "Leo", a HTS based voice model for Austrian German

Category	Phones (IPA)
Vowels (monoph.)	ɑ ɑː ɒ ɒː ɐ e ɛ eː i ɪ iː ɔ o oː øː æː ɐ œ œː ə u ʊ uː ʏ y yː
Vowels (monoph.) nasalized	ɒ̃ː ɔ̃ː ɑ̃ːɐ̯ œ̃ː
Diphthongs	aɪ̯ ɒːɐ̯ ɑːɐ̯ ɒɪ̯ ɑʊ̯ ɛɐ̯ ɛːɐ̯ ɪɐ̯ iːɐ̯ ɔɐ̯ ɔːɐ̯ oːɐ̯ ɔʏ̯ øːɐ̯ œɐ̯ ʊɐ̯ uːɐ̯ ʊːɐ̯ ʏɐ̯ yːɐ̯
Plosives (stops)	b̥ d̥ g̥ k ʔ p t
Nasal stops	m n ŋ
Fricatives	ç x f h s ʃ v z ʒ
Approximants	j
Trill	r
Lateral approx.	l

Table 1: Phone set used for Austrian German voice "Leo".

With the framework we provide a free voice model of a male, professional speaker for Standard Austrian German called "Leo". The model is built from 3,700 utterances recorded in studio quality using a phonetically balanced corpus. The phone set used in the voice can be seen in Table 1. A pronunciation dictionary with 14,000 entries, letter to sound rules and procedures for number conversion are also included. The voice is distributed with the framework, but can also be used with the Festival speech synthesis system.

4. Unit selection voices for Viennese dialects

Within the research project "Viennese sociolect and dialect synthesis" (VSDS), we developed three voices for speech synthesis modeling three Viennese varieties. In the light of personalization and regionalization of speech based interfaces it becomes indispensable to develop not only high quality speech synthesis for different languages but also for a representative set of language varieties, i.e., dialects that differ from the standard variety substantially enough to treat them alongside different languages. In performing this task, the focus lies on the necessity that the developed synthetic voices must be able to shift between the standard variety and specific dialects, similar to everyday language use [7]. In Vienna, language varieties are differentiated rather socially than regionally, therefore it would be correct to speak about sociolects. In the VSDS project, we developed three different voices: a voice representing "the Viennese dialect", one representing colloquial Viennese, and one representing the youth language in Vienna. For the recordings, we could win two renowned actors and for the recordings of youth language, we arranged a casting among pupils of vocational schools.

4.1. Speaker selection

The selection of the professional speakers was based on several criteria, amongst others: reading speed and accuracy, the accuracy of their standard Austrian pronunciation, the degree of authenticity of their sociolect, the consistency of their pronunciation (in particular, we did not want them to shift between different sociolects without being told so), and the pleasantness of the voice.

4.2. Text selection

The quality of a unit selection voice highly depends on how well the recorded material covers the set of possible diphones and prosodic contexts. Most of our recording text script for the standard Austrian variety was selected from large corpora of non-proprietary texts, such as EU parliament debate transcripts, and from the Viennese city magazine "Falter" (with their friendly approval). We were aiming at diphone coverage with the following linguistic context features: lexical stress, syllable boundaries and word boundaries. During the initial iterations of text selection, we focused on the most frequent diphones without features while taking account of some back off strategies, for example that diphones bridging a word boundary can easily be backed off by inserting a short pause.

4.3. Recording

The recordings were made in an unechoic, acoustically isolated room with a HD-recorder (44100 kHz sampling rate, 16 bit encoding) and a professional microphone. We made sure that the recording parameters (distance to microphone recording level) were the same for each session. The recordings were semi-automatically segmented at sentence level using the acoustic software S_TOOLS-STx of Acoustics Research Institute (ARI) and a script written in Perl. The speech database contains transcriptions and soundfiles corresponding to single sentences. Importantly, these are not just cut from the original recordings, but they can be dynamically exported each time some alignments change.

4.4. Voices

The release "Unit selection voices for Viennese dialects" contains data for 3 Viennese voices (Table 2). Additionally the release contains base lexica for the phonetic encoding of each variety, which covers the most important and typical words of the respective Viennese variety, and a set of letter-to-sound rules for Austrian German. The voices can be used with the Festival speech synthesis system [14], in particular the open-domain unit selection Multisyn [16]. The provided data can also be used for training of HMM-based voices for HTS [12].

[6] We have published instructions on adding a new language to Flite: http://sourceforge.net/p/at-flite/wiki/AddingNewLanguage/

Voice ID	Variety	Age group	Database size
HPO	Viennese dialect	45-60	2:55
HGA	Colloquial Viennese	60-70	3:10
JOE	Viennese youth language	15-25	2:11

Table 2: Viennese dialect unit selection voices.

5. GIDS – Goisern and Innervillgraten Audiovisual Dialect Speech Corpus

Visual speech synthesis techniques have possible applications in computer games and films. Generating visual speech directly from audio data is nowadays a state-of-the-art technique in facial animation in the computer games industry [17]. We have developed a corpus of audio-visual speech recordings to investigate visual dialect text-to-speech synthesis where we generate an acoustic and visual signal of a certain speaker from given text.

5.1. Corpus

The Goisern and Innervillgraten Dialect Speech (GIDS) Corpus is a collection of audiovisual speech recordings for research purposes. It consists of a total of 7068 sentences spoken by eights speakers from two Austrian villages, Bad Goisern (BG) and Innervillgraten (IVG). For each speaker, about two thirds of the recorded sentences are in the speaker's respective dialect and the rest is in Regional Standard Austrian German (RSAG). The dialect of Bad Goisern in the Salzkammergut region belongs to the (South)-Central Bavarian dialects, and the dialect of Innervillgraten in the East Tyrol region belongs to the Southern Bavarian dialect family as shown in Figure 1.

After a careful phonetic analysis we compiled sets of phonetically balanced sentences (656 for IVG and 665 for GOI) with respect to the phone set established for the dialect, the frequency of occurrence of each phone in the data, and the context specific variation of phones. The utterances of the recording script were extracted from a larger corpus of material consisting of 18-20 hours of recordings for each dialect with at least 10 speakers per dialect. These sentences consisted of spontaneous speech (elicited with key words) and translation tasks. We created a lexicon of words occurring in the script. The script was divided into a training and testing part. In the final audio-visual recordings we recorded 2 male and 2 female speakers per dialect, i.e., 8 speakers in total.

The recordings consist of optical 3D facial motion tracking data, captured with a NaturalPoint OptiTrack Expression system,[7] the greyscale video data also recorded by the same system, and studio quality audio.

For each of the recorded utterances, the corpus contains a RIFF wave audio file, facial marker data in the form of a matrix stored as a text file, a gray scale video from the optical system, the sentence of the utterance in plain text, a text file listing the phones spoken in the utterance including begin and end times of all phones, and a quin-phone full-context label file.

6. MMASCS – Multi-Modal Annotated Synchronous Corpus of Speech

The MMASCS corpus is a multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data

[7] http://www.naturalpoint.com/optitrack/

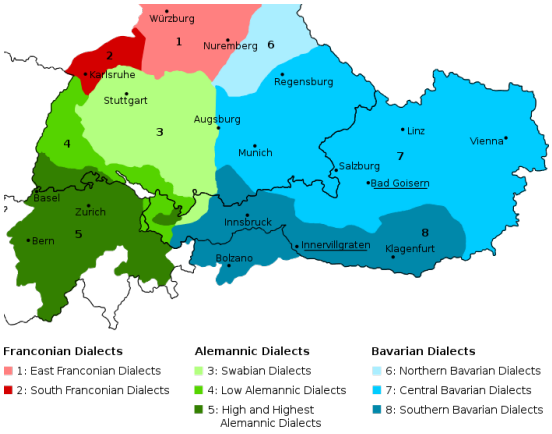


Figure 1: Upper German dialects

(a) (b)

Figure 2: Example still images from the gray scale video from the OptiTrack system (a) and the color video from the camcorder (b) showing visual markers and articulatory markers.

of normal, fast and slow speech. The tongue motion data is captured with Electro-Magnetic Articulography (EMA).

The MMASCS corpus combines facial motion capture data with intra-oral EMA data. In comparison to optical motion capturing only, this has the obvious advantage of also providing tongue motion data, which is impossible to capture optically. In comparison to EMA data only, it has the advantage of providing a larger number of tracked points on the lips, eyelids, eyebrows and other areas of the face. While it is in principle possible to use EMA coils also on the face surface, the inexpensive and easy-to-attach optical markers are much less intrusive for the speaker than the EMA coils with their cable connection (one cable per coil) to the articulograph. Another difference is that our data is for Austrian German speech. One can imagine that it might be interesting to investigate inter-lingual differences in speech motion, once a larger number of corpora (of EMA and/or facial motion data) in various languages is available (of course speaker-specific effects would need to be accounted for). Finally, our data is different in that it comprises data of speech at three different speaking rates (normal, fast and slow).

We have already used this corpus for evaluating a method to convert from non-acoustic to acoustic speech, where we could show that visio-articulatory features can improve the conversion compared to visual only features [19].

7. Recording of high quality dialect data in the field

In the future we also aim to record dialect data of 40 different dialect regions in Austria in the field using a mobile recording

studio. These recordings shall also be released under an open data or open research data license.

The selection of dialect locations and speakers as well as the phonetic and phonological analyses of 40 Austrian dialects, an essential prerequisite for dialect synthesis, is currently in progress within the SFB project "Deutsch in Österreich" ("German in Austria").

For speech synthesis we will create phonetically balanced recording scripts, record 1 male and 1 female speaker for each location, perform a semi-automatic transcription of the recordings, build and investigate acoustic models for statistical parametric synthesis, and build a synthesis front-end. To achieve high quality recordings, we will deploy a mobile recording studio. We will test the studio by recording two speakers in the course of the year 2016. The recording scripts will be adapted from our existing recording scripts for Standard Austrian German (SAG), Viennese (VD), Innervillgraten (IVG) and Bad Goisern (BG) dialect.

8. Licensing, repositories, and standards

Since we are also synthesizing a speaker's identity the data we are collecting is very personal and the speakers must be informed about possible applications of their data. Many of the speakers that we have recorded agreed to release their data within an open data or open source framework, but we can also observe that the use of speech synthesis technology is not yet as widespread that speakers are able to fully understand possible application scenarios. Independent of country dependent legal requirements as scientists and developers we have to make sure to give speakers that we are recording a realistic perspective on what can happen with their voice data. The Festvox documentation [20] contains some guidelines on these issues with a list of possible licenses from "free for any use" to "fully proprietary", but in the future we may need more sophisticated licenses that reflect the fast technological changes that we are witnessing.

Our data sets are available from our websites, but it would be beneficial to have a common repository for data distribution within the speech communication community.

The data format standards that we use for the creation of our data sets are mainly set by the popular speech synthesis frameworks such as HTS and Festival. Such a kind of implicit and bottom-up standardization seems natural for a field that is strongly driven by research, but might not be optimal from an industry point of view.

9. Conclusion

We have given an overview of open data sets and open source software that we have released for Austrian German language varieties and drew some conclusions concerning the collection and licensing of such data with a special focus on speech synthesis. Furthermore we discussed recordings that we plan to perform in the future, where we aim to cover most Austrian dialects.

10. Acknowledgements

The project "Viennese Sociolect and Dialect Synthesis" was funded by the Vienna Science and Technology Fund (WWTF). This research was also funded by the Austrian Science Fund (FWF): P22890-N23, P23821-N23, and by the BMWF - Sparkling Science project *Sprachsynthese von Auditiven Lehrbüchern für Blinde SchülerInnen* (SALB).

11. References

[1] M. Toman and M. Pucher, "An open source speech synthesis frontend for HTS," in *Text, Speech, and Dialogue (TSD) 2015*, Pilsen, Czech Republic, 2015, pp. 291–298.

[2] M. Pucher, F. Neubarth, V. Strom, S. Moosmller, G. Hofer, C. Kranzler, G. Schuchmann, and D. Schabus, "Resources for speech synthesis of viennese varieties," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010, pp. 105–108.

[3] D. Schabus and M. Pucher, "Comparison of dialect models and phone mappings in hsmm-based visual dialect speech synthesis," in *Proceedings of the 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP)*, Vienna, Austria, Sept 2015, pp. 84–87.

[4] D. Schabus, M. Pucher, and P. Hoole, "The MMASCS multimodal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May 2014, pp. 3411–3416.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[6] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[7] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, feb 2010.

[8] C. Valentini-Botinhao, M. Toman, M. Pucher, D. Schabus, and J. Yamagishi, "Intelligibility Analysis of Fast Synthesized Speech," in *Proc. Interspeech*, Singapore, September 2014, pp. 2922–2926.

[9] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009, pp. 528–531.

[10] M. Toman, M. Pucher, and D. Schabus, "Cross-variety speaker transformation in HSMM-based speech synthesis," in *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*, Barcelona, Spain, Aug. 2013, pp. 77–81.

[11] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW)*, Bonn, Germany, Aug. 2007, pp. 294–299.

[12] "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/.

[13] "hts-engine," http://hts-engine.sourceforge.net/.

[14] "Festival," http://www.cstr.ed.ac.uk/projects/festival/.

[15] "Flite," http://www.festvox.org/flite/.

[16] R. A. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[17] SpeechGraphics, "Speech Graphics - Audio-driven facial animation," http://www.speech-graphics.com/, 2015.

[18] M. Toman, M. Pucher, S. Moosmüller, and D. Schabus, "Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis," *Speech Communication*, vol. 72, pp. 176 – 193, 2015.

[19] M. Pucher and D. Schabus, "Visio-articulatory to acoustic conversion of speech," in *Proceedings of the 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP)*, Vienna, Austria, Sept 2015, p. Article No.6.

[20] "Festvox - Who owns a voice," http://festvox.org/bsv/x794.html.

Romance Corpus Phonology: from *(Inter-)Phonologie du Français Contemporain* (I)PFC to *(Inter-)Fonología del Español Contemporáneo* (I)FEC

Elissa Pustka[1], *Christoph Gabriel*[2], *Trudel Meisenburg*[3]

[1]Institut für Romanistik, Universität Wien, Spitalgasse 2, 1090 Vienna, Austria / [2]Romanisches Seminar, Universität Mainz, Jakob-Welder-Weg 18, 55128 Mainz, Germany / [3]Institut für Romanistik/Latinistik, Universität Osnabrück, Neuer Graben 40, 49069 Osnabrück, Germany

elissa.pustka@univie.ac.at, christoph.gabriel@uni-mainz.de, tmeisenb@uos.de

Abstract

The corpus project *(Inter-)Fonología del Español Contemporáneo* (I)FEC aims to document the pronunciation of Spanish in the world, including L1 and L2 speakers as well as learners of Spanish as a foreign language. Our starting point is the French research program *(Inter-)Phonologie du Français Contemporain* (I)PFC. On the basis of its nearly 20 years of experience and several pilot studies on Spanish, we present for the first time the guidelines we developed for the data collection in this project. (I)FEC works with a modular system: the basic design elicits data via a reading task (word list and text), a discourse completion task and a semi-focused interview. For some types of speakers (learners, illiterates, multilinguals), we provide supplementary tasks. In doing so, we take into account variation in both segmental and suprasegmental phenomena such as regionally confined oppositions (e.g., /s/:/θ/), the weakening of coda consonants (particularly coda /s/), word stress, syllabification and intonation.

Key words: Corpus Linguistics, Phonology, Spanish, French, Foreign Language Learning, Second Language Acquisition (SLA), Third Language Acquisition

1. Introduction

For Spanish, there is not yet a research program comparable to *(Inter-)Phonologie du Français Contemporain* (I)PFC, which documents the variation of French pronunciation worldwide, provides an open access database for the international research community as well as for school and university education and disseminates the results to a large public ([1], [2], [3], [4]). The existing corpora of spoken Spanish only provide (hardly comparable) spontaneous data (e.g., [5], [6], [7]) or data from just one speaker per location (e.g., [8]), or they focus exclusively on prosody ([9], [10], [11]). None of them elicits phonological features systematically.

For this reason, we have revised the PFC method (section 2) and adapted it to Spanish. So far, we have tested different versions of the new guidelines in pilot studies with 9 L1 speakers (originally from Madrid, Tenerife, Mexico, Nicaragua, Cuba, Peru, Paraguay, Chile and Argentina, but living in a German-speaking environment today), with 15 L1 speakers in and around Seville (Spain), and with 12 learners of Spanish as a foreign language from Osnabrück (Northern Germany) and Vienna (Austria), respectively. The resulting method box of (I)FEC is presented in section 3.

2. Experiences from (I)PFC

PFC aims to provide the largest corpus of spoken French and sees itself as a prototype for other projects: Before (I)FEC, the method had already been transferred to *Phonologie de l'Anglais Contemporain* (PAC, for the phonology of English, [12]) and to *Interphonologie du Français Contemporain* (IPFC, [4]), which aims at learners of French as a foreign language. Since 1999, data from 418 speakers (corresponding to 36 survey points) have been integrated into the constantly growing online database ([1]).

2.1. Guidelines for data collection and analysis

PFC provides easily accessible guidelines for data collection and analysis, which also encourages students and non-phonologists to join the program and to contribute to an efficient expansion of the world-wide corpus.

2.1.1. Fieldworker(s) and informants

The PFC guidelines recommend working in teams of two fieldworkers, one of them knowing the speech community well, the other one less or not at all. In doing so, PFC attempts to overcome the observer's paradox in spontaneous discussions and to create a contrast with the semi-focused interview (see 2.1.2). At each location, approximately 12 speakers are recorded, who ideally have lived there throughout all their life. There should be an equal number of male/female speakers belonging to at least two generations and, if possible, to different social environments. Socio-demographic information is collected via a questionnaire. The informants sign an agreement allowing the anonymized data to be used for research and didactic purposes.

2.1.2. Speech recordings

PFC collects recordings in four tasks, which cover the continuum between formal and informal situations and styles as well as the medial gap between read and (more or less) spontaneous speech ([13]): reading of a word list (94 items), in which five minimal pairs are repeated at the end; reading of a fictitious newspaper article containing several items from the word list; a guided interview and a free conversation. This methodological mix aims at satisfying the requests for comparability and authenticity. Furthermore, it presents an economical advantage: If an opposition that is prescribed by the pronuncia-

tion norm is not realized in the word list, it is highly probable that this opposition is not realized in any other task either.

The minimal pairs at the end of the PFC word list test regionally confined oppositions, among others /a/:/ɑ/ in *patte* 'paw' vs. *pâte* 'pasta'. The words of the minimal pairs already appear individually in the randomly arranged word list, which tests numerous other phenomena, too. The informants often comment on this task, which gives valuable evidence about their representations and attitudes towards language variation and the norm as well as their phonological awareness. Some words of the list reappear in the text *Le Premier Ministre ira-t-il a Beaulieu ?* (393 words). For certain regions (e.g., Switzerland, Canada) additional word lists and text passages exist. In the IPFC project, the learners not only read the list, but also repeat it after a model speaker, which allows testing their phonetic and phonological competencies independently from the graphic forms and their reading capacities.

In addition, PFC requests two types of spontaneous speech, an interview conducted by the fieldworker the subject does not know yet (*entretien guidé*; 20 min.) and a conversation with the other one or among informants (*discussion libre*; 30 min.). In total, the PFC recordings take approximately 60 min., the IPFC recordings about 90 min. per informant.

2.1.3. Transcription and Analysis

The recordings are aligned to the sound and transcribed orthographically in Praat ([14]) following the (I)PFC conventions. In addition to the transcription of the text, 10 min. from each, the interview and the conversation, are transcribed. The text and 5 minutes of each type of spontaneous speech are further annotated with the PFC coding system for liaison and schwa (e.g., *grand11t honneur*, *pe0212tit*). Using the PFC tool Dolmen ([15]), the coding allows the quantitative exploration of these phenomena in the online corpus and the correlation of the realization rates with external (region, age, gender, education) and internal factors (e.g., position in the word).

2.2. Problematic issues

2.2.1. Fieldworker(s) and informants

Experience has shown that the PFC ideal of teams of two is in practice often hardly feasible. We thus plan to work with single fieldworkers in (I)FEC.

Another problem is that the PFC guidelines, despite their openness, imperatively prescribe the reading of the word list and the text. Illiterates are thus a priori excluded ([2], p. 29), which is problematic with respect to the representativeness of the data, particularly concerning older rural people, migrants and other multilinguals such as parts of the speakers in the overseas departments or in francophone Africa. In the case of Spanish, the problem already arose in our pilot study in and around Seville, where elderly speakers in a small village had serious problems with reading. For this reason, we plan to use a picture list and a picture story as a supplement or in extreme cases as a substitute for the word list and the text (see 3.2).

Another shortcoming of the PFC guidelines is that the linguistic skills of multilingual informants are not entirely documented. Especially in the IPFC learner corpus, the absence of L1 data makes the detection of interferences difficult, but the problem also concerns the L2 French in numerous francophone regions. Spanish is also spoken alongside other languages, such as indigenous languages in the Americas (e.g., Quechua, Nahuatl), co-official languages in Spain (e.g., Catalan, Basque) or migrant languages (e.g., Italian or Chinese in Argentina). Multilingual informants will thus be recorded in all their languages using Aesop's fable *The North Wind and the Sun* ([16]), either as a reading task or – in the case of illiterates – by retelling an orally presented version of the text. Further supplement tasks are optional.

Finally, we adapted the agreement form to current ethical standards ([17]) by including the address of a responsible person and the name of the fieldworker. Each speaker is attributed a non-traceable code, which ensures *anonymization* rather than mere *pseudonymization* (as is the case for the current PFC practice using the initials of the subject's first and last name).

2.2.2. Spontaneous speech

According to the reports of the fieldworkers and the results of many quantitative analyses, the original idea of the PFC program, which intended to contrast spontaneous speech in two situations, proved to be difficult to achieve. Further criticism shows that the free conversation does not meet the sociolinguistic requirements of a natural environment. Autobiographical topics as suggested by the PFC guidelines present difficulties for open access publishing of the corpus and didactic applications. Finally, the large data samples have only partially been transcribed and coded until now. This experience leads us to reducing the amount of spontaneous speech in (I)FEC to a 20 min. long interview about civilization and linguistic awareness (see 3.1.4).

2.2.3. Repetition task

In the case of Spanish, the repetition task is particularly challenging due to the pluricentricity of the language. Like English, Spanish is usually taught following both European and American models; several varieties (among them, e.g., Madrid, Mexico City, and Buenos Aires) can thus be considered as pronunciation norms ([18]). This linguistic diversity is mirrored in ordinary teaching practice, where learners get input from speakers of different origin. For this reason, the 24 subjects tested in our pilot study (Osnabrück and Vienna, 2015) were presented with a mixed auditory input containing productions from model speakers of the three above-mentioned varieties. As might be expected, the learners tried to reproduce the items independently of their individual target pronunciation and, e.g., repeated *vainilla*, produced as [baj'niʃa] by the Argentinean model speaker (which is quite easy for germanophone learners due to phonemic /ʃ/ in their L1), even if they otherwise had no contact with this variety. In order to avoid choosing one model variety, we shall substitute the repetition of Spanish words with a reproduction task involving the repetition of appropriate logatomes, aiming to test the learners' aptitude to perceive and produce the sounds of (various dialects of) Spanish. A disadvantage of such a task, however, consists in the fact that it abstracts from language-specific frequency effects (e.g., /s/-weakening in frequent *gracias* vs. infrequent *ciempiés*).

3. (I)FEC guidelines

(I)FEC works with a modular system: the basic design (see 3.1) follows PFC in eliciting data via two reading tasks (word list and text) and a semi-focused interview. An innovation is the discourse completion task for eliciting prosodic data. Like

in PFC, the informants fill in a socio-demographic questionnaire (different versions for L1 speakers and learners) and a (German and Spanish) agreement. For some types of speakers (learners, illiterates, multilinguals), we provide supplementary tasks (see 3.2). Concerning the recordings as well as digitization and (partial) transcription of the data, we are currently testing the software SpeechRecorder ([19]).

3.1. FEC basic design

3.1.1. Word list

The FEC word list includes 125 words, among them 6 (pseudo-)minimal pairs, which makes it slightly longer than the PFC list. We systematically consider the realization of Spanish phonemes in (nearly) all positions. Special focus lies on the regionally confined oppositions. Furthermore, we test phonological processes, word stress and the influence of the graphic form on the pronunciation ([20], [21], [22], [23]).

Among the phenomena related to the phoneme system, one of the most common neutralizations is *seseo*, i.e. the loss of the opposition /s/:/θ/ in favor of /s/. One test ground for this is the minimal pair *la casa* 'the house' vs. *la caza* 'the hunt'. The traditional norm, postulating the Castilian model for the whole Spanish-speaking world, distinguishes [ˈkasa] and [ˈkaθa]. However, speakers of almost all American varieties as well as informants from parts of Andalusia and the Canary Islands consistently produce homophonic [ˈkasa]. Other tested oppositions comprise /ʝ/:/ʎ/ (which is abandoned in favor of /ʝ/ in so-called *yeísmo* varieties) and /ɾ/:/r/ (tap vs. trill).

In addition to the phoneme oppositions, the word list elicits phonological processes. Regarding /s/-weakening, it contains e.g. the items *la casa* [laˈkasa] 'the house' and the corresponding plural form *las casas* [lasˈkasas]. Since final /s/ is not only aspirated (realized as [h]), but often elided in various regions of the Spanish-speaking world (e.g., Andalusia, Canary Islands, *tierras bajas* 'lowlands' of Latin America), the opposition between singular and plural may become inaudible with [laˈkasa] serving for both. Among other phonological processes to be tested are vowel-weakening, hiatus resolution, glide formation and the spirantization of voiced stops.

Even a few prosodic features can be tested through the word list. We included the minimal triplet *número* [ˈnumeɾo] 'number', *numero* [nuˈmeɾo] 'I number' und *numeró* [numeˈɾo] 's/he numbered', on which the graphic accent (or its absence) marks the stressed syllable. In contrast, the stimulus *¡TOMATELO!* lacks this information because of the expressive capitalization.

At the end of the word list, we put two words which do not form a minimal pair in any norm: *barón* 'baron' and *varón* 'man'. Both should be pronounced identically [baˈɾon]. However, some American speakers present a distinction based on the graphics <b>:<v>, i.e. [baˈɾon] vs. [vaˈɾon] ([20], p. 3). The same is likely to occur in learner data due to the close intertwinement of orthography and pronunciation in instructed foreign language learning.

In what follows we reproduce the complete word list (which is not presented to the informants in this compact form, but as a list with one word per line or as a *PowerPoint* presentation with one word per slide):

1. *continúa*, 2. *reloj*, 3. *viuda*, 4. *tabúes*, 5. *estudiéis*, 6. *querría*, 7. *caída*, 8. *pacto*, 9. *miau*, 10. *chalet*, 11. *jinete*, 12. *rehusa*, 13. *numeró*, 14. *toros*, 15. *guau*, 16. *muy*, 17. *flor*, 18. *ríe*, 19. *hoy*, 20. *juzgar*, 21. *signo*, 22. *labio*, 23. *deuda*, 24. *queja*, 25. *ketchup*, 26. *o hay*, 27. *ladrón*, 28. *club*, 29. *vainilla*, 30. *la papa*, 31. *iceberg*, 32. *número*, 33. *vacuo*, 34. *ángel*, 35. *afgano*, 36. *plan*, 37. *la caza*, 38. *logro*, 39. *un yunque*, 40. *mismo*, 41. *coñac*, 42. *el vino*, 43. *admirar*, 44. *un sueño*, 45. *buey*, 46. *él vino*, 47. *tengo*, 48. *montón*, 49. *álbum*, 50. *esdrújulo*, 51. *bou*, 52. *yo lo sé*, 53. *un chico*, 54. *algo*, 55. *diurno*, 56. *ahí*, 57. *la tapa*, 58. *enfermo*, 59. *diablo*, 60. *caudal*, 61. *nadie*, 62. *¡TOMATELO!*, 63. *causa*, 64. *búho*, 65. *la tira*, 66. *llave*, 67. *perro*, 68. *caldo*, 69. *suntuoso*, 70. *guante*, 71. *cuidar*, 72. *óptimo*, 73. *ñandú*, 74. *baile*, 75. *drama*, 76. *vienes*, 77. *gracias*, 78. *oído*, 79. *la casa*, 80. *ración*, 81. *tan blanco*, 82. *ciempiés*, 83. *deshielo*, 84. *muchacho*, 85. *salud*, 86. *palas*, 87. *rosbif*, 88. *pastel*, 89. *con agua*, 90. *quería*, 91. *paz*, 92. *étnico*, 93. *champán*, 94. *honra*, 95. *un tío*, 96. *obtiene*, 97. *la quita*, 98. *baúl*, 99. *la pita*, 100. *pero*, 101. *la capa*, 102. *oye*, 103. *reír*, 104. *tenue*, 105. *lleno*, 106. *las casas*, 107. *Europa*, 108. *allí*, 109. *numero*, 110. *los otros*, 111. *cambiáis*, 112. *virrey*, 113. *diurético*

The list ends with the (pseudo-)minimal pairs:

114. *numero*, 115. *número*, 116. *numeró*, 117. *la caza*, 118. *la casa*, 119. *las casas*, 120. *ahí*, 121. *allí*, 122. *pero*, 123. *perro*, 124. *barón*, 125. *varón*

3.1.2. Text

The 381 words text constructed for the FEC project on the basis of the word list (and especially the minimal pairs) is comparable to the PFC text. 16 words from the list reappear in the text (highlighted):

Un sueño bastante animal

Normalmente nunca me acuerdo de mis *sueños*. *Pero* lo de la noche pasada me *causa* una gran incógnita: Parece un día como otro y voy caminando hacia mi trabajo. De repente, escucho el fuerte ladrido de un *perro*, que viene de *la casa* de un vecino. Es un hombre bastante raro, de quien se dice que posee un gran *número* de animales: además de algunos *perros*, gatos y pollos, como cualquiera en el barrio, también tiene una admirable colección de insectos. Mientras más me acerco a *la casa*, más aumentan los ladridos y veo que causan un caos enorme, tanto así que los demás vecinos salen de *sus casas* para ver qué es lo que pasa.

La situación se agrava aún más cuando la viuda del Doctor *Numeró*, un profesor de matemáticas ya jubilado, llama a la perrera, que llega de inmediato al lugar de los hechos para inspeccionar. ¡Qué sorpresa cuando entran en *la casa*! Tan suntuosa por fuera, y por dentro parece un zoológico obtenido por *la caza* nocturna en selvas y pantanos. Todo se sale completamente de control cuando el primero de los perreros entra a *la casa* y cae al suelo. En seguida comienza a gritar que *algo* lo pica: es un *ciempiés*. ¡*Gracias* a Dios, el resto de su familia ya se encuentra en el estómago de un bulldog! Cuando entran *los otros* perreros, el *perro* finalmente se calla. Todos los insectos restantes están esparcidos por el suelo y causan pánico entre los gatos y los pollos, que tratan de subirse a las estanterías.

Al final, el jefe de los perreros, preocupado por hacer una investigación perfecta, cuenta los animales, y se sorprende de que además de los perros, gatos y pollos se encuentran cinco tipos diferentes de insectos, o no, tal vez diez, o quince, o veinte: mariposas, abejas, libélulas, saltamontes, y además cerdos, ranas y elefantes, ranas rojas y elefantes amarillos, que

empiezan a hablar, sí, discuten sobre el peinado de los perreros y los *planes* de renovación de *la casa* … y entonces yo me encuentro *allí*, entre todos, tratando de entender las diferentes discusiones. En ese momento suena mi despertador. ¿Qué *diablos* trata de decirme mi subconsciente con semejante *sueño* sobre mis traumas de niño y deseos ocultos?

3.1.3. Discourse Completion Task

In order to systematically elicit a certain amount of intonation patterns, we added a discourse completion task ([24]). This inductive method consists of confronting the speakers with a series of hypothetical everyday situations to which they are supposed to react verbally. The speakers may phrase their verbal reactions as they wish, e.g., Interviewer: *Entra en una tienda donde nunca estuvo antes y pregunta si tienen mandarinas.* 'You enter a shop where you have never been before and you ask if they sell tangerines.' Subject: *¿Tiene(n) mandarinas?* / *¿Hay mandarinas, por casualidad?* / *Hola, ¿mandarinas, tenés?* etc. 'Do you have tangerines?' (possible responses). The learners are presented with a simplified version of this task consisting in the reproduction of a given answer.

3.1.4. Interview

The 20-minute interviews follow particular guidelines in FEC and IFEC. FEC focuses on civilization (e.g., *¿Qué lugares recomendarías visitar en tu ciudad o pueblo?* 'Which places in your town/village would you recommend to visit?') and linguistic awareness (e.g., *¿A través de qué características de la forma de hablar se reconoce a alguien de tu región?* 'Which characteristics of speech reveal the origin of a speaker from your region?'). In the case of IFEC, language learning is particularly interesting (*¿Qué te parece difícil en español?* 'What seems difficult to you in Spanish?'). The guidelines are adapted to the competence levels: A1, A2–B1 and B2–C2.

3.2. Supplementary tasks

In addition to the basic design, we currently develop supplementary tasks for learners (IFEC), illiterates and multilinguals. For learners, we plan a logatome list, the reading of a constructed text in German (accompanied by a short summarizing task), a detailed interview on linguistic representations and attitudes in German and, finally, a phonological awareness test in the course of which the learners are presented with their own productions and asked to comment on them (see [25]). For illiterates, a picture list and a picture story are in preparation, for multilinguals either the reading of *The North Wind and the Sun* or specifically constructed supplement tasks will be applied.

4. Acknowledgements

We gratefully acknowledge Duygu Durmus, Kristina Dziallas, Marie-Antoinette Goldberger, Isabella Rechberger, Laura Strokorb and Franziska Stuntebeck for collecting the data of the pilot studies as well as all speakers for their disposition to participate in these.

5. Bibliography

[1] http://www.projet-pfc.net/.

[2] J. Durand, B. Laks and Ch. Lyche, "Le projet PFC. Une source de données primaires structurées", J. Durand, B. Laks and C. Lyche (eds.), *Phonologie, variation et accents du français*, Paris, Hermès, pp. 9–61, 2009.

[3] S. Detey, J. Durand, B. Laks and Ch. Lyche, "The PFC programme and its methodological framework", S. Detey, J. Durand, B. Laks and Ch. Lyche (eds.), *Varieties of Spoken French*, Oxford, Oxford University Press, 2016.

[4] I. Racine, S. Detey, F. Zay and Y. Kawaguchi: "Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2. L'exemple du projet 'Interphonologie du français contemporain' (IPFC)", A. Kamber and C. Skupiens (eds.), *Recherches récentes en FLE*, Bern, Lang, pp. 1–19, 2012.

[5] C-ORAL-ROM: E. Cresti and M. Moneglia (eds.), *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam, Benjamins, 2005.

[6] CREA: Real Academia Española, *Corpus de referencia del español actual (CREA)*. http://www.rae.es/recursos/banco-de-datos/crea.

[7] *Corpus Oral de Español como Lengua Extranjera* (ELE), http://cartago.lllf.uam.es/corele/home_es.html.

[8] Piñeros, Carlos Eduardo, *Dialectoteca del Español,* http://dialects.its.uiowa.edu/#.

[9] E. Martínez Celdrán and A. Fernández Planas (coord.), *Atlas Multimedia de la Prosodia del Espacio Románico* (AMPER), 2003–2015. http://stel.ub.edu/labfon/amper/cast/index.html.

[10] C. Gabriel, Hamburg Corpus of Argentinean Spanish (HaCASpa), 2011. http://www.corpora.uni-hamburg.de/sfb538/en_h9_hacaspa.html.

[11] P. Prieto and P. Roseano (eds.), *Atlas interactivo de la entonación del español*, 2009–2013. http://prosodia.upf.edu/atlasentonacion/.

[12] J. Durand and A. Przewozny, "La phonologie de l'anglais contemporain: usages, variétés et structure", *Revue française de linguistique appliquée* XVII, pp. 25–37, 2012.

[13] W. Labov, *Sociolinguistic Patterns*, Philadelphia: University of Pennsylvania Press, 1972.

[14] P. Boersma and D. Weenink, *Praat. Doing phonetics by computer*. http://www.fon.hum.uva.nl/praat/ [Computer program. Version 6.0.10], 2016.

[15] J. Eychenne and R. Paternostro, "Analyzing transcribed speech with Dolmen", S. Detey, J. Durand, B. Laks and C. Lyche (eds.), *Varieties of Spoken French. A source book*, Oxford, Oxford University Press, pp. D35–D52, 2016.

[16] International Phonetic Association, *Handbook of the International Phonetic Association.* Cambridge, CUP, 1999.

[17] DFG-Handreichung "Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora", http://www.dfg.de/foerderung/antragstellung_begutachtung_entscheidung/antragstellende/antragstellung/nachnutzung_forschungsdaten/index.html.

[18] W. Oesterreicher, "Plurizentrische Sprachkultur – der Varietätenraum des Spanischen", *Romanistisches Jahrbuch* 51, pp. 281–311, 2000.

[19] Ch. Draxler and K. Jänsch, "SpeechRecorder – A Universal Platform Independent Multi-Channel Audio Recording Software", Proc. LREC 2004, Lisbon.

[20] D. L. Canfield, *La pronunciación del español en América*, Bogotá, Instituto Caro y Cuervo, 1962.

[21] J. I. Hualde, *The Sounds of Spanish*, Cambridge, Cambridge University Press, 2005.

[22] Real Academia Española / Asociación de Academias de la Lengua Española, *Nueva gramática de la lengua española. Fonética y fonología*, Madrid, Espasa Libros, 2011.

[23] C. Gabriel, T. Meisenburg and M. Selig, *Spanisch: Phonetik und Phonologie. Eine Einführung*, Tübingen, Narr, 2013.

[24] J. C. Félix-Brasdefer, "Data collection methods in speech act performance. DCTs, roleplays, and verbal reports", A. Martínez-Flor and E. Usó-Juan (eds.), *Speech act performance. Theoretical, empirical, and methodological issues*, Amsterdam, Benjamins, pp. 41–56, 2010.

[25] A. G. Osburne, "Pronunciation strategies of advanced ESOL learners", *International Review of Applied Linguistics in Language Teaching* 41, pp. 131–141, 2003.

AI vs. AU in American English compared to German

Renate Raffelsiefen, Anja Geumann

Institut für Deutsche Sprache, R5, 6-13, D-68161 Mannheim
raffelsiefen@ids-mannheim.de, geumann@ids-mannheim.de

Abstract

American English and German AI, AU observed in cognates such as *Wein, wine, Haus, house* are usually treated on a par, represented with the same initial vowel (cf. [aɪ], [aʊ] for Am. Engl. and German [1]). Yet, acoustic measurements indicate differences as the relevant trajectories characteristically cross in Am. Engl. but not in German. These data may indicate consistency with the same initial target for these diphthongs in German, supporting the choice of the same symbol /a/ in phonemic representation, as opposed to distinct targets (and distinct initial phonemes) in American English.

Index Terms: corpora, phonetics, phonology, diphthongs

1. Introduction

Phonemic theory is rooted in the intuition of a single level of abstraction, where speech sounds have identical representations for as long as phonetic differences between them can be attributed to context [2]. The question arises then of what the conditions are for determining whether or not phonetic differences can be attributed to context.

Diphthongs, which are defined by a movement from a starting position to a different finishing position within the syllable, appear to be particularly prone to coarticulation among its two members. Here we focus on the initial member of AI- versus AU-type diphthongs in words like *wine, house* in American English and *Wein, Haus* in German. In both languages the F1/F2 trajectories indicate distinct turning points associated with the respective initial members towards the positions associated with the following members (assumed to be /i/ versus /u/, respectively). We suggest that the relevant differences among those turning points for German can be attributed to the distinct position associated with the second diphthong member, whereas such an analysis seems highly questionable for the respective turning points in American English. Independent phonological evidence for the sameness of the initial members of AI and AU in German, as opposed to English, as well as additional allophonic relations of those members to independent monophthongs will also be discussed.

2. Data

The data were taken from two corpora of read speech [3], [4]. (TIMIT American English, Northern, 31 female speakers, Southern, 36 female speakers, Kielcorpus, Standard German, 26 female speakers), manually annotated. Formant values were extracted automatically with PRAAT [5] at 10 equidistant points between 5-95% of the acoustic vowel duration, Burg algorithm, 5 formants, with 5500 Hz as maximum formant search range. The contexts for the target vowels were not checked but considered to be fairly representative for a wide range of occurrences. The examples we use here in the text are not necessarily contained in the corpora.

3. Discussion

The trajectories for the German diphthongs AI and AU in Figure 1 indicate distinct turning points marking the respective targets of the initial diphthong members (cf. also [6]). The further back articulation of the initial vowel in AU compared to that in AI can be explained with reference to context, indicative of anticipatory retracting (as well as raising) of the tongue body to produce the following back high vowel. As a result of being analysable in terms of modifications of the timing of articulatory gestures conditioned by context the relevant differences qualify as allophonic, supporting the same initial phoneme in German AU and AI.

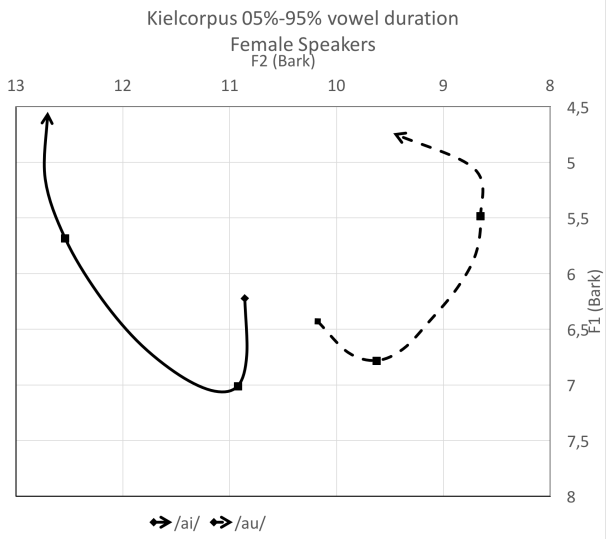


Figure 1: *Germ. AI (cont.), AU (dashed) 5% - 95% of vowel duration, square marks indicate 25% and 75% as in all the following figures, symbols indicate phonemes.*

Both the assumption that German AI and AU are biphonemic and that they share the same first vowel phoneme are supported by phonological evidence from reduplication ([7]). The data in (1) illustrate relevant word formation patterns expressing exasperation in German, where the vowel in the nucleus is repeated and the sequence is separated by /h/ (1a). The fact that for base words containing a diphthong only the initial vowel is reduplicated (cf. (1b)) indicates a biphonemic structure of diphthongs. The observation that the respective initial syllables in reduplicated words based on AI and AU are entirely homophonous in (1c) (boldfaced) supports the assumption of identical vowel phonemes to represent the

initial members in the relevant diphthongs. The observation that the bodfaced syllables are moreover homophonous to the initial syllables in reduplicated words based on the monophthong /a/ as in (1d) supports the choice of the centralized low monophthong as in *kalt* 'cold' to represent the initial diphthong member in German AI and AU.

(1a) /ja/ 'yes' -> /jaha/
(1b) /nain/ <nein> 'no' -> /nahain/
(1c) /kain/ → /**ká**.hàin/ <kein> 'no'
/kaum/ → /**ká**.hàum/ <kaum> 'barely'
(1d) /kalt/ → /**ká**.hàlt/ <kalt> 'cold'

The relation between the trajectories of German AI and AU to those representing the monophthongs as in /zat/ <satt> 'full' versus /zɑt/ <Saat> <seed> is shown in Figure 2.

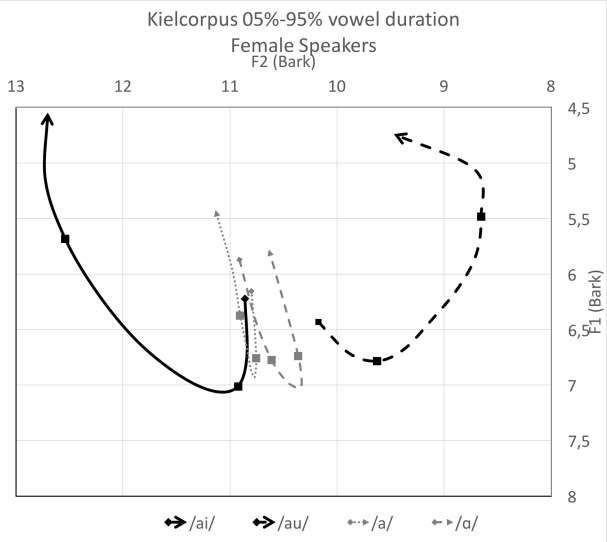


Figure 2: *Germ. AI (cont.), AU (dashed) and two vowel monophthongs.*

The data support the choice of centralized /a/ to represent the initial diphthong member in German AI and, assuming the coarticulatory influence from the following /u/ discussed above, are moreover consistent with positing this vowel as the initial member of AU. Despite the closer vicinity of the peripheral vowel /ɑ/ to the turning point marking the initial member of AU there are two reasons for positing centralized /a/ in both diphthongs. There is a general constraint against peripheral vowels in closed syllables in German, which rules out tautosyllabic /ɑu/. Also the fact that the initial vowel in both AI and AU is short in words like *Haus* and *Wein* indicates initial centralized /a/, since peripheral vowels are always subject to phonetic lengthening in stressed syllables in German.

(American) English differs from German in that the trajectories of AI and AU characteristically cross, overlapping strongly in some regions (Figure 3 Southern), less so in others (Figure 4 Northern). In contrast to German it is hence the initial vowel in the diphthong AI, which is articulated further back than that in *AU* [8, p.1572], [9, p. 162].

It is questionable whether or not this difference is consistent with positing the same phoneme for the initial diphthong member: presumably it resists explanation in terms of modifications of articulatory gestures conditioned by the relevant second members. Assuming that AI and AU in English are also biphonemic we suggest then that the phonemes representing their respective initial member differ.

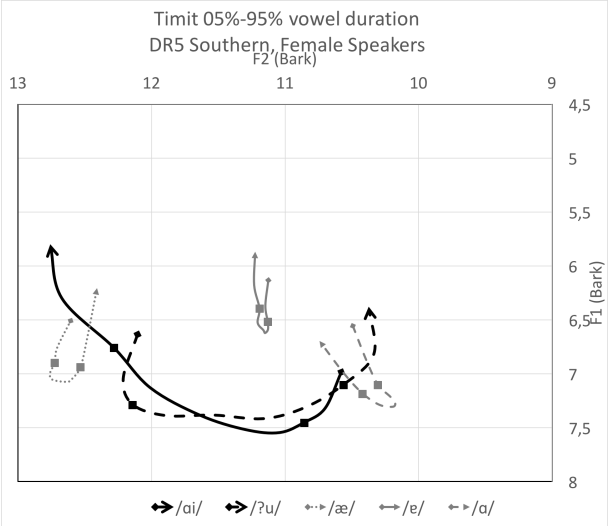


Figure 3: *Am.E. (Southern) AI (cont.), AU (dashed) and three vowel monophthongs.*

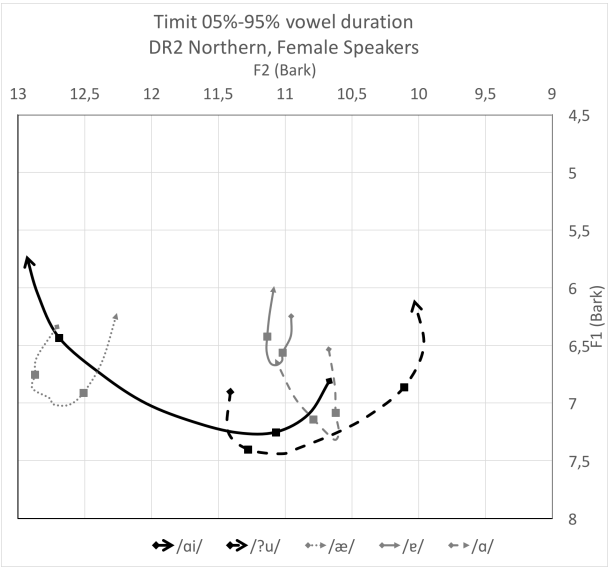


Figure 4: *Am.E. (Northern) AI (cont.), AU (dashed) and three vowel monophthongs.*

As for establishing allophonic relations to monophthongs the vowel as in *father*, transcribed as /ɑ/ in Figure 3, appears to be a plausible candidate for the initial member in AI (but not AU). AI is therefore phonemically represented as /ɑi/ in Figure 3. The allophonic relation between the initial member of AU and monophthongs appears to be much harder to establish. The vowel /æ/ (as in *gather*) might be the most plausible candidate to represent the initial member of AU in Southern Am.E., whereas /ɐ/ (as in *mother*) might be more plausible for Northern Am. E..

The latter choice is weakly supported by historical alternations (2a), which indicate the loss of the second diphthong member to reduce complex syllable structure in certain contexts, and a historical sound change (2b), which results from the loss of the second diphthong member /u/

before labials. In both cases of historical reduction of AU involving the loss of the second member, the monophthong /ɐ/ as in *mother* persists.

(2a) profound – prof/ɐ/ndity 'profundity', south -s/ɐ/thern 'southern'

(2b) pl/ɐ/m 'plum' (cf. *Pflaume*), d/ɐ/ve 'dove' (cf. *Taube*)

The initial member of the diphthong AU is represented with a question mark in Figure 3 and 4, to express our lack of certainty. In general there is a question of how to determine whether or not differences seen in phonetics can be attributed to context and therefore are consistent with positing a single phoneme.

4. Bibliography

[1] Peter Ladefoged, *Vowels and Consonants*. Second ed. Blackwell Publ.: Malden, MA et.al., 2005.

[2] Ferdinand de Saussure, *Cours de linguistique générale*. Paris: Payot, 1916.

[3] John S. Garofolo, et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. LDC93S1. Web Download. Linguistic Data Consortium: Philadelphia, 1993.

[4] Klaus J. Kohler, (Ed.) "Phonetisch-Akustische Datenbasis des Hochdeutschen", *Kieler Arbeiten zu den PHONDAT-Projekten 1989-1992. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 26, 1992.

[5] Paul Boersma, David Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.0.17, retrieved 21 April 2016 from http://www.praat.org/.

[6] Klaus J. Kohler, (1977) *Einführung in die Phonetik des Deutschen.* Erich Schmidt Verlag: Berlin.

[7] Renate Raffelsiefen, Fabian Brackhane, "Motivating phonemic structure: the case of diphthongs in German", *Ms*. IDS Mannheim, 2016

[8] Thomas Gay, "Effect of speaking rate on diphthong formant movements", *Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1570-1573, 1968.

[9] Hartwig Eckert, William Barry, *The Phonetics and Phonology of English and Pronunciation: A Coursebook*. Wissenschaftlicher Verlag Trier: Trier, 2005.

Acoustic and articulatory manifestations of final lengthening and voicing contrasts for German learners of English as a second language

Oxana Rasskazova [1], Malte Belz [1] , Christine Mooshammer [1,2], Jelena Krivokapić [2,3]

[1] Humboldt-Universität zu Berlin, Germany
[2] Haskins Laboratories, USA
[3] University of Michigan, USA
oxana.rasskazova@hu-berlin.de, malte.belz@hu-berlin.de

Abstract

The present acoustic and articulatory study investigates whether German native speakers show prosodic and segmental transfer effects of the L1 when speaking English as L2. The focus lies on prosodic lengthening patterns as well as voicing contrast in word-final position in English, a pattern that is difficult for many German learners due to their native syllable-final obstruent devoicing rule.

Index Terms: prosody, final lengthening, final devoicing, English as a second language

1. Introduction

Most segments show temporal variation for prosodic changes, see [1, 2, 3, 4]. For example, there are indications that consonants as well as vowels tend to be longer at prosodic boundaries than phrase-medially, e.g., [5, 6]. However, different segments and segment classes are not always affected in a uniform way. For example, [7] and [8] show that tense but not lax vowels in German stretch in stressed syllables and compress for fast speech rate.

In our previous acoustic and articulatory studies [9, 10], we investigated the effects of phrasal boundaries on the temporal characteristics of preceding segments in German as well the interaction of tenseness and final lengthening. The results show that, contrary to speech rate and stress effects, lax vowel in phrase-final position lengthen, although less than tense vowels. The articulatory closing gesture for consonants as well as the duration of constriction phase is longer for phrase-final compared to phrase-medial position.

In the present production study we will investigate two aspects that might contribute to a pronounced L2 accent: In our first experiment we will focus on whether German native speakers deviate from their lengthening patterns found for German when speaking English. Both, English and German show final lengthening effects, but it is unclear whether lengthening also interacts with tenseness. As a null hypothesis, we expect that subjects show the same prosodic features as in German when speaking English, e.g., lax vowels will show a lengthening effect in phrase-final positions, but less than tense vowels (Hypothesis 1).

Furthermore, we are interested in whether German native speakers transfer the phonological pattern of word-final obstruent devoicing into English or whether they produce a voicing contrast in word-final positions. [11] found that some acoustic patterns of L1 German, such as consonant closure durations for voiceless versus voiced obstruents, were transferred to L2 English, whereas other acoustic parameters correspond to phonological patterns typical for English, such as longer vowel durations before voiced obstruents. However, the interaction between tenseness of the preceding vowel and post-vocalic voicing was not addressed in [11]. Since lax vowels are temporally more inflexible in German than in English [7] we assume that there will be a stronger transfer of German word-final obstruent devoicing for words with lax vowels than for words with tense vowels (Hypothesis 2). We investigate whether German native speakers neutralize voiced word-final consonants in English by means of acoustic as well as articulatory data.

2. Method

2.1. Speakers and Stimuli

Acoustic and articulatory data of 8 native German participants were recorded in a sound proof cabin by means of EMA (AG 501, Carstens Electronics). Sensors were attached to tongue tip, mid and back, the jaw, the lips and four reference sensors for compensating for head movements. The speakers (4 male, 4 female) are 23–28 years old and advanced English learners. German and English sentences with target words containing minimal pairs differing in vowel tenseness and consonant voicing were presented on a monitor, with 5 repetitions, respectively. Target sentences were presented in a random order and mixed with filler sentences. The target words were embedded in two boundary strength contexts: phrase-medial (1) and phrase-final (2), e.g.:

(1) We should wait for a ***beat*** in any event. I wouldn´t know when to start.

(2) We should wait for a ***beat***. In any event, I wouldn´t know when to start.

For a comparison of the lengthening effect in L1 German and L2 English, we compared the minimal pairs *beat/bit* and *Beet/Bett* (Engl. ‘bed’ (botan.), ‘bed’). For the voicing contrast part, we analyzed the English minimal pairs *beat/bead*, *bit/bid*, *seat/seed*, and *sit/Sid*.

2.2. Analysis

Tongue tip movements were labelled for the closing gesture duration towards alveolar consonants and closure duration in the target word using *mview* (Mark Tiede, Haskins Laboratories). Closing duration is defined as the time span of closing movement onset and plateau onset for the word-final consonant by using a 20% threshold criterion based on the tangential velocity of the tongue tip signal. Closed phase

duration is defined as the time span between plateau onset and plateau offset.

The acoustic measurements of vowel duration preceding voiced and voiceless consonants, consonant closure duration, duration of voicing into closure and VOT was carried out in Praat [12]. All statistics were carried out using R 3.3.0 [13] with the packages lme4 [14] and lmerTest [15].

3. Results

3.1. Comparing final lengthening in L1 German and L2 English for German native speakers

To analyze whether German native speakers exhibit the same degree of final lengthening for the minimal pairs *beat/bit* and *Beet/Bett*, we calculated a linear mixed-effect model for acoustic vowel duration (logarithmic) with target language (German vs. English), tenseness (tense vs. lax), and phrasal condition (medial vs. final) and their interactions as fixed effects and participants as random effects (cf. Fig. 1). Significant differences are found for phrase-final vs. phrase-medial position ($\beta = .33$, se = .03, $p < .001$), for English vs. German targets ($\beta = .19$, se = .03, $p < .001$), and for tense vs. lax vowels ($\beta = .58$, se = .03, $p < .001$), as well as for the interaction of position and vowel ($\beta = -.17$, se = .04, $p < .001$).

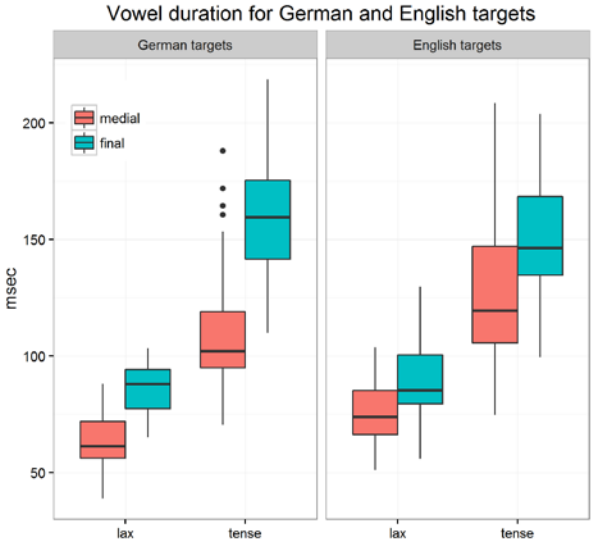


Figure 1: *The effect of tenseness and phrasal condition on acoustic vowel duration in L1 (German,* Beet/Bett*) and L2 (English,* beat/bit*). Red (left) boxes represent targets in phrase-medial position; blue (right) boxes represent targets in phrase-final position.*

The same minimal pairs are also analyzed for the articulatory consonant closing and plateau duration. We calculated a mixed-effect model with the same fixed and random effects structure as for vowel duration. For logarithmic closing duration, no significant effect is found for phrase-final position or tenseness. However, English targets exhibit a significant longer closing duration than German targets ($\beta = .58$, se = .04, $p < .001$), cf. Fig. 2. The interaction between position and tenseness is significant ($\beta = .3$, se = .06, $p < .01$), because closing duration only lengthens following tense vowels in final position.

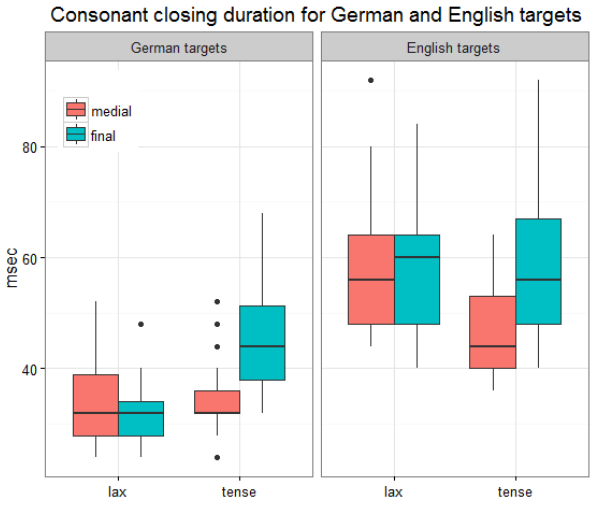


Figure 2: *The effect of tenseness and phrasal condition on articulatory consonant closing duration in L1 (German,* Beet/Bett*) and L2 (English* beat/bit*). Red (left) boxes represent targets in phrase-medial position; blue (right) boxes represent targets in phrase-final position*

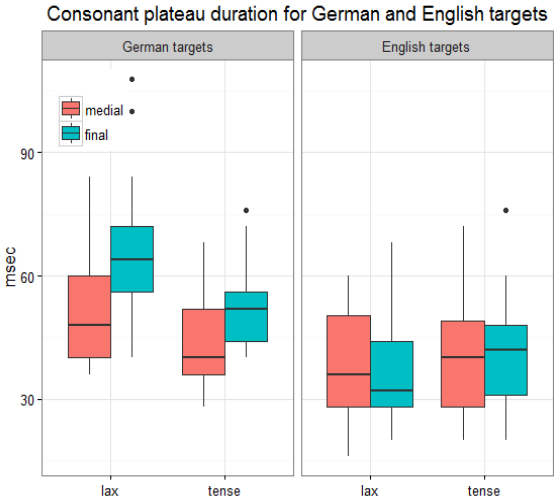


Figure 3: *The effect of tenseness and phrasal condition on articulatory consonant plateau duration in L1 (German,* Beet/Bett*) and L2 (English,* beat/bit*). Red (left) boxes represent targets in phrase-medial position; blue (right) boxes represent targets in phrase-final position.*

The model for consonant plateau duration (cf. Fig. 3) displays a significant effect on plateau duration for phrase-final position ($\beta = .21$, se = .05, $p < .001$). Plateau duration of English targets is significantly smaller than for German targets ($\beta = -.38$, se =.05, $p < .001$). Closures following tense vowels are significantly shorter than following lax vowels ($\beta = -.17$, se = .05, $p < .01$). For English targets in phrase-final position the plateau duration is significantly smaller than for the corresponding German targets ($\beta = -.21$, se = .08, $p < .05$). The interaction between targets and tenseness is significant ($\beta = .28$,

se = 07, p < .001), as plateau duration for English targets shows no effects of tenseness and position.

3.2. Voicing contrast in L2 (English)

To investigate whether German subjects produce a voicing contrast in word-final positions when speaking L2 English, we analyzed the acoustic measures vowel duration preceding voiced vs. unvoiced consonants, consonant closure duration, duration of voicing into closure and VOT and the articulatory measures consonant closing and plateau duration.

A linear mixed-effects model for vowel duration on a logarithmic scale with phrasal condition, tenseness and voicing as well as their interactions as fixed effects, and targets as well as subjects with random slopes for voicing as random effects reveals significant main effects of (phrase-final) condition (β = .28, se = 3.5, p < .001) and tense vowels (β = .48, se = .03, p < .001), but not for voiced vs. voiceless consonants and for none of the interactions (cf. Fig. 4).

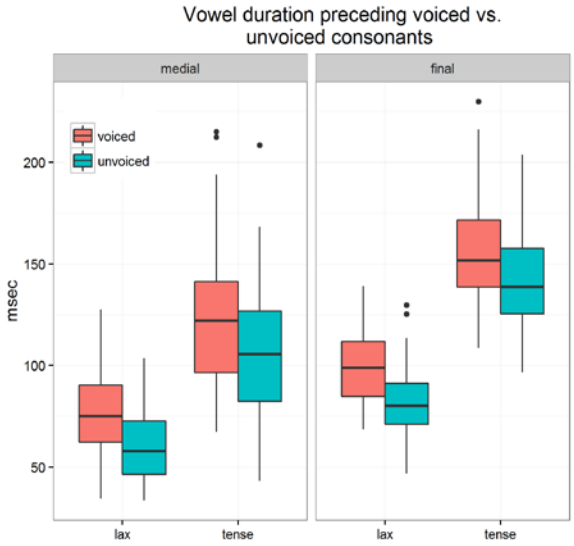


Figure 4: *Lax vs. tense acoustic vowel duration preceding voiced vs. unvoiced consonants in phrase-medial vs. phrase-final position. Red (left) boxes represent voiced targets; blue (right) boxes represent unvoiced targets.*

Logarithmic closing duration show a significant final lengthening effect (β = .16, se = .03, p < .001). There are no significant main effects of tenseness and voicing, but a three-way interaction of tenseness, position, and voicing (β = .17, se = .07, p < .05, cf. Fig. 5).

Logarithmic plateau duration show no main effects for voicing, tenseness or position. There is a three-way interaction of phrase-final position, preceding tense vowel, and voiced consonant targets (β = -.27, se = .13, p < .05).

For the acoustic measures during the stops, the logarithmic duration of voicing during consonant closure (cf. the red bars in Fig 6) is shorter for unvoiced targets (β = -.26, se = .07, p < .001). This effect is enhanced in final positions (β = -.26, se = .1, p < .05).

The logarithmic acoustic closure duration (cf. the green bars in Fig. 6) is longer in phrase-final position (β = .26, se =09, p < .01) and smaller for tense compared to lax vowels (β = -.5, se = .1, p > .001). There are interactions of position and tenseness (β = .44, se = .13, p < .01) and of tenseness and voicing (β = .36, se = .13, p < .01) because no voicing contrast is found for consonants following lax vowels in medial position.

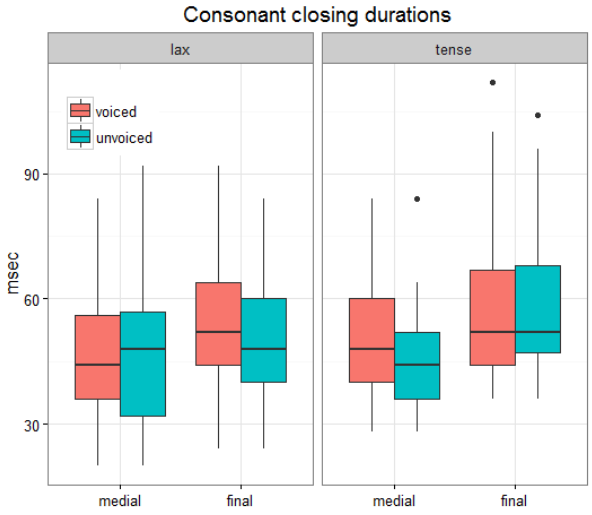


Figure 5: *Articulatory closing duration for consonants following lax vs. tense vowels in phrase-medial vs. phrase-final position. Red (left) boxes represent voiced targets; blue (right) boxes represent unvoiced targets.*

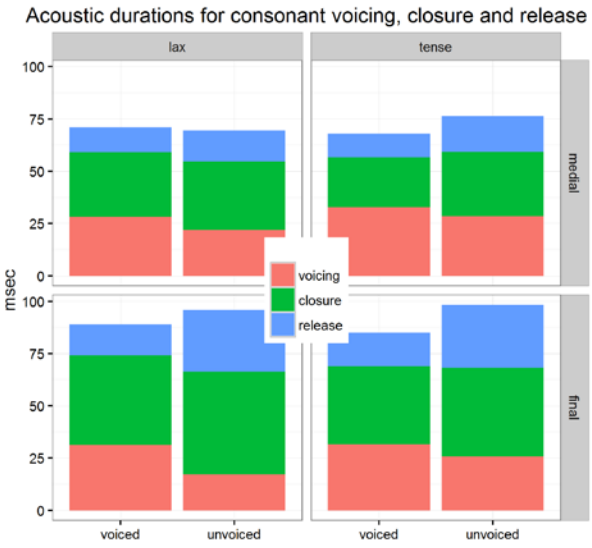


Figure 6: *Acoustic durations for voicing into closure (red), consonant closure duration (green) and consonant release burst including aspiration (blue), split for tenseness and position.*

The logarithmic aspiration duration (cf. the blue bars in Fig. 6) exhibit a longer duration for unvoiced targets (β = -.26, se = .07, p < .001), and an interaction of position and voicing (β = -.26, se = .1, p < .05) as this effect is stronger for phrase-final positions.

4. Discussion and Conclusion

German speakers show similar prosodic behavior concerning final lengthening in acoustic vowel duration when speaking L1 German and L2 English, confirming Hypothesis 1. For both target languages, vowel duration is lengthened in phrase-final position, but lax vowels lengthen less than tense vowels. For the English targets, German learners exhibit longer consonant closing durations (cf. Fig. 2), with a stronger lengthening effect for closing duration following tense vowels in phrase-final position. Unexpectedly, consonant plateau duration for the English targets is significantly shorter than for the German targets in final position. Generally, our results on final lengthening could be explained by either a transfer effect from L1 German to L2 English, by the advanced L2 English proficiency, or because German and English are similar with respect to final lengthening. In order to tease these possible causes apart, comparable data by native speakers of English are needed.

The second part of this study adheres to voicing contrast. The articulatory data for consonant closing and plateau duration of voiced vs. unvoiced targets are inconsistent. Together with the acoustic vowel length, no differences in voicing can be found. Looking at individual speakers, only two of the 8 subjects produce a pronounced difference in vowel duration before voiced vs. voiceless consonants. Thus, vowel length does not differ consistently for voiced vs. unvoiced consonants, despite the fact that pre-consonantal vowel lengthening is a reliable voicing cue in L1 English, cf. [11]. Significant differences in the voicing contrast manifest in the acoustic parameters during the alveolar stop: duration of voicing during stop consonants, consonant closure duration and aspiration. Voiced targets show a greater amount of voicing duration than unvoiced targets. The acoustic variables show more consistent differences for voicing than the articulatory data since voicing into closure and aspiration reflect glottal activity and lingual-glottal coordination.

Especially interesting are the interactions we found: Consonants after lax vowels in phrase-medial position show a reduced voicing distinction. This could be a reflex of the phonotactic asymmetry in German, see e.g., [16] that in word-medial position voiced obstruents are much more restricted after lax vowels than after tense vowels (e.g., words like *Ebbe)*. The result that the voicing distinction is more pronounced in final position is in line with [17] who found evidence for incomplete voicing neutralization in utterance-final positions in German.

In sum, this study presents first results on the prosodic variation and the voicing distinction, produced by German learners of English, and considered measurements of acoustic and articulatory data, different phrasal positions as well tenseness of vowels, factors that have not been taken into account in other studies, e.g., [11]. Future work will compare these results to native English speakers and will test whether and how deviations from the English norm contribute to a perceived foreign accent.

5. References

[1] M. E. Beckman, J. Edwards, and J. Fletcher, "Prosodic structure and tempo in a sonority model of articulatory dynamics," in *Papers in Laboratory Phonology II*, Papers in Laboratory Phonology, M. E. Beckman and J. Kingston, Eds. Cambridge: Cambridge University Press, 1992, vol. 2, pp. 68–86.

[2] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple prosodic boundaries," *Journal of Phonetics*, vol. 26, no. 2, pp. 173–199, 1998.

[3] D. Byrd, J. Krivokapić, and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effects," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1589–1599, 2006.

[4] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.

[5] R. Berkovits, "Durational effects in final lengthening, gapping, and contrastive stress," *Language and Speech*, vol. 37, no. 3, pp. 237–250, 1994.

[6] D. Byrd, "Articulatory Vowel Lengthening and Coordination at Phrasal Junctures," *Phonetica*, vol. 57, pp. 3–16, 2000.

[7] P. Hoole and C. Mooshammer, "Articulatory analysis of the German vowel system," in *Silbenschnitt und Tonakzente*, Linguistische Arbeiten, P. Auer, P. Gilles, and H. Spiekermann, Eds. Tübingen: M. Niemeyer, 2002, vol. 463, pp. 129–152.

[8] C. Mooshammer and S. Fuchs, "Stress distinction in German: simulating kinematic parameters of tongue-tip gestures," *Journal of Phonetics*, vol. 30, no. 3, pp. 337–355, 2002.

[9] M. Belz, O. Rasskazova, M. Weirich, A. Riemenschneider, J. Krivokapić, and C. Mooshammer, "Artikulatorische und akustische Untersuchungen zur finalen Längung im Deutschen." [Online]. Available: http://www.online.uni-marburg.de/pundp11/talks/Belz etal.pdf

[10] ——, "Final lengthening in German," Cornell University Ithaca, NY, 2016. [Online]. Available: http://www.labphon.org/labphon15/long abstracts/LabPhon15 Revised abstract 61.pdf

[11] B. L. Smith, R. Hayes-Harb, M. Bruss, and A. Harker, "Production and perception of voicing and devoicing in similar German and English word pairs by native speakers of German," *Journal of Phonetics*, vol. 37, no. 3, pp. 257–275, 2009.

[12] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341–345, 2001.

[13] R Core Team, "R: A language and environment for statistical computing," Wien, 2016.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[15] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest: tests in linear mixed effects models," 2015.

[16] C. Féry, "Final Devoicing and the stratification of the lexicon in German," in *Proceedings of HILP 4*, 1999.

[17] H. G. Piroth and P. M. Janker, "Speaker-dependent differences in voicing and devoicing of German obstruents," *Journal of Phonetics*, vol. 32, no. 1, pp. 81–109, 2004.

What is the fate of Scottish Vowel Length Rule in Glasgow?

Tamara Rathcke[1], *Florent Chevalier*[2], *Jane Stuart-Smith*[3]

[1] English Language and Linguistics, University of Kent
[2] Modern Languages and Cultures, University of Glasgow
[3]English Language, University of Glasgow

t.v.rathcke@kent.ac.uk, florent.chevalier@glasgow.ac.uk, jane.stuart-smith@glasgow.ac.uk

Abstract

This paper studies the longitudinal development of a vowel timing alternation known as the "Scottish Vowel Length Rule" in a distinctive variety of Scottish English spoken in Glasgow by working-class men and women. Combining apparent-time and real-time evidence, we show that the implementation of the Rule has changed over time, though unlike in many other varieties of Scottish English, the factors shaping its fate seem to be internal rather than external. Overall, Glaswegian English behaves like a quantity language and controls for prosodic timing effects while preserving the phonological timing alternation; and this is despite a marginal, quasi-phonemic status of the Rule.

Index Terms: SVLR, sociolinguistic real-time corpus, sound change, prosodic timing, Glaswegian

1. Introduction

Glaswegian English, like many other varieties of Scottish English, is well known for its quasi-phonemic patterning of the vowel duration, the so-called 'Scottish Vowel Length Rule' (SVLR, [1]). SVLR-vowels are generally short, and lengthen only before voiced fricatives, /r/ and at morpheme boundaries. Aitken's [1] original formulation applied the Rule to all vowels, but more recently Scobbie et al [2] only found evidence for /i ʉ/ and /ai/ participating in this timing alternation. SVLR stands in contrast to the Postvocalic Voicing Effect (PVE) frequently observed in other varieties of English, e.g. spoken in England and North America, where a vowel is lengthened before voiced consonants but shortened before voiceless ones ([3]). The primary difference between SVLR and PVE concerns the complexity of their constraints: while PVE requires just one constraint, namely the voicing of postvocalic consonants, SVLR additionally relies on the specification of the manner of articulation of the consonants (fricative vs nasal/oral stop) and, if the consonant is a sonorant, its place of articulation (central vs. lateral).

The complexity of the SVLR-constraints is possibly one of the main reasons why the Rule has often been documented to be weakening in situations of high contact with Anglo-English, and giving place to the timing alternations of PVE (e.g. [4, 5]). However, the number of real-time studies addressing this type of sound change is still limited, and there has been little research into potential internal factors influencing this change. Since the timing alternations of SVLR are considered to result in quasi-phonemic vowel quantity in Scottish English ([6]), we might expect SVLR to interact with prosodic timing as in other quantity languages ([7]). In many quantity languages, prosodic timing as well as phonemic vowel quantity place different functional demands on the implementation of vowel duration which might reach ceiling effects due to a combination of accentual, phrase-final and quantity-related lengthening ([8]). Accordingly, durational demarcation of some of the linguistic functions may be compromised. Due to a high functional load of duration for phonology, some quantity languages show only marginal prosodic timing effects (e.g. [7]). However, sound changes towards vowel quantity neutralization in phrase-final positions have also been documented (e.g. [8]).

In this paper, we are wondering about the fate of SVLR in Glasgow where the dialect contact to other varieties of English is traditionally rather limited and where we could expect SVLR to be more resistant to change induced by the external factors ([2]). In a previous investigation ([9]), we addressed this question using a sample of young and middle-aged male speakers recorded in the 1970s and 2000s. The present paper extends the previous results to a larger sample that includes female (as well as male) speakers of the two age groups and decades of recording.

2. Method

2.1. Corpus and speakers

The sample for this paper was drawn from a real-time corpus of Glaswegian vernacular; it contains recordings of spontaneous speech made as early as 1917 as well as more recent ones from 2000s and is stratified by speaker age ([9]).

Our speakers were men (***m***) and women (***f***) in their teens (***Y***-group) and forties (***M***-group) who were recorded for sociolinguistic projects in Glasgow in 1970s (***70***) and 2000s (***00***). We analysed the data of 16 male speakers (4 per group, [9]) and 12 female speakers (3 per group, [10]). 2 out of the 12 females and 5 out of the 16 males had high levels of contact to Anglo-English.

2.2. Data annotation and analysis

All sentences containing words with the SVLR-monophthongs /i ʉ/ in stressed positions were analysed, though words with a postvocalic /r/ were not included. We followed the same labelling routine as in our previous study ([9]) and coded for the SVLR- and the PVE-environments as well as prosodic timing factors (prominence and position within the phrase). The first author annotated the male speaker set ([9]), the second author the female speaker set ([10]).

With the measured vowel duration as the dependent variable, linear mixed effects models were fitted. *Speaker* and *word* were random factors; the predictors were speaker *group*, dialect *contact*, *vowel*, *PVE* and *SVLR* environment, phrasal

position and *prominence* levels; the covariates were lexical *frequency*, number of *syllables* per target word and number of *segments* per target syllable. We tested for all meaningful 3- and 2-way interactions of the main predictors.

3. Results and Discussion

Significant results relevant to the research questions of this study are displayed in Figures 1-3a/b. With regards to the external influence of the dialect contact to Anglo-English (Figure 1), t-tests showed no statistically reliable difference between PVE-long and PVE-short contexts in high-contact speakers and even slightly longer vowels in PVE-short than PVE-long contexts in low-contact speakers ($t=2.0$, $p<0.05$). These findings reinforce the conclusion we discussed in our previous work ([9]) that dialect contact is an unlikely factor to influence the longitudinal development of vowel timing in Glasgow, in contrast to other Scottish English varieties ([4, 5]).

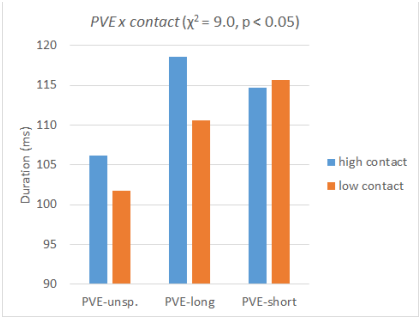


Figure 1: *2-way interaction of dialect contact and PVE.*

As expected, SVLR interacts with prosodic timing in many ways. The short/long distinction reaches a larger magnitude under increased prominence: SVLR-long vowels are substantially longer when accented (20 ms, $t=7.5$, $p<0.001$) whereas SVLR-short vowel show only a small lengthening effect (10 ms, $t=2.3$, $p<0.001$).

Unlike in our previous study ([9]), we do not find evidence for a neutralized short/long SVLR-contrast in phrase-medial, unaccented positions; this might be related to a relatively small number of such vowels, and the lack of a consistency check across male and female datasets.

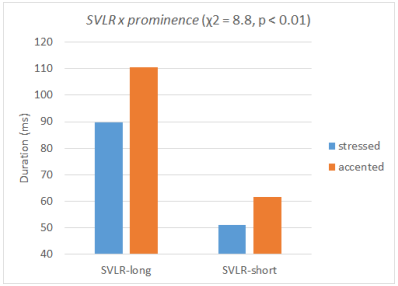


Figure 2: *2-way interaction of prominence and SVLR.*

The results in Figure 3 corroborate our previous finding and show that middle-age male and female speakers born in 1920s have significantly longer SVLR-vowels in phrase-final positions than all other groups (all comparisons $t>2.0$, $p<0.05$). This finding is indicative of an internally induced change ([8]).

Overall, Scottish English spoken in Glasgow behaves like a true quantity language and controls for the amount of prosodically induced lengthening, despite a rather marginal, quasi-phonemic status of SVLR.

4. Acknowledgements

This work was partly funded by the Leverhulme Trust (RPG-142 grant to the last author).

References

[1] Aitken, A.J., (1981) The Scottish Vowel Length Rule. In M. Benskin, M.L. Samuels (eds.), *So meny People, Longages and Tonges: Philological Essays in Scots and Mediaeval English presented to Angus McIntosh*. Edinburgh: The Middle English Dialect Project, pp. 131-157.

[2] Scobbie, J. M, Turk, A., Hewlett, N. (1999) Morphemes, Phonetics and Lexical Items: The Case of the Scottish Vowel Length Rule. *Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco*, pp. 1617-1620.

[3] House, A.S., Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of Acoustical Society of America 25(1)*, pp. 105-113.

[4] Hewlett, N., Matthews, B., and Scobbie, J.M. (1999): Vowel duration in Scottish English speaking children. *Proceedings of the XVth ICPhS, San Francisco*, pp. 2157-60.

[5] Watt D., Ingham C. (2000): Durational evidence of the Scottish Vowel Length Rule in Berwick English. In: Nelson, D. and P. Foulkes (eds) *Leeds Working Papers in Linguistics 8*, pp. 205-228.

[6] Scobbie, J.M., Stuart-Smith, J. (2008) Quasi-phonemic contrast and the fuzzy inventory: examples from Scottish English. In: *Contrast in Phonology: Theory, perception acquisition*. Berlin: Mouton de Gruyter, pp. 87-113.

[7] Nakai S., Turk A., Suomi K., Granlund S., Ylitalo R., Kunnari S. (2012): Quantity constraints on the temporal implementation of phrasal prosody in Northern Finnish. *Journal of Phonetics 40*, pp. 796–807.

[8] Nakai, S. (2013). An explanation for phonological word-final vowel shortening: Evidence from Tokyo Japanese. *Laboratory Phonology 4(2)*, pp. 513 – 553.

[9] Rathcke, T., Stuart-Smith, J.(2016). On the Tail of the Scottish Vowel Length Rule in Glasgow. *Language and Speech 59(3)*, pp. 404-430.

[10] Chevalier, F. (2016). *Temps réel, temps apparent et genre en variation phonétique: l'évolution de la quantité vocalique à Glasgow au cours du XXème siècle* (Unpublished master's thesis). University of Poitiers, France.

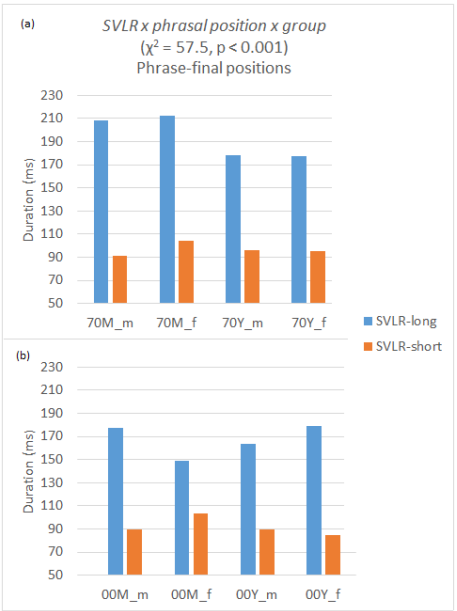


Figure 3: *3-way interaction of SVLR, phrasal position and speaker group.*

Entrainment analysis of categorical intonation representations

Uwe D. Reichel[1], *Jennifer Cole*[2]

[1]Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
[2]Department of Linguistics, Northwestern University, United States of America
uwe.reichel@nytud.mta.hu, jennifer.cole1@northwestern.edu

Abstract

Most studies on prosodic entrainment focus on coarse parametric variables as f0 mean and standard deviation. Only recently first attempts were made to measure entrainment also for categorical intonation representations namely pitch accent types [1]. We propose further metrics for this purpose adopted from text similarity measurement and alignment. These metrics were applied to quantify the similarity of automatically derived intonation contour class sequences in cooperative and competitive dialogs. In line with previously reported results for parametric variables we found also for the categorical representation higher similarities and thus more entrainment in the cooperative dialogs than in the competitive ones. The introduced metrics can be of use for any entrainment research on categorical data as e.g. for ToBI label sequences.

Index Terms: entrainment, intonation stylization, string similarity, local alignment

1. Introduction

In conversation speakers accommodate more and more to each other. This phenomenon is called entrainment and can be observed on various phonetic and linguistic levels. On the linguistic level entrainment affects amongst others the choice of words [2] and syntactic constructions [3, 4]. On the phonetic level entrainment was revealed in dialog data and shadowing experiments for speaking rate [5, 6], intensity [5, 6], voice quality [6], and pitch [7, 8, 5, 9]. Entrainment turned out to be stronger in case of mutual positive attitude of the interlocutors, than in case of negative attitude [10]. Furthermore, entrainment has been shown to increase the success of conversation in terms of low inter-turn latencies and a reduced number of interruptions [6, 2]. Consequently, more entrainment has been reported in cooperative than in competitive dialogs e.g. with respect to intonation contour shapes [11]. These findings are in line with theoretical models such as the Communication Accommodation Theory [12] stating that entrainment enhances social approval and communication efficiency.

For intonation entrainment research is so far mostly restricted to parametric variables, most of them coarse as for example f0 mean and standard deviation over utterance stretches. Only few attempts have so far been made to measure entrainment for a higher-level categorical intonation representation. [1] measured global entrainment over entire dialogs in terms of perplexity and Kullback-Leibler divergence on ToBI [13] pitch accent and boundary tone trigrams. Furthermore, they addressed local entrainment in temporally closely related speech chunks using the Levenshtein distance between tone sequences.

This study aims to contribute to these new entrainment analyses of categorical intonation representations in the following way:

- It will be shown, how such a representation can be generated in a bottom-up way (section 3).
- We will introduce similarity measures for this representation, that capture local entrainment within neighboring speech chunk pairs.
- These measures provide a better account to sub-sequence and crossing alignments of tone sequences than does a Levenshtein distance based approach (section 4).

The employed similarity metrics are: Jaccard index, Cosine index, Szymkiewicz-Simpson coefficient, as well as a similarity measure derived from local alignment.

We applied these metrics to cooperative and competitive dialog data (section 2) to see whether the found similarity values are in line with the findings on parametric data mentioned above. Concretely, we hypothesize to find more entrainment in cooperative than in competitive dialogs expressed by higher values of all proposed similarity metrics.

2. Data

We used parts of the Illinois Game Corpus [14] that contains *Tangram* game dialogs by American English speakers in cooperative and competitive settings. The tangram is a puzzle consisting of seven pieces that can be combined to various shapes. Both dialog partners were separately presented with Tangram silhouettes that were reciprocally hidden from the view of the other partner. The task was to decide whether the silhouettes are the same or different by verbally describing them to each other. In the cooperative setting the partners solved this common goal in a joint effort. In the competitive setting, the partners were required to solve this task competitively, and the one solving it first was declared to be the winner. For more details about the recording setting please consult [15]. For the current study a subset of ten dialogs by five interlocutor pairs was used, of which three were Female-Female pairs and two were Male-Female pairs. Each interlocutor pair took part in a cooperative and a competitive condition, thus our data comprises paired samples of five cooperative and competitive dialogs. Mean dialog duration amounts to 6.5 minutes.

The dialogs were manually text-transcribed and chunk-segmented, and partly manually dialog-act annotated using the tag set of [16]. The data was signal-text aligned by the WEBMAUS webservice [17, 18] and was part of speech tagged using the Balloon toolkit [19]. Both alignment and part of speech labels serve to automatically locate prosodic events, i.e. phrase boundaries and potential pitch accent locations as described in [20].

F0 was extracted by autocorrelation (PRAAT 5.3.16 [21], sample rate 100 Hz). Voiceless utterance parts and F0 outliers were bridged by linear interpolation. The contour was then

smoothed by Savitzky-Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a base value [22]. This base value was set to the F0 median below the 5th percentile of an utterance and serves to normalize F0 with respect to its overall level.

3. Categorical intonation representation

For intonation stylization we adopt the parametric CoPaSul approach of [20], which is illustrated in Figure 1. Within this framework intonation is stylized as a superposition of linear global contours, and third order polynomial local contours. The domain of global contours approximately related to intonation phrases is determined automatically by placing prosodic boundaries at speech pauses and punctuation in the aligned transcript. The domain of local contours is determined by placing boundaries behind each content word so that the resulting segments generally contain at most one pitch accent.

The global linear component is given by the F0 baseline fitted through f0 values at the bottom of the time varying f0 range as explained in [23]. The baseline is then subtracted from the F0 contour, and a third order polynomial is fitted to the F0 residual within each local segment. Time is normalized to the range from -1 to 1 so that time 0 is placed in the mid of the content word's syllable bearing the lexical stress.

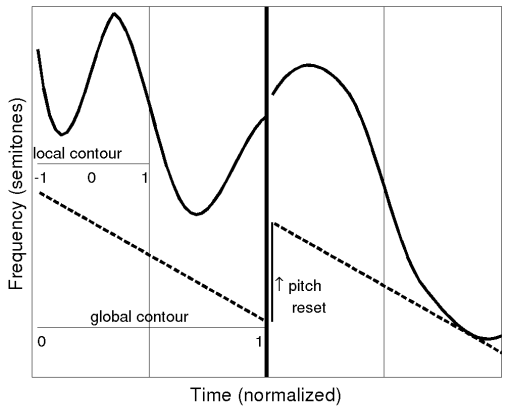


Figure 1: *CoPaSul: Contour-based parametrical superpositional F0 stylization.*

In order to derive a categorical representation from this parametric stylization, the slopes of the global contours as well as the polynomial coefficients of the local contours are clustered by Kmeans. Following [20] the optimal number of contour classes was initialized by subtractive clustering [24]. The resulting three global and four local contour classes are shown in Figure 2.

4. Entrainment measurements

As one can see in Figure 3, the contour class distributions, unigrams as well as bigrams, are highly determined by the dialog act of the speech chunk. This is reflected by significantly higher information radii (two-sided Welch tests, $p < 0.001$) of these distributions when comparing them between different dialog act chunks as opposed to same dialog act chunks. These findings are in line with [25] who discuss dialog-related differences in intonation parameters in the context of Ohala's Frequency Code

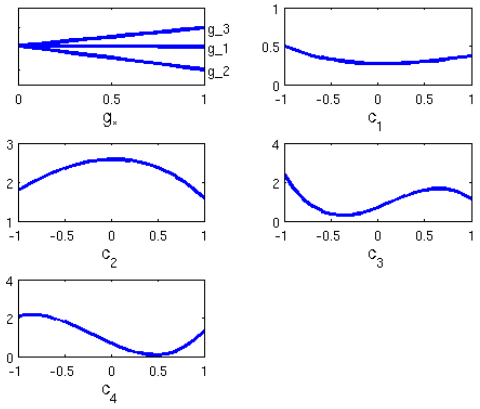


Figure 2: *Global (g_i) and local (c_j) contour classes resulting from polynomial coefficient clustering.*

framework [26]. In order to disentangle entrainment and dialog act dependencies, we applied the similarity measures only on speech chunks of the same dialog act.

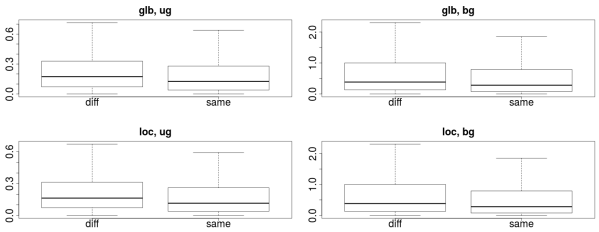


Figure 3: *Information radii of contour class unigram and bigram probability models within and across dialog act types.*

4.1. Similarity of contour class inventories

The similarity of the contour class inventories X and Y of speech chunk pair was quantified by three standard string-based similarity metrics [27]: the Cosine similarity, the Jaccard index [28] and the overlap ratio (Szymkiewicz-Simpson coefficient [29]), which are defined as follows:

$$\text{Cosine } C(X,Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}},$$

$$\text{Jaccard } J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|},$$

$$\text{Overlap } O(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)}.$$

$|S|$ refers to the cardinality of a set S, i.e. in our case the number of different contour class types. All indices range from 0 (no similarity) to 1 (total similarity).

4.2. Similarity of contour class sequences

We adopted the idea of [1] to measure similarity of contour class sequences by means of alignment. Since sequences usually differ in length, and since these length differences add up to the

overall distance, it is advisable to normalize the distance with respect to length. [1] propose the following transformation of the Levenshtein distance $d(x, y)$ between the sequences x and y to a similarity score $s_r(x, y)$ ranging between 0 and 1 partly normalized with respect to length:

$$s_r(x, y) = \frac{m - d(x, y)}{m},$$

where $m = \max[\text{length}(x), \text{length}(y)]$, i.e. the length of the longer sequence and thus the upper limit of the number of edit operations. Note that x and y here do not refer to sets as the capital letters in the previous section, but to contour class sequences. As one can see in Figure 4, this similarity measure has two shortcomings: first, it does punish sequences of different length even if one sequence is entirely contained within the other. Thus two possible domains of entrainment, utterance duration and intonation, are merged to a single metrics. Second, it does punish sequences with cross matching subsequences. Thus, it cannot account for cases where interlocutors choose the same intonation contours but at different positions within their utterances. To disentangle duration and intonation and to capture cross matches we propose an alternative measure based on local alignment:

$$s_l(x, y) = \frac{\text{length(localigned(z))}}{\text{length}(z)},$$

where $z = \arg\min_{z \in \{x,y\}}[\text{length}(z)]$. The similarity $s_l(x, y)$ of an intonation class sequence pair is thus the proportion of the locally aligned parts of the shorter sequence in that pair. As s_r also s_l similarity scores range from 0 to 1, Figure 4 gives an example. Since all members of the shorter sequence x are (with cross matches) contained in the longer sequence y, $s_l(x, y)$ amounts 1. In contrast, the Levenshtein distance between x and y amounts 6 which yields a similarity $s_r(x, y) = \frac{7-6}{7} = 0.14$, and thus a quite different result, that underestimates the fact, that x is entirely contained in y.

x	**e**	**f**	**g**	**a**	**b**		
y	**a**	**b**	c	d	**e**	**f**	**g**

Figure 4: Alignment of two sequences x and y of differing length. x is with cross correspondences entirely contained in y. Levenshtein distance: 6; Levenshtein-derived similarity $s_r(x, y) = 0.14$; local alignment derived similarity $s_l(x, y) = 1$.

The proposed local alignment is implemented by an adaptation of the dynamic programming Smith-Waterman algorithm [30]. The alignment score matrix H spanned by the sequences x and y with length m and n, respectively (cf. left half of Figure 5) is filled as follows:

$$H[i, 0] = 0, 0 \leq i \leq m$$
$$H[0, j] = 0, 0 \leq j \leq n$$
$$H[i, j] = \max \left\{ \begin{array}{rcl} 0 & : & \text{Lower bound} \\ H[i-1, j-1] + s(x_i, y_j) & : & \text{Match/Mismatch} \\ \max_{k>0}[H(i-k, j) + W_k] & : & \text{Deletion} \\ \max_{l>0}[H(i, j-l) + W_l] & : & \text{Insertion} \end{array} \right\},$$
$$1 \leq i \leq m, 1 \leq j \leq n$$

$s(a, b)$ is a similarity function and W_i a gap scoring scheme [31]. Both allow for a high flexibility in the alignment process. For our purpose we restrict it to align only matching subsequences. Thus everything but zero-substitutions should result in a cell value below or equal 0 so that this operation will not contribute to the alignment. This is realized by setting W_i as well as $s(a, b)$ for $a \neq b$ constant to $-l$, where l is the length of any of the sequences to be aligned. Only zero-substitutions ($a == b$) are rewarded by $s(a, b) = 1$.

All matching subsequences are then retrieved from this matrix by the following iteration:

while $\max(H) > t$

– trace back from the cell containing this maximum the path leading to it until a zero-cell is reached
– add the subsequence collected on this way to the set of aligned sequences
– set all traversed cells to 0

This iteration is illustrated in Figure 5. The threshold t defines the required minimum length of aligned subsequences. It is set to 2 in this study. $t = 1$ would result in a complete alignment of any pair of permutations of x. The traversed cells need to be set to 0 after each iteration step to prevent that one subsequence would be related to more than one alignment pair.

This approach allows for two more restrictions: to prevent cross alignment not just the traversed cells $[i, j]$ but for each of these cells its entire row i and column j needs to be set to 0. Second, if only the longest common substring is of interest, then the iteration is trivially to be stopped after the first step.

	-	a	b	c	d	**e**	**f**	**g**
-	0	0	0	0	0	0	0	0
e	0	0	0	0	0	**1**	0	0
f	0	0	0	0	0	0	**2**	0
g	0	0	0	0	0	0	0	**3**
a	0	1	0	0	0	0	0	0
b	0	0	2	0	0	0	0	0

	-	**a**	**b**	c	d	**e**	**f**	**g**
-	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0
a	0	**1**	0	0	0	0	0	0
b	0	0	**2**	0	0	0	0	0

Figure 5: Iterative longest common subsequence (LCS) detection in local alignment. While the matrix maximum is above a threshold, start at this maximum and trace back until a 0 cell is reached and set all traversed cells to 0. This yields in the first iteration step (**left**) the alignment of **e f g**, and in the second step (**right**) the alignment of **a b**.

5. Results

In line with mentioned findings of previous studies and with our hypothesis all similarity measures yield higher values in the cooperative than in the competitive dialogs (two-sided Welch tests, $p < 0.001$). This is shown in Figure 6.

6. Discussion

We introduced several similarity metrics from natural language processing to measure entrainment in categorical intonation data. The results indicate higher entrainment for both intonation inventory as well as tone sequencing which is well in line with finding on the parametric level. This we take as an indication that the proposed metrics are of value in prosodic entrainment research. We argue that local alignment based similarity is better suited for entrainment measurements than the transformed standard Levenshtein distance since it cancels out sequence length differences and can cope with cross correspondences. It is highly flexible due to several tuning parameters

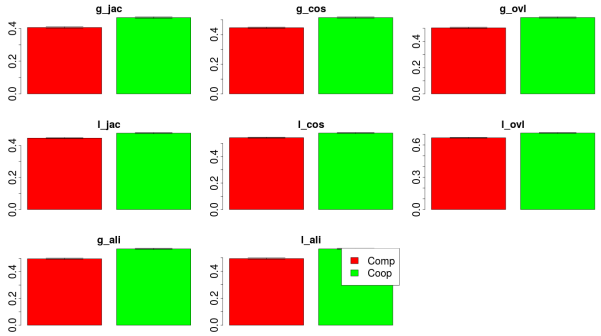


Figure 6: *Similarities of global (g_*) and local (l_*) contour class inventories in competitive (COMP, red) and cooperative (COOP, green) dialogs. jac – Jaccard index, cos – cosine similarity, ovl – overlap ratio, ali – local alignment.*

given by the similarity function, the gap penalty scoring, the score thresholding, and the procedure how to trace back the alignment score matrix, so that it can be customized to the respective research needs.

In this study the categorical intonation representation was derived in a bottom-up way. Nevertheless, the measures can be applied to any categorical data including expert-driven intonation representations as ToBI annotations.

7. Acknowledgments

The work of the first author is financed by a grant of the Alexander von Humboldt society. The data collection was funded by the Volkswagen Stiftung.

8. References

[1] A. Gravano, v. Beňuš, R. Levitan, and J. Hirschberg, "Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement," in *Proc. Interspeech*, Dresden, 2015, pp. 578–582.

[2] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, 2008, pp. 169–172.

[3] A. Cleland and M. Pickering, "The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure," *Journal of Memory and Language*, vol. 49, pp. 214–230, 2003.

[4] S. Gries, "Syntactic priming: A corpus-based approach," *Journal of Psycholinguistic Research*, 2005.

[5] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 3081–3084.

[6] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *NAACL HLT '12 Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, 2012, pp. 11–19.

[7] S. Gregory and S. Webster, "A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions," *J. Pers. Soc. Psychol.*, vol. 70, pp. 1231–1240, 1996.

[8] S. Gregory, K. Dagan, and S. Webster, "Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality," *J. Nonverbal Behavior*, vol. 21, pp. 23–43, 1997.

[9] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Language and Speech*, vol. 55, pp. 231–248, 2011.

[10] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 793–796.

[11] J. Cole and U. Reichel, "Prosodic entrainment – the cognitive encoding of prosody and its relation to discourse function," Keynote at Framing speech satellite workshop of the Speech Prosody conference, Boston, 2016.

[12] H. Giles and N. Coupland, *Language: Contexts and Consequences*. Pacific Grove, CA: Brooks/Cole, 1991.

[13] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, Cambridge, Massachusets, 1980.

[14] PAGE, "Prosodic and Gestural Entrainment in Conversational Interaction across Diverse Languages," http://page.home.amu.edu.pl/.

[15] U. Reichel, N. Pörner, D. Nowack, and J. Cole, "Analysis and classification of cooperative and competitive dialogs," in *Proc. Interspeech*, Dresden, Germany, 2015, p. paper 3056.

[16] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.

[17] F. Schiel, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proc. ICPhS*, San Francisco, 1999, pp. 607–610.

[18] T. Kisler, U. Reichel, F. Schiel, C. Draxler, B. Jack l, and N. Pörner, "BAS Speech Science Web Services - an update of current deve lopments," in *Proc. LREC 2016*, Portoro, Slovenia, 2016, pp. 3880–3885.

[19] U. Reichel, "PermA and Balloon: Tools for string alignment and text processing," in *Proc. Interspeech*, Portland, Oregon, USA, 2012, p. paper no. 346.

[20] ——, "Linking bottom-up intonation stylization to discourse structure," *Computer, Speech, and Language*, vol. 28, pp. 1340–1365, 2014, dOI: 10.1016/j.csl.2014.03.005.

[21] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Tech. Rep., 1999, 132–182.

[22] A. Savitzky and M. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[23] U. Reichel and K. Mády, "Comparing parameterizations of pitch register and its disconti nuities at prosodic boundaries for Hungarian," in *Proc. Interspeech 2014*, Singapore, 2014, pp. 111–115.

[24] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation," *Journal of Intelligence & Fuzzy Systems*, vol. 2, no. 3, pp. 267–278, 1994.

[25] K. Mittelhammer and U. Reichel, "Characterization and prediction of dialogue acts using prosodic features," in *Elektronische Sprachverarbeitung 2016*, ser. Studientexte zur Sprachkommunikation, O. Jokisch, Ed. Dresden, Germany: TUDpress, 2016, vol. 81, pp. 160–167.

[26] J. Ohala, "The frequency code underlies the sound symbolic use of voice pitch," in *Sound Symbolism*. Cambridge: Cambridge University Press, 1994.

[27] W. Gomaa and A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[28] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines," *Bulletin de la Socit Vaudoise des Sciences Naturelles*, vol. 37, pp. 241–272, 1901.

[29] D. Szymkjewicz, "Une contribution statistique a la géographie floristique," *Acta Soc. Bot. Polon.*, vol. 34, no. 3, pp. 249–265, 1934.

[30] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[31] M. Vingron and M. Waterman, "Sequence alignment and penalty choice. Review of concepts, case studies and implications," *Journal of Molecular Biology*, vol. 235, no. 1, pp. 1–12, 1994.

Prosodic Marking of Information Status in Task-Oriented Dialogues

Christine T. Röhr, Tabea Thies, Stefan Baumann, Martine Grice

IfL Phonetik, University of Cologne, Germany

{christine.roehr; tabea.thies; stefan.baumann; martine.grice}@uni-koeln.de

Abstract

In the present paper we investigate the effect of information status on accent placement and accent types used in semi-spontaneous speech. As an elicitation method we use a 'spot-the-difference' task which provides natural (dialogue) but still controlled (task-oriented) speech data. The task has been shown to be an ideal testbed for the relation between prosody and discourse meaning. However, it has not been used in a fine-grained study of information status yet. This is done in the present study by applying the *RefLex* annotation scheme, which differentiates between a referential and a lexical level of givenness. The semi-spontaneous speech data indicate a systematic but probabilistic relation between prosodic prominence and an item's level of givenness. That is, the correlation between increasing newness and increasing prominence is predominantly reflected in a more frequent use of nuclear pitch accents as well as a less frequent use of deaccentuation. Both the referential and lexical levels of givenness turn out to have an incremental effect on the degree of an item's prosodic prominence. Consequently, the *combined* degree of givenness of a referent (reflected by a combination of RefLex labels) indicates an overall prominence value of the item's prosodic realization.

Index Terms: production, semi-spontaneous speech, information status/givenness, RefLex, prosody, accent placement, pitch accent

1. Introduction

It is well known that in intonation languages like German the marking of information status (given-new dimension) is an important linguistic function of prosody.

In particular, Chafe [1] and Prince [2] have shown that it is not sufficient to distinguish only given and new information but to take at least a third intermediate class of givenness into account, sometimes described as accessible or inferable information. In the last three decades, various labelling schemes have been designed to enable annotations of more fine-grained differences in an item's information status (e.g. different types of accessible information). However, following Baumann & Riester [3] none of these schemes have proven detailed enough to capture and distinguish all sorts of informational distinctions which are necessary to explain even the most elementary intonational patterns. They argue that for an adequate analysis of an item's information status in spoken language two levels of givenness have to be investigated: a referential and a lexical level. Accordingly they developed a new, two-layered, type of annotation system for information status of referring (and non-referring) expressions (called *RefLex*), which, moreover, does not only distinguish given and new but also intermediate classes of givenness/novelty. Referential information status is assigned at the level of DP and PP, whereas lexical information status applies at the word level or modified NP level. In Table 2 an overview of the scheme – divided into a referential and a lexical level – is presented. The overview is a simplification (cf. [4]) of a more comprehensive account and describes only those labels which play a role in the present study. For the entire scheme, consult [3]. Detailed annotation guidelines are about to be published.

With regard to prosody, differences in an item's level of givenness have been shown to be marked by nuclear pitch accent placement (e.g. [5], [6], [7]) and/or pitch accent type (e.g. [8], [9], [10], [11], [12], [13], [14]): The less given an item, the higher the prosodic prominence produced.

Baumann & Riester [4] investigated the impact of the information status categories at a referential and a lexical level (as proposed in their RefLex scheme) on the prosodic realization. For read speech they generally confirmed the relationship between information status and prosody showing a stepwise increase in prosodic prominence from given to new items, predominantly ordered according to the information status at the lexical level. However, the results have been found to be less clear in spontaneous speech.

In order to find out which combinations of information status levels serve as triggers for which intonational categories further speech corpora have to be built/used and annotated according to the RefLex annotation scheme. The aim of the present study is to provide semi-spontaneous speech data by using a rather natural test setting with two interlocutors (task-oriented dialogues). We thus set up a 'spot-the-difference' task (resembling the 'diapix' task by [15]) which involves two similar but not identical pictures, and two participants who cannot see each other's picture. The two subjects have the task of working together to find ten differences between the pictures, involving either missing or replaced items. This task has the advantage to elicit balanced speech from each participant, i.e. there are no predetermined 'Giver' and 'Receiver' roles, as in typical Map Task dialogues. It also facilitates the elicitation of repeated mentions of segmentally controlled expressions in a variety of syntactic positions and utterance types.

For the present study we generally hypothesize that the 'newer' (or less given) an item is (both at the referential and the lexical level) the more it is made prosodically prominent by a speaker. More precisely, we assume relative differences in the intonational marking of information status, indicated by different distributions and/or probabilities of prosodic categories. The categories we are looking at are a) accent placement and b) accent type for nuclear and prenuclear pitch accents (categorized according to GToBI [16]) - assuming an increase in prominence from left to right:

a) no accent < (postnuclear accent <) prenuclear accent < nuclear accent (cf. [17])

b) no accent < low accent (L*) < falling accent (H+L*, H+!H*), high accent (!H*, H*) < rising accent (L*+H, L+H*) (cf. also [18])

The RefLex categories are claimed to express an increase in the level of an item's newness from left to right:

a) referential level: r-given < r-bridging < r-unused < r-new

b) lexical level: l-given < l-accessible < l-new

2. Method

2.1. Test Material

We designed two pictures for a spot-the-difference task (see Fig.1). Both pictures show a picnic-setting: A girl and a boy are arranged on a meadow nearby a tree and a wooden hut.

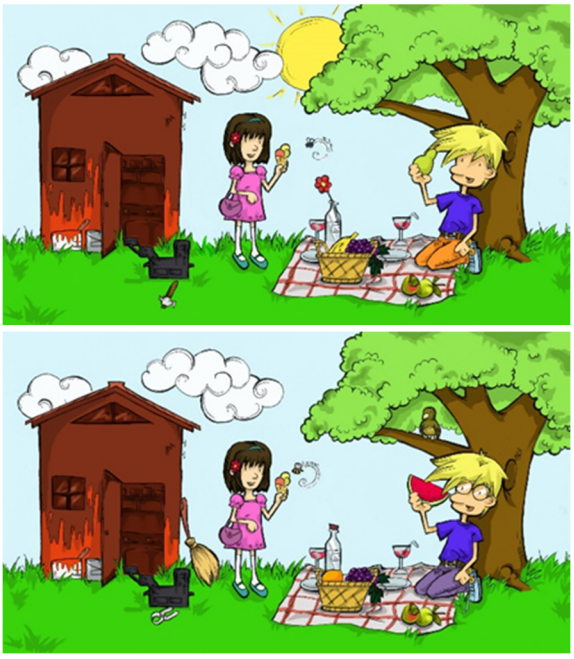


Figure 1: Pictures used for spot-the-difference task

Missing items	Replaced items		
Sonne sun	*Birne* pear	vs.	*Melone* melon
Blume flower	*Fliege* fly	vs.	*Biene* bee
Besen broom	*Hammer* hammer	vs.	*Zange* tongs
Brille glasses	*Banane* banana	vs.	*Orange* orange
Vogel bird	*orange* orange	vs.	*lila (Hose)* purple (trousers)

Table 1: Differences between the pictures of 'spot-the difference' task

Altogether the two pictures contain fifteen different target items/words that involve in total ten differences: On the one hand, there are five items that are only present in one of the two pictures (e.g. top picture: sun and flower) which are absent in the other picture. On the other hand, there are items in five particular positions that are in the other picture replaced by other items (e.g. the item in the boy's hand, top picture: pear vs. bottom picture: melon). A list of the ten differences or rather 'missing' and 'replaced' target items is given in Table 1.

The items on the pictures needed to be familiar and easily identifiable. Therefore, the target items were chosen from the *LEMO* database [19] which contains a set of 260 pictures that are standardized with regard to name agreement, image agreement, familiarity, and visual complexity.

2.2. Experimental setup

The experiment took place at the I*f*l Phonetik of the University Cologne and was composed of three parts: a priming phase, a practice section and the main experiment.

In the priming phase we familiarized the subjects with the target items that are shown in the pictures of the main experiment. This was necessary in order to guarantee an easy recognition of the images of the items in question and a uniform naming by all subjects. For the priming we used in total 40 images from the LEMO database (including the 15 target items of the experiment). The priming was conducted separately for each subject. We presented the priming elements successively on a computer screen and the subject's task was to read out loud the name of a depicted item. Subjects were instructed to only use those denotations in the main experiment for the particular items.

In the practice section we familiarized the subjects with the task. They were seated opposite to each other in a sound attenuated room. However, we placed a partition wall between the subjects so that they could not see each other. Both subjects received the same pair of two similar pictures involving five differences. To get an idea of the task of the main experiment they needed to discuss the five differences between the two pictures.

In contrast to the practice section in the following main experiment each subject only received one of the two pictures (see Fig.1). Since the subjects were not able to see each other's picture their task was to work together in order to discover the ten differences between their pictures. The main experiment was over as soon as the subjects identified all differences. The conversations between the subjects during the main part of the experiment were recorded acoustically using a headset condensator microphone for each subject.

2.3. Subjects

We recorded 12 native speakers of Standard German (six female and six male) in six dialogues, arranged in two female pairs, two male pairs and two mixed-gender pairs. They grew up in North Rhine-Westphalia or Rhineland-Palatinate and were aged between 19 and 29 years (mean = 22.8, SD = 2.7).

2.4. Analysis

In a first step, we produced transcripts of the dialogues. In a second step, we annotated the information status of nominal and also adjectival expressions in the conversations according to the RefLex annotation scheme (see Table 2). That is, at a referential level labels were applied to DPs, PPs and APs. The information status of the words within these phrases were separately labelled at a lexical level. Information that

expresses a contrast was additionally marked with a '(c)' attached to the lexical label.

In a third step we segmented and annotated the acoustic data in *Praat* [20]. At four segmental levels we a) annotated every spoken word, b) determined their part of speech (except for verbal expressions), c) classified the type of phrase that has been labelled at the referential level and d) marked the primary stressed syllable of all nouns, adjectives and verbs. The part of speech labels refer to the guidelines of the *Stuttgart-Tübingen-TagSet* [21]. At the phrasal level we distinguished noun phrases, prepositional phrases, adjective phrases, adverbial phrases and pronominal phrases.

Furthermore, we analyzed the prosodic realization of all sentences at two different levels. On a level of accent placement we marked for every word whether it was realized with no accent (coded as 0), with a postnuclear accent (1), a prenuclear accent (2) or a with a nuclear accent (in subordinate clauses coded as 3 and in main clauses coded as 4). On a tonal level we marked the positions of realized pitch accents and boundary tones and categorized their tonal configuration according to GToBI. In a last step the RefLex annotations were transferred to Praat.

In this paper we will present a descriptive analysis of the prosodic marking of the annotated RefLex categories. The results presented here are based on pooled GToBI and RefLex categories, even though we used the more fine-grained categories during the annotation process. That is, we basically distinguish between low (L*), falling (H+L*, H+!H*), high (!H*, H*) and rising (L*+H, L+H*) pitch accent types. RefLex categories were pooled according to the simplified overview given in Table 2.

Referential level (indicated by 'r-')	
r-given	coreferring anaphor that is present (immediately or displaced) in previous discourse context or contained in the text-external context (-sit)
r-bridging	non-coreferring anaphor dependent on previously introduced scenario
r-unused	globally unique discourse-new (non-anaphoric) entity which is generally known (-known) or identifiable from its own linguistic description (-unknown)
r-new	indefinite non-unique discourse-new entity
Lexical level (indicated by 'l-')	
l-given	markable is a repetition (-same), synonym (-syn), hypernym (-super) or holonym (-whole) of a previous expression
l-accessible	markable has an identical word stem (-stem) or is a hyponym (-sub) or meronym (-part) of a previous expression
l-new	markable is not related to another expression within the last five intonation phrases or clauses

Table 2: Simplified overview of annotation tags of the RefLex annotation scheme (cf. [3], [4])

3. Results

As an overall result, we found that the examined RefLex categories had an effect on the prosodic marking.

The distribution of accent *placement* (on all RefLex-annotated words) both as a function of the referential level and the lexical level of givenness shows that a word is more likely to get accented the less given it is. Figure 2 indicates that a decrease in referential and lexical givenness is reflected by a clear increase in the use of nuclear (and tendentially also prenuclear) accents. Accordingly, the number of words that are not accented progressively decreases. The data reveal similar results for a separate analysis of nouns and adjectives. However, the effect of information status on nouns applies more clearly to nuclear accents, while for adjectives it primarily shows differences in the distribution of prenuclear accents (due to structural reasons).

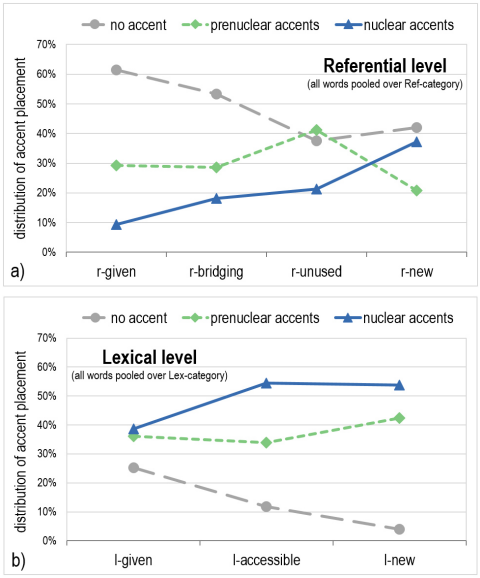


Figure 2: Relative distribution of (prenuclear and nuclear) accents on all RefLex-annotated words ordered according to their assumed level of givenness on a) a referential and b) a lexical level

In order to investigate the interaction between referential and lexical categories of givenness with regard to the prosodic marking we further analyzed the effect of combined RefLex categories.

Results for accent *placement* (see Fig.3) show that both levels have an impact on the resulting degree of prosodic prominence. Thus, if an item is both referentially and lexically *new*, there is a high probability that the item is marked by a nuclear accent (which is highly prominent in general). However, if an item is referentially *given* but lexically *new*, it is most consistently marked by a prenuclear accent, which may be considered secondary in its degree of prominence (see [17]). That is, the resulting prominence value of the prosodic realization seems to reflect the combined degree of givenness of an item, represented by the two levels of information status.

For the distribution of accent *types* (and their inherent level of prominence, see [18]) it is the *lexical* level that turns out to be decisive (see Fig.4). This is reflected by a reverse distribution of rising and high accents, i.e. rising accents (assumed to be most prominent) become more frequent, whereas high accents become less frequent with increasing newness on the lexical level.

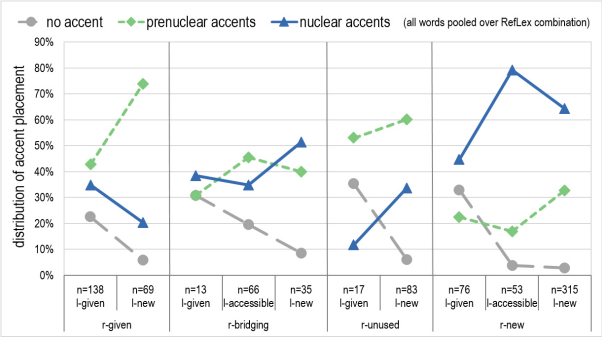


Figure 3: Relative distribution of (prenuclear and nuclear) accents on all RefLex-annotated words ordered according to their level of givenness (RefLex combination)

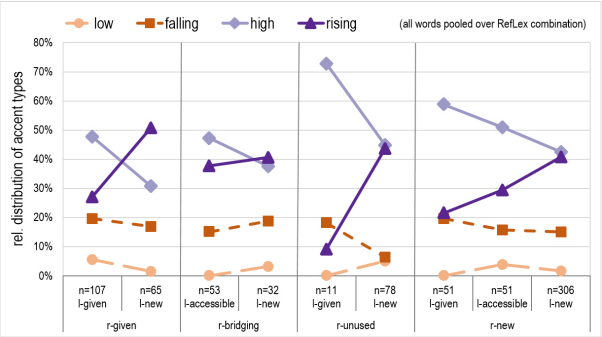


Figure 4: Relative distribution of accent types (pre-nuclear and nuclear accents pooled) on RefLex-annotated words ordered according to their level of givenness (RefLex combination)

4. Conclusions

Although we investigate (semi-)spontaneous speech, a style that has often been claimed to be less clear-cut than read speech (see e.g. [4]), our study provides clear evidence for the relation between levels of givenness and their prosodic realization.

The data suggest that accent placement is a more decisive prosodic marker of information status than accent type. Thus, the hypotheses on the distribution of accents in terms of their placement could be confirmed, showing a general tendency of a stepwise increase in prosodic prominence from given to new expressions *at both levels (referential and lexical)*. As a consequence, the *combined* degree of givenness of an item (reflected by combined RefLex labels) results in an overall prominence value of the prosodic realization.

Furthermore, the RefLex scheme has been shown to be a promising tool for the investigation of the relation between an item's information status and its prosodic realization. Our study has confirmed the relevance of both a referential and a lexical level of givenness.

Finally, the spot-the-difference task has turned out to be a useful source of task-oriented, (semi-)spontaneous speech for the examination of information status and prosody.

5. References

[1] W. Chafe, "Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View," in *Subject and Topic*, C. Li, Ed. New York: Academic Press, 1976, pp. 25-56.

[2] E. F. Prince, "Toward a Taxonomy of Given-New Information," in *Radical Pragmatics*, P. Cole, Ed. New York: Academic Press, 1981, pp. 223-256.

[3] S. Baumann and A. Riester, "Referential and lexical givenness: Semantic, prosodic and cognitive aspects," in *Prosody and Meaning*, G. Elordieta and P. Prieto, Eds. Berlin & New York: Mouton De Gruyter, 2012, pp. 119-162.

[4] S. Baumann and A. Riester, "Coreference, Lexical Givenness and Prosody in German," in *Lingua* [Special Issue "Information Structure Triggers"], vol. 136, J. Hartmann, J. Radó, S. Winkler, Eds. 2013, pp.16-37.

[5] D. J. Allerton, "The Notion of 'Givenness' and its Relation to Presupposition and Theme," *Lingua*, vol. 44, pp.133-168, 1978.

[6] J. Terken and J. Hirschberg, "Deaccentuation of Words Representing 'Given' Information: Effects of Persistence of Grammatical Role and Surface Position," *Language and Speech*, vol. 37, pp. 125-145, 1994.

[7] C. Féry and F. Kügler, "Pitch accent scaling on given, new and focused constituents in German," *Journal of Phonetics*, vol. 36, no. 4, 680-703, 2008.

[8] J. B. Pierrehumbert and J. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan and M. E. Pollack, Eds. Cambridge: MIT Press, 1990, pp. 271-311.

[9] K. Kohler, "Terminal Intonation Patterns in Single-Accent Utterances of German: Phonetics, Phonology and Semantics," *AIPUK*, vol. 25, pp. 115-185, 1991.

[10] S. Baumann, M. Grice, "The Intonation of Accessibility," *Journal of Pragmatics*, vol. 38, no. 10, pp. 1636-1657, 2006.

[11] A. Chen, E. den Os, E. and J. P. de Ruiter, "Pitch accent type matters for online processing of information status: Evidence from natural and synthetic speech," *The Linguistic Review*, vol. 24, pp. 317-344, 2007.

[12] P. Schumacher and S. Baumann, "Pitch accent type affects the N400 during referential processing," *NeuroReport*, vol. 21, no. 9, pp. 618-62, 2010.

[13] C. T. Röhr and S. Baumann, "Prosodic marking of information status in German," *Proceedings of the 5th International Conference on Speech Prosody*, vol. 100019, pp. 1-4, 2010.

[14] C. T. Röhr and S. Baumann, "Decoding information status by type and position of accent in German." *Proceedings of the ICPhS XVII*, 2011, pp. 1706-1709.

[15] A. R. Bradlow, R. E. Baker, A. Choi, M. Kim and K. J. van Engen. "The Wildcat Corpus of Native- and Foreign-Accented Englisch," *Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 2-3072, 2007.

[16] M. Grice, S. Baumann, and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology," in *Prosodic Typology. The Phonology of Intonation and Phrasing*, S. Jun, Ed. Oxford: OUP, 2005, pp. 55-83.

[17] S. Baumann, "The Importance of Tonal Cues for Untrained Listeners in Judging Prominence," *Proceedings of the 10th ISSP*, pp. 21-24, 2014.

[18] S. Baumann and C. T. Röhr, "The perceptual prominence of pitch accent types in German," *Proceedings of the ICPhS XVIII*, vol. 298, pp. 1-5, 2015.

[19] J. G. Snodgrass and M. Vanderwart. "A standardised set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6. 174-215, 1980.

[20] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341-345, 2001.

[21] A. Schiller, S. Teufel, C. Stöckert and C. Thielen, "Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS", 1995.

Anticipatory V-to-V Coarticulation in German Preschoolers

Elina Rubertus [1], *Dzhuma Abakarova* [1], *Jan Ries* [1], *Aude Noiray* [1]

[1] Laboratory for Oral Language Acquisition, University of Potsdam
rubertus@uni-potsdam.de

Abstract

This study investigates lingual V-to-V anticipatory coarticulation in German preschoolers and adults using ultrasound measures. In light of conflicting results in the literature, the aim was to study effects in larger cohorts and with a widespread set of vowels. Results provide evidence for V-to-V coarticulation in children as well as adults, independent of the intervocalic consonant. Interestingly, coarticulation degree decreases with age.

Index Terms: Language acquisition, coarticulation, ultrasound, speech production

1. Introduction

Coarticulation, generally defined as the articulatory overlap of speech sounds with one another, provides an opportunity to bridge the gap between phonology and phonetics, as abstract phonemes are assembled to a continuous speech stream. It seems that various mechanisms guide this process: Recasens [1] summarizes that the nature of coarticulatory processes and their magnitude are influenced by mechanical constraints on the one hand and articulatory preprogramming mechanisms on the other hand. More precisely, he found that the extent to which a vowel may overlap with a preceding consonant in CV-syllables highly depends on the articulatory constraints associated with the consonant, i.e. its resistance [2]. However, in anticipatory V-to-V coarticulation in VCV sequences, resistance of C did not show an accordingly large influence ([1], [3]). Recasens interprets this as evidence for V-to-V anticipatory coarticulation to mainly result from articulatory preprogramming.

To become fluent speakers, young children have to both develop a refined control of their speech production system and learn to plan their articulation to achieve their native language's coarticulation patterns. However, albeit studied quite frequently, coarticulation in child speech remains poorly understood because of contradictory results in previous studies (e.g., [4] versus [5]). Due to the lack of non-invasive articulatory measures, child speech has been mostly examined acoustically (except for [6], [7], [8]).

The present study is part of a larger project that aims to track the developmental course of coarticulation mechanisms in German children, investigating multiple age groups and combining traditional acoustic measurements with direct measures of articulation via ultrasound imaging and labial shape tracking.

Here, we more specifically focus on articulatory investigations of lingual anticipatory V-to-V coarticulation in German preschoolers. Some studies on child speech reported a systematic change of the first vowel depending on the second vowel ([9]; [4], [10] for 9;5-year-old child; [11]). Others did not find such effect ([10] for 4;8-year-old child; [12]). Except for [12] all studies only included acoustic measurements. For adults, there is strong evidence for anticipatory V-to-V coarticulation and for this effect to be at least partially modulated by the intervocalic C's resistance ([1], [3], [13], [14]). However, as Recasens [1] emphasizes, the impact of the consonant's resistance is a lot smaller in anticipatory V-to-V coarticulation than it is in CV- and even in carry-over V-to-V coarticulation.

In light of previous literature, our study addresses the following questions: First, do we observe anticipatory V-to-V coarticulation in children as well as adults? If we find that the tongue position during the first vowel varies as a function of tongue position during the second vowel, it will bring evidence for anticipatory V-to-V coarticulation. Second, is the magnitude of V-to-V coarticulation modulated by the degree of resistance of the intervening consonant? If so, we expect smaller V-to-V coarticulation in cases for which consonantal resistance is stronger (i.e. alveolars) than when resistance is minimal (i.e. labials). And finally, does the coarticulatory pattern and magnitude change in the course of development? This hypothesis will be tested by looking at possible differences across cohorts.

2. Method

2.1. Participants

In this study, two cohorts of children including 18 3-year old children (10 females, age range: 3;05 – 3;08 (Y;MM), mean: 3;06), 13 5-year old children (7 females, age range: 5;04 – 5;07, mean: 5;06) and 16 adults (8 females, age range: 19-34 years mean: 25;08) were tested. Participants grew up in a monolingual German environment and none of them reported any language-, hearing, or visual problems.

2.2. Stimulus material

C_1VC_2ə pseudowords were embedded in carrier phrases with the German female article /aɪnə/ such as "eine bide". The set of consonants used consisted of /b/, /d/, and /g/, the vowel set of the tense and long vowels /i/, /y/, /u/, /a/, /e/, and /o/. C_1Vs were designed as a fully crossed set of Cs and Vs while the second C_2ə syllable was added in a way that C_1 was never equal to C_2. Anticipatory V-to-V coarticulation was measured between the vowel and the preceding schwa. Children repeated every word 3 times, resulting in 108 trials per child. Those 108 trials were presented in 6 randomized blocks. For adults the additional consonant /z/ was included but is not analyzed here. With 3 repetitions of each word their data set included 218 trials presented in 9 randomized blocks.

2.3. Experimental procedure

Participants were recorded with SOLLAR (Sonographic and Optical Linguo-Labial Articulation Recording system [15]). This child-friendly platform allows for simultaneous recordings of tongue movement (Sonosite Edge, sr.: 48Hz), lip movement

(video camera SONY HDR-CX740VE, sr.: 50Hz) and audio speech signal (microphone Shure, sampling rate: 48kHz). The relatively small ultrasound probe was positioned straight below participants' chin to record the tongue on the midsagittal plane. It was fixed on a custom-made probe holder to be flexible in the vertical dimension allowing for natural jaw movement but prevent motion in lateral and horizontal translations. The acoustic recordings served as a reference to detect the relevant time points in the ultrasound video.

Stimuli were recorded by a German female model speaker beforehand. In the experiment, the task for participants was to repeat the auditorily presented stimuli. For children, the repetition task was presented as a game to stimulate their interest.

2.4. Data processing

First, acoustic data were phonetically labeled using Praat [16]. The time points relevant to our analysis are the temporal midpoint of schwa and the temporal midpoint of the vowel.

These time points were subsequently used to find the corresponding frames in the ultrasound video signal. For each relevant time point, tongue contours were semi-automatically detected with scripts custom-made for MATLAB [17] as part of the SOLLAR platform (see **Fehler! Verweisquelle konnte nicht gefunden werden.**). For each relevant contour, the x-coordinate of the highest point of the tongue dorsum was automatically extracted and used for subsequent coarticulation analyses.

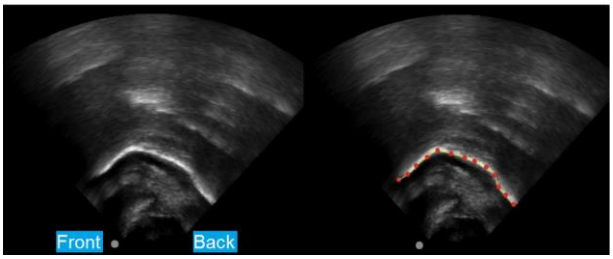


Figure 1. *Midsaggital view of the tongue surface. Left: Tongue contour without labels. Right side: Tongue contour labeled in red.*

3. Results

In line with previous acoustic studies, we first investigated coarticulation using Locus Equations (LE). Most notably, Sussman and colleagues (e.g., [18]) computed linear regressions for the second formant (F2) between the vowel onset and its midpoint to test for linear relationships between consonant and vowels in CV sequences. They found that the slopes of the regressions varied with the amount of CV coarticulation. We transposed LE to the articulatory domain and used the horizontal position of the highest point of the tongue instead of F2. Instead of examining the degree of coarticulation between the consonant and the vowel, we report on the relationship between the schwa and the vowel.

Figure 2-5 display the resulting regression lines. While the slopes for the different consonant contexts are roughly the same within each cohort, slopes are highest for 3-year-old children, intermediate for 5-year-olds and lowest for adults, suggesting less V-to-V coarticulation in older cohorts.

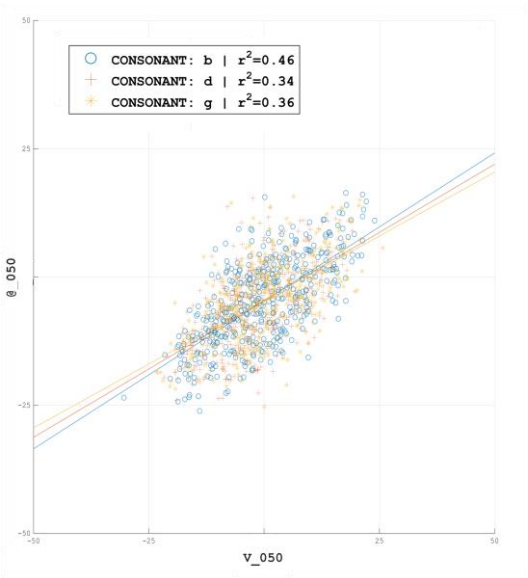


Figure 2. Linear regressions between schwa midpoint and vowel midpoint for 3-year-olds. Slopes are b: 0.58, d: 0.53, g: 0.5.

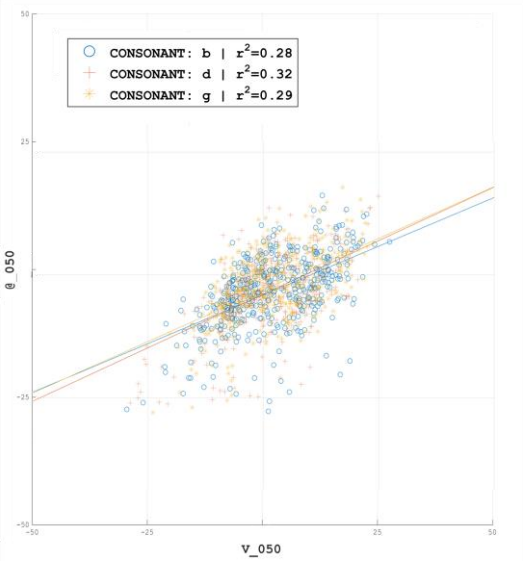


Figure 3. *Linear regressions between schwa midpoint and vowel midpoint for 5-year-olds. Slopes are b: 0.4, d: 0.44, g: 0.42.*

For a more precise analysis we also fit a linear mixed effects model to investigate the relationship between the horizontal position of the highest point of the tongue dorsum during schwa (dependent variable) and during the vowel (independent variable), using R [19] and lme4 [20]. As fixed effects, the horizontal position of the highest point of the tongue at the temporal midpoint of the vowel, cohort, and consonant were included with interaction terms. As random effects, we included intercepts for participants and words, as well as by-word random slopes for the effect of cohort. Residual plots were visually inspected and did not show deviations from homoscedasticity or normality. The goodness of fit was determined using likelihood ratio tests and p-values were obtained with lmerTest [21]. The linear model's output for the main effects and the interactions is dis-

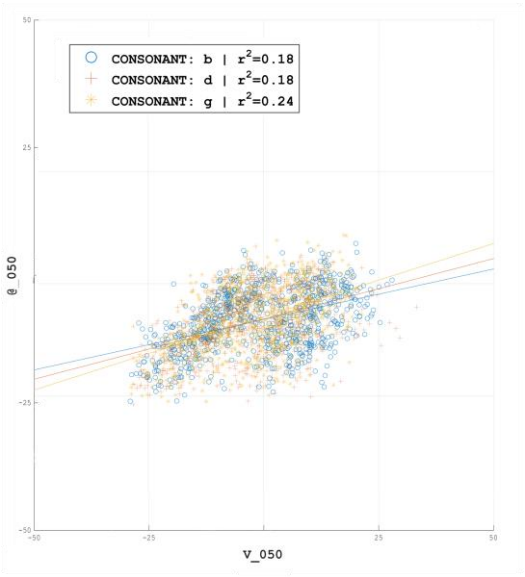


Figure 4. *Linear regressions between schwa midpoint and vowel midpoint for adults. Slopes are b: 0.23, d: 0.27, g: 0.33.*

played in **Fehler! Verweisquelle konnte nicht gefunden werden.**. The cohort of 5-year olds and the labial consonant served as the base lines for cohort and consonant respectively.

The position of the tongue (represented by the x-coordinate of the highest point of the tongue dorsum) during the vowel significantly affects its respective position during the schwa. There is not any significant difference in tongue position during schwa between the 3- and the 5-year-olds, but between the 5-year-olds and adults. Neither the consonant /d/ nor /g/ differ significantly from /b/ in their effect on schwa. More interestingly though, the effect of the vowel on the schwa seems to be modulated by age as shown by the significant interactions between the vowel and the cohorts. The effect of the vowel on schwa is not significantly affected by the nature of the intervening consonant as shown by non-significant interactions between vowel and consonants. The interaction between the 3-year old cohort and the consonant /g/ is only marginally significant suggesting that the effect of /g/ on the schwa is different between 3- and 5-year-olds. None of the three-way interactions reached significance.

Table 1. *Output of the linear mixed model.*

Effect	Estimate	SE	t	p
Intercept	-4.714	1.30	-3.61	***
Vowel midpoint	0.56	0.02	20.13	***
3-year-olds	0.27	1.39	0.2	
Adults	-3.35	1.52	-2.21	*
Consonant /d/	1.17	1.16	1.01	
Consonant /g/	0.41	1.12	0.37	
Vowel midpoint: 3-year-olds	0.09	0.03	2.69	**
Vowel midpoint: Adults	-0.43	0.03	-15.01	***
Vowel midpoint: Consonant /d/	0.01	0.04	0.31	
Vowel midpoint: Consonant /g/	0.01	0.04	0.28	
3-year-olds: Consonant /d/	-0.54	0.53	-1.015	
Adults: Consonant /d/	-1.28	0.93	-1.37	
3-year-olds: Consonant /g/	-1.01	0.53	-1.89	.
Adults: Consonant /g/	0.018	0.9	0.02	
Vowel midpoint: 3-year-olds: Consonant /d/	-0.04	0.05	-0.92	
Vowel midpoint: Adults: Consonant /d/	-0.01	0.04	-0.26	
Vowel midpoint: 3-year-olds: Consonant /g/	-0.07	0.05	-1.42	
Vowel midpoint: Adults: Consonant /g/	0.05	0.04	1.28	

Sign. codes: ***: p < 0.001, **: p < 0.01 *: p < 0.05, .: p < 0.1

4. Discussion

The significant main effect of tongue position at vowel midpoint in the linear mixed effects analysis suggests an overall effect of the vowel on the schwa, hence, the presence of V-to-V coarticulation. Interestingly, the non-significant interactions between the vowel midpoint and the different consonants negate a dependency of the V-to-V coarticulation magnitude on the identity of the intervening consonant. In our stimuli, the consonants /b/, /d/, and /g/ were used because they vary in coarticulatory resistance with /b/ being least resistant, /g/ intermediate and /d/ most resistant. Results of our earlier CV-coarticulation analysis are neatly in line with this hierarchy. In light of previous literature ([1], [3], [13], [14]), it might then be expected that the effect of V2 on V1 is lower in cases of an intervening /d/, intermediate for an intervening /g/ and highest for an intervening /b/. However, our data suggest that the resistance of the consonant does not significantly affect V-to-V coarticulation. This result might be interpreted as an even stronger case of Recasens' observations [1] that the consonant's effect on coarticulation amount is limited in anticipatory V-to-V coarticulation. In a next step it would be interesting to investigate carry-over coarticulation from V to the final schwa in our data to see whether the consonants' resistance has a larger impact here. Recasens' [1] interpretation of this result, that V-to-V coarticulation is

mainly a question of articulation preprogramming as opposed to CV-coarticulation which is much more a matter of mechanical constraints, suggests that children's task to master V-to-V coarticulation is essentially learning to plan their articulation.

From a developmental perspective, our current results suggest that there are actually changes in the degree but not in the pattern of coarticulation with age. Planning is thus probably not adult-like yet. The significant interactions between age cohorts and vowel midpoint depict that 3-year-olds' V-to-V coarticulation magnitude differs from 5-year-olds' and that 5-year-olds' in turn differs significantly from adults'. Going back to the regression analysis, the pattern of slopes across age cohorts shows a developmental trend towards less coarticulation with increasing age. Slopes are highest for the 3-year-olds, intermediate for the 5-year-olds and lowest for adults. While our current results are very clear about this age effect, previous investigations displayed different pictures: Repp [10] for example found no V-to-V coarticulation in his younger (4;8 years) but in his older participant (9;5 years) suggesting an increase of coarticulation. Barbier [12] did not find anticipatory V-to-V coarticulation in 4-year-olds. Nittrouer [4] did find V-to-V coarticulation in 3-7-year-olds but no age effect and Boucher [9] actually found an effect of age comparable to our result. The only previous study investigating lingual anticipatory V-to-V coarticulation in children with direct articulatory measurements as we did is Barbier [12]. However, their stimulus material differs substantially from ours in that they included two full vowels instead of a schwa and a vowel. Schwa is generally more malleable than other vowels (e.g., [22]), which might explain why there are effects in our study but not in [12]. A strong claim such as Nittrouer et al.'s [4, p.387] "children's gestures are organized into separate syllabic units, as adult gestures are." is nevertheless challenged by the present findings which suggest that there is still ongoing change towards adult-like (syllabic) patterns in 5-year-olds.

While there are substantial changes in the magnitude of V-to-V coarticulation, the coarticulatory patterns, i.e. the influence of the intervening consonant on V-to-V coarticulation, does not change with age as proposed by the non-significant three-way-interactions between vowel midpoint, cohorts, and consonants.

5. Conclusion

To answer the three initial questions, we studied lingual anticipatory V-to-V coarticulation in German 3- and 5-year-olds as well as adults measuring the highest point of the tongue. Our results converge towards the conclusion that both children and adults exhibit V-to-V coarticulation independent of the nature of the intervocalic consonant. Further, the degree of V-to-V coarticulation decreases across life.

We are currently developing more refined measures to explore whether an influence of the intervocalic consonant may be manifest in the global tongue contour or curvature degree.

6. Acknowledgements

This work was supported by the DFG grant number: 1098, a great team of enthusiastic researchers at LOLA and a lot of friendly and patient families.

7. References

[1] D. Recasens, "An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences." (1987). *Journal of Phonetics*, vol. 15, pp. 299-312, 1987.

[2] D. Recasens, "V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study." *Journal of Phonetics*, vol. 12, pp. 61-73, 1984.

[3] D. Recasens. "Vowel-to-vowel coarticulation in Catalan VCV sequences." *Journal of the Acoustical Society of America*, vol. 76, pp. 1624-1635, 1984.

[4] S. Nittrouer, M. Studdert-Kennedy and S. T. Neely. "How children learn to organize their speech gestures: Further evidence from fricative-vowel syllables." *Journal of Speech and Hearing Research*, vol. 39, pp. 379-389, 1996.

[5] J. R. Green, C. A. Moore, and K. J. Reilly. "The sequential development of jaw and lip control for speech." *Journal of Speech, Language, and Hearing Research*, vol. 45(1), pp. 66-79, 2002.

[6] L. Ménard and A. Noiray. "The development of lingual gestures in speech: Comparing synthesized vocal tracts with natural vowels." *Fait de Langue*, vol. 37, pp. 189-202, 2011.

[7] A. Noiray, L. Ménard, and K. Iskarous. "The development of motor synergies in children: Ultrasound and acoustic measurements." *The Journal of the Acoustical Society of America*, vol. 133(1), pp. 444-452, 2013.

[8] A. Smith. "Speech motor development: Integrating muscles, movements, and linguistic units." *Journal of communication disorders*, vol. 39(5), pp. 331-349, 2006.

[9] K. M. Boucher, "Patterns of anticipatory coarticulation in adults and typically developing children", M.S. thesis, Brigham Young University, Provo, UT, 2007.

[10] B. Repp. "Some observations on the development of anticipatory coarticulation." *Journal of the Acoustical Society of America*, vol. 79(5), pp. 1616-1619, 1986.

[11] A. Noiray et al. "Emergence of a vocalic gesture control: Attunement of the anticipatory rounding temporal pattern in French children." E*mergence of language Abilities*, S. Kern, F. Gayraud and E. Marsico, Eds. Newcastle upon Tyne: Cambridge Scholars Publishing, 2008, pp. 100-116.

[12] G. Barbier et al. "Speech planning as an index of speech motor control maturity." *14th Annual Conference of the International Speech Communication Association*, 2013.

[13] C. A. Fowler and L. Brancazio. "Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation." *Language and Speech*, vol. 43(1), pp. 1-41, 2000.

[14] S. E. G. Öhman. "Coarticulation in VCV utterances: Spectrographic measurements." *Journal of the Acoustical Society of America*, vol. 39, pp. 151-168, 1966.

[15] A. Noiray, J. Ries, M. Tiede, M. (2015). "Sonographic & Optical Linguo-Labial Articulation Recording system (SOLLAR)," presented at Ultrafest VII, Hong Kong, 2015.

[16] Boersma, P. & Weenink, D. "Praat: doing phonetics by computer" [Computer program]. Version 6.0.20, retrieved 3 September 2016 from http://www.praat.org/.

[17] MATLAB Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States.

[18] H. M. Sussman, K. A. Hoemeke, H. A. McCaffrey. "Locus equations as an index of coarticulation for place of articulation distinctions in children." *Journal of Speech and Hearing Research*, vol. 34, pp. 769-781, 1992.

[19] R Core Team. "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2016.

[20] D. Bates et al. "Fitting linear mixed-effects models using lme4." *Journal of Statistical Software*, vol. 67(1), pp. 1-48, 2015

[21] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. "lmerTest: Tests in linear mixed effects models." R package version 2.0-32. https://CRAN.R-project.org/package=lmerTest, 2016.

[22] C. Browman and L. Goldstein, "Targetless schwa: An articulatory analysis" Haskins Laboratories, New Haven, CT, Status Report on Speech Research, SR-101/102, 1990.

An acoustic analysis of German initial laterals in the L2 speech of Bosnian migrants living in Vienna

Carolin Schmid

Acoustics Research Institute, Austrian Academy of Sciences, Vienna

carolin.schmid@oeaw.ac.at

Abstract

In this paper, the phonetic realization of L2 German initial laterals produced by L1 Bosnian speakers living in Vienna is investigated and compared to L1 Standard Austrian German (SAG) and L1 Viennese Dialect (VD) realizations, both in read speech and in spontaneous speech. This pilot study is embedded in the larger context of my PhD dissertation on the sociophonetic aspects of language contact in the speech of Bosnian migrants living in Vienna, which will be concerned with language acquisition as well as language attrition. The results show that Bosnian speakers realize a more velarized lateral in initial word position than both SAG and VD speakers, both in read and in spontaneous speech and in all vowel contexts. Thereby, the velarization is stronger in read speech than in spontaneous speech.

Index Terms: laterals, L2 speech, sociophonetics

1. Introduction

The present preliminary study within the framework of my PhD thesis deals with the acquisition of a new phonetic category in L2. More precisely, this study examines the acquisition of a phoneme category, when there is only *one* category in the L2, whereas there are *two* categories in the L1. Speakers of L1 Bosnian in Vienna are confronted to two German varieties, namely Standard Austrian German (SAG) and the Viennese Dialect (VD). Bosnian features two lateral phonemes, a palatal and a velarized lateral phoneme, the latter having an alveolar allophone [1]. The two phonemes are also distinguished orthographically: <l> stands for the velarized lateral, <lj> for the palatalized lateral. Laterals in the German contact varieties spoken in Vienna are structured in other phonetic categories than in L1 Bosnian: The category of the palatalized lateral is absent in both L2 varieties. The second category is mapped in a different way: whereas it is a broader category in L1 Bosnian, including mainly velarized laterals, but depending on the phoneme context also containing a plain alveolar variant, there is only one lateral phoneme in a smaller category produced in SAG: the alveolar lateral approximant. The VD shares the lateral phoneme inventory with SAG, but contrary to SAG, this phoneme category is broader, because it includes the velarized lateral, which is also present in Bosnian, as a variant. In the VD, the lateral phoneme is primarily produced as an alveolar lateral approximant and its velarized variant occurs mostly at the word-initial and –final position and between back vowels. Indeed, velarization is negatively connoted by Austrian listeners, as a salient feature of the negatively evaluated VD, and therefore it is avoided especially by female speakers [2, 3], as speakers are usually aware of this negative connotation. Thus, this phoneme category is basically shared by Bosnian and the VD, but the use of the same variants within this category is different.

Even if L1 Bosnian speakers were aware of the negative evaluation of the VD, they may not understand its stereotyped function as a dialect marker. Additionally, due to the relative phonetic vicinity of the laterals in Bosnian and German (especially if Bosnian speakers are used to hear the velarized lateral in the VD), they may just have transferred the lateral from their Bosnian to their German pronunciation (see the findings about similar sounds in L1 and L2 of [4]). In the present study I will investigate whether cross-linguistic influence [5] will lead to transfer phenomena (as described in [6,7]) from L1 Bosnian to L2 German.

I hypothesize that by and large, Bosnian speakers will realize laterals in average more velarized than speakers of SAG and also more velarized than VD speakers, who don't produce a velarized lateral in all instances and often generally avoid it because of its negative evaluation. Furthermore, I hypothesize that women perform better than men concerning pronunciation in L2. To my knowledge, this hasn't been shown so far (as stated in [8]), but [9] states that "female L2 learner may be more concerned about pronunciation accuracy than their male counterparts." Regarding standard vs. dialect pronunciation, women are described to use a variety closer to the standard norms than men do (see [10]), and these findings could also be valid for L2 pronunciation and have to be verified. One additional assumption for the present study concerns orthography. In read speech the laterals are likely to be realized in a more velarized way, because the German letter <l> for the lateral is the letter used for the velarized lateral in the Bosnian orthographic system. As described in [11], orthography can have a hindering effect on an accurate, targetlike L2 pronunciation. Thus I suppose that the grapheme <l> will trigger the activation of the Bosnian velarized lateral category and also supress the realization of a palatalized lateral phoneme.

To answer these questions, acoustic analyses of spontaneous speech and read speech of L1 Bosnian speakers as well as of L1 SAG and VD speakers is analyzed and compared. In this preliminary study I focus on laterals in initial word position in order to find out what cues most noticeably the pronunciation of German laterals by Bosnian L1 speakers. Sounds at the word initial position are prosodically more pronounced than in other positions and are thus more marked in the perception of listeners.

To shortly summarize the articulation of laterals and to explain the acoustic measurements applied in this study: Laterals are articulated with a closure at the mid-sagittal line of the vocal tract, built with the tip or blade of the tongue against the alveolar ridge, in the case of alveolar laterals, or dentally, in

the case of velarized laterals. The production of palatalized laterals implies a larger closure in the area of the palatum. Thereby, in all laterals, the airstream passes laterally at the sides of the tongue, causing antifromants. Responsible for the darker quality of the velarized lateral is a longer cavity behind the closure in the vocal tract, which is caused by a secondary articulation in terms of a velar/pharyngal constriction and leads to lower F2 and higher F1. According to [12], laterals are perceived as alveolar laterals, when their F2 is above 1500 Hz and as velarized laterals when their F2 is below 1200 Hz. F3 is associated to the front cavity of the vocal tract (the cavity before the first closure) and is especially important for the realization of palatalized and velarized laterals, where it should be relatively high, around 3000 Hz, together with an high F2 just below the F3 for palatalized laterals.

Table 1: examples for read sentences (i= front vowel, a= back vowel, l=lateral).

Vowel context	Sentence
(1) i + l + i	Die Liebe ist… *(Love is…)*
(2) a + l + a	Da lachen ja alle. *(That makes everybody laugh.)*
(3) a + l + i	Wo Liebe ist, da… (*Where love is, there…)*
(4) i + l + a	Die Locken… *(The curls…)*

For the acoustic investigation of the laterals in the present study, and the primarily interesting analyze of the degree of velarization of the laterals, especially F2 is of interest. That is, because it displays a large variation between alveolar and velarized laterals and can be measured more precisely than F1 (overlapping/confusion with F0) and F3 (often in the area of antiformants).

2. Method

2.1. Speakers

German read and spontaneous speech was recorded from three female and two male Bosnian speakers, as well as from three female and male SAG speakers and two female and three male speakers of the VD, who served as control groups.[1] The Bosnian speakers were born and raised in the region of today's Bosnia and emigrated to Vienna during the war in Bosnia, when they were between 20 and 35 years old. The SAG and the VD speakers have the same age, thus at least 40 years, and were born and raised in Vienna. They differ in their educational background, namely in that the SAG speakers and their parents have a higher school education (academic education) as the VD speakers (see [13] for the selection criteria of SAG- and VD-speakers).

[1] Unfortunately, one male Bosnian speaker and one female VD speaker had to be excluded from the analysis because of insufficient recording quality.

2.2. Recordings

2.2.1. Read speech

Read speech was recorded in order to control the lateral context and to have comparable data for all speakers. 20 sentences, containing a word with a lateral in initial stressed position, were read twice. As shown in Table 1, the laterals occurred in balanced vowel context: (1) between front vowels, (2) between back vowels, (3) after a front, before a back vowel, and (4) after a back, before a front vowel. The sentences were partially written as cloze text, in order to avoid monotonous read speech.

2.2.2. Spontaneous speech

Spontaneous speech was recorded by means of a semi-structured biographical interview, including questions about social networks and language use and –attitudes. The interviewer of the L1 Bosnian speakers was a German L1 speaker, the SAG speakers were interviewd by an SAG speaker and the VD speakers by a dialect speaker. The recordings of spontaneous speech allowed us to analyze the pronunciation of laterals in a less controlled, but a more natural setting, in order to verify or falsify the results obtained from read speech. Additionally, the information of the biographical interviews will be used in further studies for qualitative analyses of the pronunciation performance of the Bosnian speakers.

2.2.3. Parameter extraction

The transcription of the recordings as well as the segmentation of the laterals, the words in which the laterals occurred and the neighboring vowels were performed manually in STx [14]. In doing so, the lateral segment border were verified through intensity variation, formant transitions and changes in the waveform. Formant transitions were not excluded from the lateral segment. In a next step, formant frequencies (F1-F3) of the laterals and the vowels were measured using LPC (window length 46 ms, 95% overlap), verified and, if necessaire, manually corrected. Additionally, duration of the words, the laterals and the vowels and spectral intensity of the laterals and the vowels were extracted.

Table 2: mean F1 and F2 values for each L1 group and splitted according to speaker gender.

Speaker group		Read speech		Spontaneous speech	
		F1	F2	F1	F2
Bosnian speakers	female	351	1401	402	1497
	male	300	1361	369	1343
SAG speakers	female	255	1737	319	1753
	male	279	1552	319	1466
VD speakers	female	283	1747	346	1756
	male	237	1549	292	1441

2.3. Statistical analyses

Descriptive statistical analyses (Kruskal-Wallis-tests) were performed in R [15], in order to get an overview of the different lateral realizations in the respective L1 groups. In the recordings of the L1 Bosnian speakers a total amount of 200 laterals in read speech and 320 laterals in spontaneous speech could be analyzed, 238 laterals in read speech as well as in spontaneous speech were included from the recordings of the SAG speakers, and the VD speakers are represented with 198 laterals in read and 157 laterals in spontaneous speech.

3. Results

Table 2: mean F1 and F2 values for each L1 group and splitted according to speaker gender.

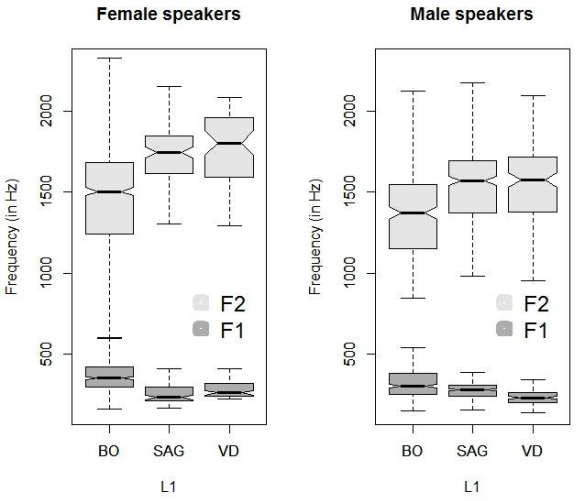


Figure 1: F1 and F2 in read speech.

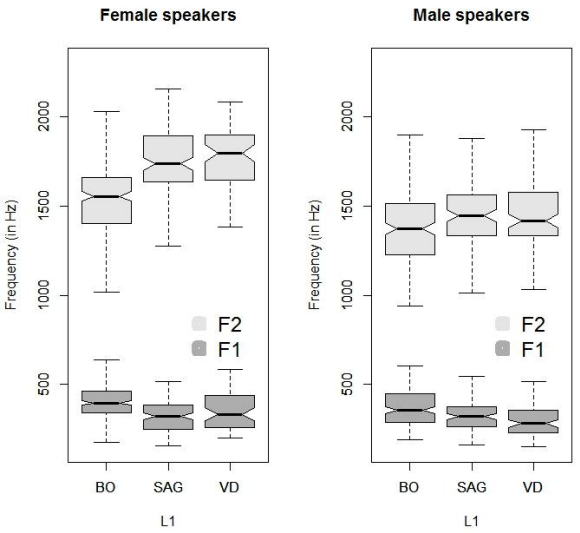


Figure 2: F1 and F2 in spontaneous speech.

First results of all measured formant values revealed no differences in F3 of the laterals for the different speaker groups, so that this formant was excluded from the analyses of the present study and only F1 and F2 were statistically analyzed. A close look at Table 2 (as well as Figure 1 and 2) reveals that Bosnian speakers have higher F1 and lower F2 values in both spontaneous and read speech than SAG and VD speakers. Whereas F1 is not significant, F2 differs between the L1 groups (women and men, respectively) in read speech ($p<0.001$), and also with $p<0.001$ in the realizations of women in spontaneous speech. Male Bosnian speakers also differ from SAG speakers with $p<0.001$, and from VD speakers with $p<0.5$ in spontaneous speech. Comparing the two speaking styles, a difference between the three L1-groups can be noticed: Bosnian speakers realize higher F2 values in spontaneous speech than in read speech ($p<0.01$ for female speakers, and in the pronunciation of male speakers a tendency can be observed which is, however, not significant), whereas SAG und VD speakers show higher F2 values in read speech ($p=0.002$ for SAG male speakers, $p=0.003$ for VD male speakers).

If we distinguish between back and front vowel context in this study, we observe that Bosnian male speakers behave similar to VD male speakers in the back vowel context. Most interestingly, both differ from SAG speakers by having lower F2 values. In spontaneous speech, Bosnian speakers as well as VD speakers show significantly lower F2 values of their laterals as SAG speakers ($p<0.05$). In the realization of read speech laterals in back vowel context, Bosnian speakers differ significantly from both L1 German groups ($p<0.001$). In the front vowel context, the laterals of the VD speakers are higher and similar to those of the male SAG speakers, and even though the F2 values of Bosnian speakers tend to be lower, there is no significant effect between the L1 groups, neither in read nor in spontaneous speech. Female speakers of L1 Bosnian on the other hand always produce laterals with a lower F2 than their L1 German counterparts (which, in spontaneous speech, is significant only in the front vowel context with $p<0.05$, and in read speech in both vowel contexts, with $p<0.01$).

Besides the results for the L1 Bosnian speakers, the computations also reveal interesting results for SAG and VD speakers. The mean formant values shown in Table 2 indicate that those two groups are equally alike in their production of laterals. However, a closer look at the investigated lateral conditions reveals a difference between SAG and VD speakers in back and front vowel context. Overall, VD speakers tend to differentiate those two contexts to a greater extent than SAG speakers, the mean F2 values are more distant. In the back vowel context, the laterals of male VD speakers have lower F2 values than SAG speakers (in spontaneous speech this difference is significant with $p<0.01$, whereas in read speech there is only a tendency for F2 to be lower in VD). In the same way, female VD speakers tend to produce lower F2 values in laterals in read speech, but not in spontaneous speech.

Additionnally to formant frequencies, duration and intensity of the lateral segments were measured. No significant effects emerged concerning these parameters, even though L1 Bosnian speakers tend to produce longer laterals in German utterances. This longer duration may be due to L2 speech, but will be investigated with additional data in more detail in further studies, for example in different word positions or concerning the realization of the two Bosnian lateral phonemes in their L1 Bosnian. The intensity measures will also be refined and measured seperately for different frequency bands.

4. Discussion

The Bosnian speakers of L2 German analyzed in this preliminary study differ in their lateral pronunciation especially in terms of F1 and F2 from L1 speakers of SAG and VD. F1 in laterals is globally realized with a higher frequency and F2 with a lower frequency, both by female and male speakers. This indicates a stronger velarization of the German laterals by Bosnian speakers. It has to be verified now, whether this is a "simple" transfer phenomenon of L1 Bosnian laterals, or whether it is a merging between the two language varieties, Bosnian and German. If it is merging, the lateral realizations in L2 German should be less velarized than the lateral realizations in L1 Bosnian, or both L1 and L2 laterals should be less velarized than the laterals pronounced by a control group of Bosnian L1 speakers still living in Bosnia and not speaking German. The influence of the VD will be analyzed by means of the interview, by taking a closer look at the questions about German language awareness and about social networks [16]. But since the results for Bosnian shown in the present investigation differ from those for VD speakers, even a strong and positive identification with VD speakers cannot acount for the still stronger velarization of Bosnian speakers.

Concerning gender, there is no evidence for a gender specific pronunciation of L2 laterals in this study. The analysis of more date will show whether Bosnian male and female speakers really behave equally in terms of L2 pronunciation.

Yet an interesting finding of this preliminary study is the difference between the lateral production in read and spontaneous speech of Bosnian speakers. In spontaneous speech, laterals were indeed more velarized than laterals of SAG and VD speakers, but less velarized than in read speech. The influence of orthography could at least partially explain this result. However, the present analysis couldn't show evidence for the existence of more palatalized laterals in spontaneous speech recordings (and thereby not orthographically influenced realization) of Bosnian speakers compared to their read spech recordings. However, this L2 differences between the recording tasks could also be explained by an interviewer effect, resulting from the Standard speaking German interviewer. This interview situation might have caused an accomodation process [17] which can only be evaluated in further studies, by investigating the lateral realizations of Bosnian speakers over time or by performing additional interviews with new interviewers (for example one with an L1 VD speaker and one with an L1 Bosnian speaker).

5. References

[1] D. Petrovic, and S. Grubisic, *Fonologija srpskoga jezika.* Beograd: SANU, 2010.

[2] S. Moosmüller, C. Schmid and C. Kasess, "Alveolar and Velarized Laterals in Albanian and in the Viennese Dialect," *Language and Speech*, vol. 58, no. 4, pp. 1-28, 2015.

[3] S. Moosmüller, C. Schmid and J. Brandstätter, "Standard Austrian German," *Journal of the International Phonetic Association*, vol. 45, no. 3, pp. 339-348.

[4] J. E. Flege, "The production of <new> and <similar> phones in a foreign language. Evidence for the effect of equivalence classification," *Journal of Phonetics*, vol. 15, pp. 47–65, 1987.

[5] M. Sharwood-Smith, "Crosslinguistic aspects of second language acquisition," *Applied Linguistics*, no. 4, pp. 192-199.

[6] R. Bundgaard-Nielsen, "Vocabulary sizes matters: The assimilation of second language Australian English vowels to first language Japanese vowel categories," *Applied Psycholinguistics*, vol. 32, no. 1, pp. 51-67, 2011.

[7] K. Aoyama, J. E. Flege, S. G. Guion, R. Akahane-Yamada and T. Yamada. "Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/," *Journal of Phonetics*, vol. 32, pp. 233–250, 2004.

[8] A. Moyer, "The social nature of L2 pronunciation," In J. Levis and A. Moyer, *Social Dynamics in Second Language Accent.* Bosten: De Gruyter Mouton, pp. 11-30, 2014.

[9] A. Moyer, "The puzzle of gender effects in L2 phonology," *Journal of Second Language Pronunciation*, vol. 2, no. 1, pp. 8-28, 2016.

[10] W. Labov, *Principles of linguistic change, Volume 2: Social factors.* Oxford: Blackwell, 2001.

[11] B. Bassetti, "Effects of hanyu pinyin on pronunciation in learners of Chinese as a foreign language," In A. Guder, X. Jihang and Y. Wan, *The cognition, learning and teaching of Chinese characters*. Beijing, China: Beijing Language and Culture University, pp. 156-179, 2007.

[12] D. Recasens, "A cross-language acoustic study of initial and final allophones of /l/," *Speech Communication*, vol. 54, pp. 368–383, 2012.

[13] S. Moosmüller, *Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck*. Köln, Weimar: Böhlau, 1991.

[14] P. Balazs, A. Noll, W. Deutsch and B. Laback, "Concept of the integrated signal analysis software system STx". *Jahrestagung der Österreichischen Physikalischen Gesellschaft 2000*, 2000.

[15] #R gui www.r-project.org

[16] L. Milroy, *Language and social networks*. Oxford: Blackwell, 1980.

[17] B. Soukup, "Zum Phänomen <Speaker Design> im österreichischen Deutsch," In A. N. Lenz and M. Glauninger, *Standarddeutsch im 21. Jahrhundert – Theoretische und empirische Ansätze mit einem Fokus auf Österreich*. Göttingen: V&R, pp. 59-79, 2015.

Wie Deutschschweizer Lernende die stimmhaften Obstruenten des Italienischen aussprechen

Stephan Schmid

Phonetisches Laboratorium, Universität Zürich (Schweiz)
stephan.schmid@uzh.ch

Abstract

Im Gegensatz zum Italienischen gibt es in Deutschschweizer Dialekten keine stimmhaften Obstruenten, sondern (stimmlose) *lenes*, die sich von den (ebenfalls stimmlosen) *fortes* hauptsächlich in der Dauer unterscheiden. Aufgrund der gängigen Modelle der L2-Phonologie ist anzunehmen, dass Deutschschweizer Lernende stimmhafte Obstruenten des Italienischen an die *lenes* ihrer L1 'assimilieren'; eine kontrastive Analyse legt zudem nahe, dass dies bei /dz/, /dʒ/ und s+C-Verbindungen besonders häufig der Fall sein könnte.

Diese Hypothesen wurden im Wesentlichen durch eine akustische Analyse von 644 italienische Obstruenten bestätigt, die von 10 Zürcher Mittelschülern realisiert wurden. Dabei ergaben sich gewisse Unterschiede zwischen den einzelnen Sprechern, aber auch aufgrund der Faktoren 'Konsonant' und 'Kontext'; letztere können z.T. anhand von Markiertheitsüberlegungen erklärt werden.

Schlüsselbegriffe: Stimmhafte Obstruenten, Zweitspracherwerb, Italienisch, Schweizerdeutsch

1. Einleitung

Die Phonetik und Phonologie des Zweitspracherwerbs ist ein blühender Forschungszweig, wie nicht zuletzt verschiedene theoretische Modelle zeigen, von denen drei in der Folge kurz erwähnt werden (1.1.). Theoretische Annahmen und eine kontrastive Analyse zwischen den Obstruenten des Zürichdeutschen und des Italienischen (1.2.) führen zur Formulierung von drei Hypothesen und zwei Forschungsfragen (1.3.). Nach der Beschreibung des Vorgehens bei der Erhebung und Auswertung der Daten (2.) folgt die Darstellung der Resultate (3.), wobei die prozentuale Entstimmung der italienischen Obstruenten anhand der Faktoren 'Sprechende', 'Konsonant' und 'phonotaktischer Kontext' aufgezeigt wird. Eine kurze Diskussion der Ergebnisse schliesst den Beitrag (4.).

1.1. Theoretische Modelle des Lauterwerbs in einer Zweitsprache

Dass die phonologischen Strukturen der Erstsprache als eine Art 'Sieb' bei der Wahrnehmung der Laute einer Zweitsprache wirken, wurde bereits von Trubetzkoy [1, S. 47] erkannt und wird auch von neueren Theorien des fremdsprachlichen Lauterwerbs postuliert. So geht das *Speech Learning Model* (SLM) davon aus, dass Unterschiede zwischen 'ähnlichen' Sprachlauten in L1 und L2 durch die Lernenden kaum wahrgenommen werden, was zu einer Art 'Äquivalenz-Klassifikation' führt [2, S. 239]. Auch das *Perception Assimilation Model* nimmt in seiner Zweitspracherwerbsvariante (PAM-L2, [3], S. 22) an, dass "naive Hörer" Laute einer L2 an artikulatorisch ähnliche Phoneme der Erstsprache 'assimilieren'.

Neben diesen eher phonetisch und z.T. psychologisch ausgerichteten Theorien gibt es in der phonologischen L2-Forschung aber auch linguistisch orientierte Ansätze wie etwa die *Markedness Differential Hypothesis* (MDH) von Fred Eckman [4, S. 97-100], die sich auf Erkenntnisse aus der Sprachtypologie stützt. Gemäss dieser Hypothese ist z.B. das Merkmal [±stimmhaft] einfach zu erwerben, wenn das entsprechende Segment sich in einer intervokalischen Umgebung befindet; hingegen sind wortfinale stimmhafte Obstruenten schwierig, falls sie nicht schon in der Erstsprache vorkommen.

1.2. Obstruenten im Zürichdeutschen und im Italienischen

Das Obstruenteninventar des Zürcher Dialekts wird in der Tabelle 1 aufgeführt (vgl. [5], S. 144):

	Bilabial	Labio-dental	Alveolar	Palato-alveolar	Velar
Plosiv	p b̥		t d̥		k g̊
Frikativ		f v̥	s z̥	ʃ ʒ̊	x ɣ̊
Affrikate		p͡f	t͡s	t͡ʃ	k͡x

Tabelle 1. *Obstruenten im Zürichdeutschen*

In dieser Darstellung sind alle phonetischen Zeichen für 'stimmhafte' Obstruenten mit einem Diakrit für Entstimmung (einem kleinen Kreis unter- oder oderhalb des IPA-Symbols) versehen. Dadurch wird ausgedrückt, dass die entsprechenden Phoneme zwar eine Position einnehmen, die in den phonologischen Systemen vieler Sprachen von stimmhaften Obstruenten besetzt ist (in der Tat stehen sie in Opposition zu einem homorganen stimmlosen Konsonanten). Allerdings sind diese als *lenes* bezeichnete Konsonanten durchwegs stimmlos: der Kontrast zu den homorganen *fortes* wird in erster Linie durch Dauerunterschiede und in zweiter Linie durch die Intensität realisiert. Das Merkmal *fortis* vs. *lenis* (bzw. die 'Spannungskorrelation') tritt konsequent bei allen Plosiven und Frikativen auf, während bei den Affrikaten nur die unmarkierte *fortis* erscheint. Eine Neutralisierung der Opposition erfolgt schliesslich bei zwei aufeinander folgenden *lenes*: bei solchen Verbindungen werden beide Konsonanten als so genannte Halb-*fortes* realisiert [5, S. 248].

Das Obstruenteninventar des Standarditalienischen wird in der Tabelle 2 dargestellt [6, S. 132]. Mit Ausnahme von /z/ kommen alle italienischen Obstruenten sowohl als einfache Konsonanten als auch als Geminaten vor. Die in der Tabelle als stimmhaft aufgeführten Phoneme werden in der Tat mit Beteiligung von Glottisschwingungen artikuliert. Die Stimmhaftigkeitskorrelation wird im Italienischen relativ stark bean-

sprucht: zwar klafft bei den palato-alveolaren Frikativen eine Lücke, dafür besteht bei den Affrikaten eine Opposition zwischen stimmlosen und stimmhaften Phonemen, während das Zürichdeutsche keine strukturell äquivalenten *lenes* aufweist (vgl. Tabelle 1).

	Bilabial	Labio-dental	Alveolar	Palato-alveolar	Velar
Plosiv	p b		t d		k g
Frikativ		f v	s z	ʃ	
Affrikate			t͡s d͡z	t͡ʃ d͡ʒ	

Tabelle 2. *Obstruenten im Italienischen*

1.3. Hypothesen und Forschungsfragen

Aufgrund der theoretischen Betrachtungen in 1.1. und der kontrastiven Analyse in 1.2. kann man folgende Hypothesen aufstellen:

- H1: Aufgrund der Ähnlichkeit zwischen den italienischen stimmhaften Obstruenten und den zürichdeutschen *lenes* neigen Lernende dazu, bei gleichem Artikulationsort und gleicher Artikulationsart die L2-Konsonanten an solche der L1 zu 'assimilieren' (vgl. SLM und PAM-L2).
- H2: Aufgrund der Lücke im zürichdeutschen Phoneminventar wird die Entstimmung bei den Affrikaten /d͡z/ und /d͡ʒ/ am stärksten ausfallen.
- H3: Bei aufeinanderfolgenden Obstruenten wird die Tendenz zur Entstimmung ebenfalls deutlicher auftreten.

Neben diesen spezifischen Hypothesen können aber auch Forschungsfragen von allgemeinerer Natur gestellt werden:

- F1: Gibt es sprecherspezifische Unterschiede bezüglich der Entstimmung der italienischen Obstruenten?
- F2: Gibt es (neben den aufgrund der spezifischen Sprachenkombination auftretenden Schwierigkeiten) Erscheinungen, die im Sinne von Eckmans Markiertheitshypothese interpretiert werden können?

2. Die empirische Untersuchung

2.1. Die Population

Für die Untersuchung der drei spezifischen Hypothesen und die beiden allgemeinen Forschungsfragen wurde ein Lesesprache-Korpus erhoben. Versuchspersonen waren 10 Schüler eines Zürcher Gymnasiums (davon 6 weiblichen und 4 männlichen Geschlechts). Die Schüler waren 17 Jahre alt und hatten seit zwei Jahren wöchentlich drei Stunden Italienischunterricht. Die Sprechenden werden in der Folge mit den ersten drei Buchstaben ihres Vornamens gekennzeichnet (für weitere Angaben siehe [7].

2.2. Das Korpus

Die Sprachaufnahmen fanden während des Italienischunterrichts in einem von der Schule zur Verfügung gestellten Raum statt und wurden mit einem digitalen Audio-Recorder sowie einem Ansteckmikrophon mit Kugelcharakteristik durchgeführt (Abtastrate: 44.1 kHz, Quantisierung: 16 bit).

Die Schüler lasen 19 italienische Sätze mit Wörtern, in welchen die 7 Phoneme /b d g v z d͡z d͡ʒ/ eingebaut waren, und zwar in den folgenden sechs Positionen: i) im absoluten Anlaut der Äusserung (##_V); ii) im Wortanlaut nach Vokal (V#_V); iii) im Wortanlaut nach Sonorant (C#_V): iv) im Wortinnern als intervokalische Simplex (V_V); v) im Wortinnern als intervokalische Geminate (V:_V); vi) nach stimmhaftem Sibilant (z_V).

Von den 670 möglichen Realisierungen (10 Schüler x 67 Wörter mit stimmhaften Obstruenten) konnten aber nur 644 ausgewertet werden, da in 26 Fällen Performanzfehler beim Lesen der Sätze auftraten.

Die Audio-Dateien wurden in Praat [8] anhand von TextGrids segmentiert und annotiert, um anschliessend für die ausgewählten Segmente den prozentualen Anteil der entstimmten Phase an der Gesamtdauer zu messen.

3. Ergebnisse

3.1. Globaler Entstimmungsanteil

Von den 644 gemessenen Obstruenten wurden ingesamt 217 (33.7%) als voll stimmhaft realisiert, während nur 12 (1.8%) ganz entstimmt waren. In 415 Fällen (64.4%) trifft man hingegen 'Zwischenlösungen' an, wobei der Entstimmungsanteil von 11% bis zu 97% reicht. In der Folge teilen wir die Realisierungen in zwei Kategorien ein, nämlich in 'stimmhaft' und '(teilweise) entstimmt'. Bei einem zu 86% entstimmten /d/ gehen wir davon aus, dass dieses in der Phonologie der Lernersprache dem L1-Phonem /d̥/ entspricht und dass die im Sprachsignal zu erkennende Periodizität zu Beginn der Verschlussphase auf einer Koartikulationserscheinung mit dem vorhergehenden Vokal (*VoiceOFFsetTime*) beruht.

Insgesamt ist also ein gewisser Einfluss des 'phonologischen Siebs' der Muttersprache nicht von der Hand zu weisen. Die Schüler sind aber zum Teil durchaus in der Lage, die italienischen Obstruenten gemäss der Norm der Zielsprache zu artikulieren. Allerdings ist in den Daten naturgemäss eine gewisse Variabilität zu erkennen, so dass in der Folge nicht nur Unterschiede zwischen den Sprechenden betrachtet werden sollen (3.2.), sondern auch der Einfluss der Faktoren 'Konsonant' (3.3.) und 'Kontext' (3.4.).

3.2. Der Faktor 'Sprechende'

Versuchen wir zunächst, die Forschungsfrage 1 zu beantworten: Gibt es sprecherspezifische Unterschiede bezüglich der Entstimmung der italienischen Obstruenten?

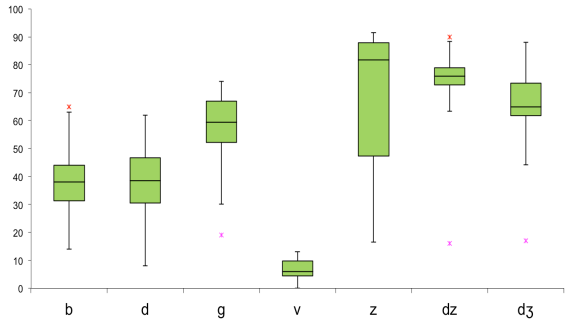


Abbildung 1: Entstimmungsanteil der Obstruenten bei den verschiedenen Sprechern

Das Säulendiagramm in Abbildung 1 zeigt bei 8 Schülern einen grösseren Anteil der (teilweise) entstimmten gegenüber den stimmhaften Realisierungen. Nur bei Eli und San treffen wir ein umgekehrtes Verhältnis an (ca. ein Fünftel der Obstru-

enten werden entstimmt), wobei dieser Befund vermutlich mit der Mehrsprachigkeit der beiden Individuen zusammenhängt: San hat einen italienischsprachigen Hintergrund und Eli spricht in der Familie auch Schwedisch – eine Sprache, die im Prinzip stimmhaft realisierte Obstruenten kennt, auch wenn diese in stimmloser Umgebung entsonorisiert werden [9].

Analysiert man die (teilweise) entstimmten Obstruenten bei den restlichen 8 Sprechenden, so findet man eine beträchtliche Varianz: die Schülerin Van realisiert z.B. ein zu 92% entstimmtes [d͡z], daneben aber auch ein nur zu 14% entstimmtes [b]. Im Durchschnitt beträgt der entstimmte Anteil an der Dauer der Realisierungen 47%, mit einer Spanne von 37% bei And bis zu 57% bei Yit (Standardabweichung 7%). Untersucht man den Zusammenhang zwischen der Anzahl der (teilweise) entstimmten Realisierungen und dem durchschnittlichen aperiodischen Anteil im Signal, so ergibt sich eine klare Korrelation ($r=0.76$, $R^2=0.57$).

Der globale Einfluss des Faktors 'Sprechende' lässt sich prüfstatistisch allerdings nicht mit einem parametrischen Verfahren testen, da die Bedingungen der Normalverteilung und Varianzgleichheit nicht erfüllt sind. Immerhin ergibt aber der nicht-parametrische Kruskal-Wallis-Test signifikante Unterschiede zwischen den 10 Schülern ($\chi^2(9)=93{,}584$, $p<.001$).

3.3. Der Faktor 'Konsonant'

Inwieweit der Grad an Entstimmung durch den jeweiligen Konsonanten bestimmt wird, lässt sich mit unseren Daten auch nur anhand eines nicht-parametrischen Tests überprüfen, da aufgrund der fehlenden Normalverteilung und Varianzgleichheit keine ANOVA durchgeführt werden kann. Auch in diesem Fall zeigt der Kruskal-Wallis-Test aber einen signifikanten Effekt für den Faktor 'Konsonant' ($\chi^2(6)=266{,}048$, $p<.001$).

Aufschlussreicher ist eine rein deskriptive Analyse der Daten, wie sie anhand der Boxplots in der Abbildung 2 vorgenommen werden kann:

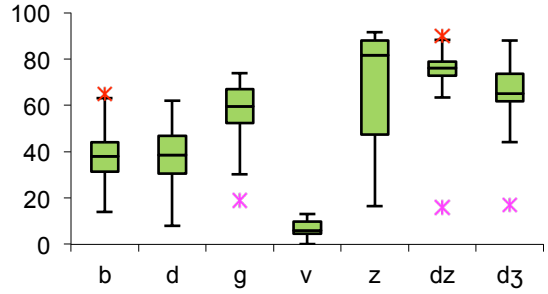


Abbildung 2: Entstimmungsanteil aufgrund des Faktors 'Konsonant'

Betrachtet man zunächst die drei stimmhaften Plosive auf der linken Seite der Grafik, so fällt auf, dass [g] im Vergleich zu [b] und [d] einen höheren Entstimmungsanteil aufweist – ein Befund, der sich nicht auf eine Interferenz der L1 zurückführen lässt. Hingegen ergibt sich eine interessante Parallele zur Phonemtypologie: in der Datenbank PHOIBLE [10] kommen labiale und koronale stimmhafte Plosive in 71% bzw. 72% der 2155 dokumentierten Sprachen vor, während /g/ nur in 64% der Sprachen vertreten ist. Die häufigere Entstimmung der velaren gegenüber den anderen Plosiven passt somit zu Eckmans MDH (Forschungsfrage 2). Die typologische Markiertheit (und somit die Schwierigkeit des entsprechenden Konsonanten im L2-Erwerb) findet ihre Erklärung in der Aerodynamik der Sprachproduktion: es ist schwierig, bei einem kurzen blockierten Ansatzrohr zu phonieren [11].

Auf der rechten Seite der Abbildung 2 zeigen die beiden Boxplots der Affrikaten /d͡z/ und /d͡ʒ/ einen relativ hohen Entstimmungsanteil mit eher geringer Streuung. Diese höheren Entsonorisierungswerte bestätigen unsere zweite Hypothese H2 (vgl. 1.3): aufgrund der strukturellen Lücke im L1-Lautinventar scheinen die Lernenden die unbekannten L2-Phoneme an die homorganen stimmlosen Affrikaten /t͡s/ und /t͡ʃ/ zu assimilieren. Abbildung 3 illustriert die totale Entstimmung dieser Konsonanten am Beispiel der Aussprache des italienischen Wortes *mezzogiorno* 'Mittag'.

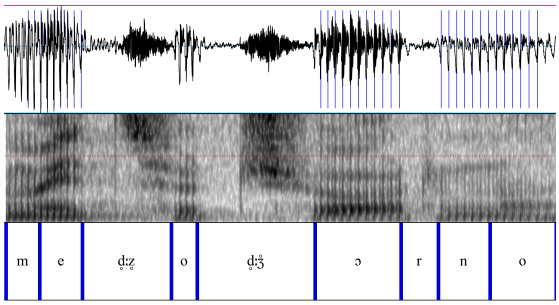


Abbildung 3: Wellenform und Spektrogramm der Aussprache von mezzogiorno *durch Sprecher* Dar

Gänzlich verschieden sind hingegen in der Mitte der Abbildung 2 die Boxplots der beiden Frikative /v/ und /z/: beim labiodentalen Konsonanten finden wir einen niedrigen Entstimmungsanteil mit geringer Varianz, während die Werte des Sibilanten höher liegen und eine grössere Streuung aufweisen. Die Sprachtypologie liefert hier keine Erklärung, da gemäss PHOIBLE beide Phoneme in ca. 30% der Sprachen vorkommen. Vielmehr sind zwei unterschiedliche phonologische Eigenschaften der L1 zu beachten. Auf der einen Seite verfügt das Zürichdeutsche über einen labiodentalen Approximanten /ʋ/ [5], an welchen der italienische Frikativ assimiliert werden kann. Auf der anderen Seite steht /z/ im Korpus oft in Verbindung mit anderen Obstruenten, weshalb die in 1.2. erwähnte schweizerdeutsche Neutralisierung der *lenes* zu Halb-*fortes* zum Tragen kommt, wodurch auch die dritte Hypothese H3 bestätigt wird (vgl. 1.3.).

Der Fall von /z/ führt uns zum dritten Faktor, der den unterschiedlichen Entsonorisierungsanteil der stimmhaften Obstruenten in den italienischen Wörtern beeinflusst, nämlich deren phonotaktische Umgebung.

3.4. Der Faktor 'Kontext'

Wie bei den anderen beiden Faktoren konnte auch der der Einfluss des phonotaktischen Kontexts nicht mit einem parametrischen Test überprüft werden, da die Voraussetzungen aufgrund der vorgängigen Inspektion der Daten anhand der Tests von Shapiro-Wilk und Levene nicht gegeben waren. Der nicht-parametrische Kruskal-Wallis-Text ergibt jedoch auch für den Faktor 'Kontext' einen signifikanten Effekt ($\chi^2(5)=76{,}027$, $p<.001$).

Eine gewisse Abhängigkeit des Entstimmungsanteils ist auch rein deskriptiv in den Boxplots der Abbildung 4 erkennbar. Intervokalisch findet man im Wortinnern (V_V, V:_V) die niedrigsten Mittelwerte (35%, 38%). Die Wortgrenze scheint nach Sonorant (C#_V) keine Rolle zu spielen, nach

Vokal (V#_V) hingegen schon; weniger überraschend ist der etwas höhere Mittelwert (45%) am Anfang der Äusserung (##V_V), da in diesem Kontext phonatorisch zuerst eine 'Einstimmung' nötig ist. Insgesamt könnte man dieses Ergebnis als (wenn auch nicht besonders deutliche) Evidenz für Eckmans Markiertheitshypothese betrachten.

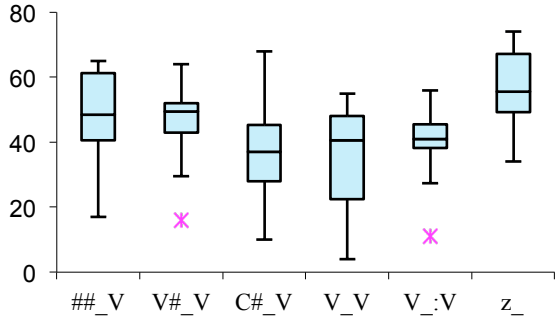


Abbildung 4: Entstimmungsanteil aufgrund des Faktors 'Phonotaktischer Kontext'

Einfacher zu interpretieren ist der Boxplot des Kontexts z_V ganz rechts in Abbildung 4. Wie bereits in 3.3. bemerkt wurde, fördert die Verbindung von zwei stimmhaften Obstruenten die Entsonorisierung von beiden Segmenten, wie die Mittelwerte von 68% für /z/ und von 57% für den darauf folgenden Konsonanten bezeugen. Auch in diesem Fall geht somit die Interferenz der L1 (Neutralisierungsregel des Zürichdeutschen) einher mit einer allgemeinen Markiertheitsbedingung (Schwierigkeit der andauernden Schwingung der Stimmlippen während der Artikulation von zwei Obstruenten). Die Abbildung 5 illustriert die Entstimmung der Konsonantenfolge /zv/ im italienischen Wort *svizzero* 'Schweizer'.

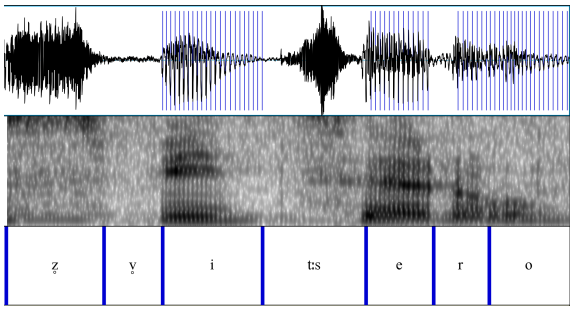


Abbildung 5: Wellenform und Spektrogramm der Realisierung des Wortes svizzero *durch Sprecher* Mik

4. Diskussion und Schluss

Insgesamt haben die Resultate die drei eingangs aufgestellten Hypothesen bestätigt. Die Tatsache, dass die meisten Schüler die stimmhaften Obstruenten des Italienischen eher (zumindest teilweise) entsonorisieren als stimmhaft aussprechen, weist darauf hin, dass sie die L2-Phoneme aufgrund ihrer Ähnlichkeit an die *lenes* ihrer L1 assimilieren (H1). Ebenfalls bestätigt wurden H2 und H3: die Entstimmung fiel deutlicher aus bei den Affrikaten (aufgrund der strukturellen Lücke in der L1) sowie bei den Verbindungen s+C (aufgrund der Neutralisierungsregel der L1).

Auch die beiden Forschungsfragen können zumindest zum Teil beantwortet werden. Trotz gemeinsamer Tendenzen können auch gewisse Unterschiede zwischen den Sprechern ausgemacht werden (F1), wobei insbesondere der Bilingualismus von zwei Schülern deren niedrigeren Entstimmungsanteil erklärt. Sprachliche Markiertheit im Sinne von Eckman kann als Erklärung angeführt werden für Unterschiede beim Faktor 'Konsonant' (grössere Entstimmung bei velaren Plosiven) sowie beim Faktor 'Kontext' (grössere Entstimmung am Anfang der Äusserung und bei Obstruentenverbindungen). Ob nun die 'Assimilation' der L2-Phoneme an die ähnlichen L1-Konsonanten auf der Stufe der Perzeption erfolgt (wie SLM und PAM-L2 suggerieren) oder ob die Interferenz der L1 erst bei der effektiven Aussprache erfolgt (wie die artikulatorisch basierten Markiertsheitsüberlegungen nahelegen würden), kann durch diese Studie nicht geklärt werden, da ausschliesslich die Sprachproduktion untersucht wurde.

5. Dank

Herzlich bedanken möchte ich mich bei den Schülerinnen und Schülern des Realgymnasiums Rämibühl in Zürich sowie bei der Italienischlehrerin Letizia Könz. Mein grösster Dank geht an Sarah Wachter für die Durchführung der Sprachaufnahmen und die Annotation der Audiodateien. Merci auch an Sandra Schwab für die Hilfe bei der statistischen Auswertung der Daten.

6. Bibliographie

[1] N. Trubetzkoy, *Grundzüge der Phonologie*. Prague: Cercle Linguistique, 1939.

[2] J. Flege, "Second speech learning. Theory, findings, and problems", in W. Strange (ed.), *Speech perception and linguistic experience*, Timonium MD: York Press, pp. 233-277, 1995.

[3] C. Best and M. Tyler, "Nonnative and second language speech perception: Commonalities and complementarities", in O. Bohn and M. Munro (eds.), *Language experience in second language speech learning: In honor of James Emil Flege*, Benjamins: Amsterdam, pp. 13-33, 2007.

[4] F. Eckman, "Typological markedness and second language phonology", in J. Hansen Edwards and M. Zampini (eds.), *Phonology and second language acquisition.* Benjamins: Amsterdam, pp. 95-115, 2008.

[5] J. Fleischer and S. Schmid, "Zurich German", *Journal of the International Phonetic Association*, vol. 36, no. 3, pp. 243-255, 2006.

[6] P.M. Bertinetto and M. Loporcaro, "The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome", *Journal of the International Phonetic Association*, vol. 35, no. 2, pp. 131-151, 2005.

[7] S. Schmid and S. Wachter, "Le ostruenti sonore nella pronuncia dell'italiano da parte di apprendenti svizzero-tedeschi", *Studi AISV*, vol. 1, pp. 191-206, 2016.

[8] P. Boersma and D. Weenink, *Praat: doing phonetics by computer* [Computer program]. Version 4.1.5, retrieved 1 June 2015 from http://www.praat.org/.

[9] O. Engstrand, "Swedish", in *Handbook of the International Phonetic Association*, Cambridge: Cambridge University Press, pp. 140-142, 199.

[10] S. Moran, D. McCloy and R. Wright, *PHOIBLE online.* Leipzig: Max Planck Institute for Evolutionary Anthropology (available at http://phoible.org).

[11] J . Ohala, "What's cognitive, what's not, in sound change", in G. Kellerman and G. Morrissey (eds.), *Diachrony within synchrony: language history and cognition.* Frankfurt a.M.: Peter Lang, pp. 309-355, 1992.

Monolingual and trilingual production of Northern Standard German vowels

Heike Schoormann[1], *Wilbert Heeringa*[2], *Jörg Peters*[1]

[1]Institute for German Studies, University of Oldenburg, Oldenburg, Germany
[2]Fryske Akademy, Ljouwert, the Netherlands

heike.schoormann@uol.de, wjheeringa@gmail.com, joerg.peters@uol.de

Abstract

Studies on vowel productions of speakers from bilingual communities report L1-L2 interactions but also monolingual-like realizations ([1], [2], [3]). Where the languages differed in communicative range and size of the speech community, monolingual-like productions of early bilinguals were found in the languages with the wider communicative range and larger speech community. We compare the acoustic realizations of Northern Standard German (NSG) vowels in monolingual speakers from Hanover, representing the larger speech community of Northern Germany, and in trilingual speakers from the Saterland, speaking the local variant of High German, Low German, and Saterland Frisian. To examine whether the NSG vowels of the Saterland speakers approached the vowels of the monolingual speakers in terms of spectral and durational features, we elicited all stressed NSG monophthongs in /hVt/ context. Our data show an orientation towards the larger speech community of Northern Germany in the productions of the trilinguals. Vowel productions which neither differed across the trilinguals' three languages nor from the monolinguals suggest contact-induced phonetic convergence towards NSG. The observed bidirectional interaction of the trilinguals' three vowel systems further supports the claim that all vowel categories are organized in a common phonological vowel space.

Index Terms: Northern Standard German, vowel production, trilingualism, phonetic interference, Low German

1. Introduction

Research on second-language acquisition and bilingualism shows that the sounds of the languages acquired influence each other mutually in the production of L1 and L2 categories (cf. [4], [5]). The observed cross-linguistic interactions suggest that the vowels of the L1 and L2 are organized in a common phonological space (cf. [6], [7], [8], [1]). Despite language-specific categories and near-monolingual-like performance in one or both of the acquired languages, early and simultaneous bilinguals from bilingual communities showed some effect of cross-linguistic interference in the production of vowel categories ([1]; [2]). In a recent study on the vowel productions of speakers from a Welsh-English community, [3] studied the substrate effect of Welsh and present a case in which interference is observable in terms of large-scale phonetic convergence in the context of regional bilingualism (cf. [9]). In [1] and [3] monolingual-like productions were observed in the languages of the early bilinguals that have the wider communicative range and larger speech community, i.e. Spanish and English.

Similar to the findings in [3], the trilingual speakers of Saterland Frisian (SF), Low German (LG), and Northern Standard German (NSG) studied by [10] showed mergers of vowel categories used in two or all three of their languages. The trilingual speakers were recruited from Scharrel, a village in the municipality of Saterland, located in the northwestern part of Lower Saxony in Northern Germany. Whereas no systematic differences were reported in the trilinguals' phonetic realizations of corresponding categories between the two languages confined to the close-knit community of Saterland, SF and the local variety of LG, deviant realizations were found for several vowel categories in NSG, which is spoken by several millions of speakers in Northern Germany.

The question that arises from the previous acoustic investigation of vowel productions in Saterland trilinguals is whether the deviant realizations of NSG vowels found in the Saterland point into the direction of NSG vowels produced by monolingual speakers of NSG. The present study expands the acoustic investigation in [10] by studying the substrate effect of SF and LG on the standard language through a comparison with monolingual speakers of the standard variety, which are more representative of the wider speech community of Northern Germany. To this end, the NSG vowel productions of the trilingual Saterland speakers reported in [10] are compared with those of monolingual speakers of NSG from Hanover. Hanover is the state capital of Lower Saxony, about 170 kilometers from Scharrel. The High German variety of Hanover is commonly considered most typical of NSG and the language used in the Northern German media.

We hypothesize that the NSG vowel productions of the Saterland speakers approach the productions of the monolingual speakers in terms of spectral and durational features. In particular, we expect that those vowel categories which showed deviant realizations in NSG in the study of [10] can be identified as instances of contact-induced phonetic convergence towards NSG.

2. Method

2.1. Participants

Twenty-three male native speakers participated in the study, 11 trilinguals from Scharrel and 12 monolinguals from Hanover. The trilingual speakers had lived in Scharrel all their lives and thus had extensive exposure to all three languages from birth. All subjects considered SF as their mother tongue and primary home language. Even though they differ somewhat in their reports on the order and the age of acquisition of LG and High German, the Scharrel subjects may all be categorized as early sequential trilinguals in the sense of [11] because all speakers were exposed to the three languages from

early childhood on within the Saterland (SF, LG, NSG), through contact with people outside of the Saterland (LG and NSG), and through the media (NSG). The monolingual Hanover subjects lived and grew up within the Hanover region with a maximum distance from the city center of about 50 kilometers. Only one subject deviates from this profile, growing up in Lüneburger Heide about 90 kilometers north of Hanover but having lived in Hanover for the larger part of his life. The 23 subjects were all aged between 50 and 75.

2.2. Material and recording procedure

The vowel systems of the three languages spoken in the Saterland share the majority of vowel categories. All 15 monophthongs of High German (/i: y: u: e: ø: o: ɛ: a: ɪ ʏ ʊ ɛ œ ɔ a/) have an equivalent in SF and LG. In addition, SF and LG have two long lax open-mid vowels, /œ:/ and /ɔ:/.

In spoken NSG the vowel phoneme /ɛ:/, which is believed to be mostly due to spelling pronunciation, tends to be merged with /e:/ and thus realized as a close-mid vowel (cf. [12], [13, p.50], [14], [15, p. 172f.]). In careful speech, however, the opposition between /e:/ and /ɛ:/ in hVt context may be upheld even by North German speakers (cf. [16, p. 79]). In the trilingual inventory of the Saterland speakers, the distinction between /ɛ:/ and /e:/ may get further support by the fact that SF and LG have two more open-mid long lax vowels, /œ:/ and /ɔ:/, which contrast with both the long tense vowels /ø:/ and /o:/ and the short lax vowel /œ/ and /ɔ/.

The 15 High German vowel categories were recorded in monosyllabic /hVt/ context. For all sessions, a native monolingual speaker of NSG guided the participants through the experiment. The /hVt/ words were elicited via rhymes in sequences of High German triggers followed by the /hVt/ target word. For this matter, the informants were first instructed to read aloud the High German trigger word displayed on the computer screen and to produce the rhyming /hVt/ target word (e.g. Boot 'boat' – Hoot) subsequently, with only the frame H_t presented on the computer screen (cf. [17], [18]). Target words were elicited as rhyming words and not displayed directly to account for a possible influence of the written form on the production data. All informants were instructed to not overarticulate but to pronounce the target word in a more habitual style.

The sequences of trigger and target words were presented in a controlled randomized order to secure that a vowel was never directly succeeded by the same vowel in the following sequence. Each sequence of a trigger and the rhyming target word was presented three times per speaker, thus eliciting a sample size of 45 tokens per subject (15 monophthongs × 3).[1] The recordings were monitored for the target pronunciation and intonation to ensure that all /hVt/ words were elicited with a falling contour. Where mistakes occurred, individual sequences were repeated at the end of each recording session. The first three valid productions of each target vowel and speaker were analyzed. Six practice sequences preceded all blocks. The recordings were made with a Tascam HD P2 digital recorder and a head-mounted microphone (DPA 4065 FR) in a quiet room and digitized at a sampling rate of 48 kHz.

2.3. Acoustic analysis

All acoustic analyses were done with the Praat software package ([19]). Measured acoustic variables included F1 and F2 at vowel midpoint, as well as vowel duration. In addition, we calculated the duration ratios for the vowel pairs of short/lax and long/tense monophthongs. Only the vowel midpoint frequencies were included in the analysis since neither High German nor SF monophthongs are diphthongized (cf. [16, p. 86], [20], [21]). Onset and offset of the vocalic segment were labeled manually for each /hVt/ word. Vowel onset was measured at the zero-crossing before the first positive peak in the periodic waveform. Vowel offset was set at the last negative-to-positive zero-crossing before the (abrupt) reduction in amplitude and/or cessation of periodicity in the waveform before the stop closure.

A Praat script was used to automatically estimate the frequencies of the first and second formant. The window length was set to 0.025 seconds. Formant settings for the LPC analysis were adapted upon visual inspection for each realization individually in the script by de- or increasing the LPC order in steps of 1 (default order of 10) and the maximum frequency in steps of 500 Hz (default 5000 Hz). Outliers due to measurement errors were corrected by hand.

A normalization of the data is necessary to mitigate variation caused by physiological differences among the different speakers while preserving sociolinguistic variation. We followed the normalization method applied in Guion (2003). In a first step we converted the Hertz data to the Bark scale using Traunmüller's (1990) formula (1).

$$z = [26.81/(1+1960/F_i)] - 0.53 \quad (1)$$

where F_i is the value for a given formant *i*.

Subsequently, we normalized the Bark formant values through the multiplication with a speaker-specific k factor, which is derived by dividing one fixed subject's average F3 ($F_3\ S_{median}$) of the open vowel (/a/) by speaker *j*'s mean F3 ($F_3\ S_j$), using the formula in (2).

$$\text{mean } F_3\ S_{median}/\text{mean } F_3\ S_j = k_j \quad (2)$$

All calculations in this study are based on normalized formant values.

2.4. Statistical processing

Function lmer from the *lme4* package [22] was used to perform linear mixed effects analyses in R [23]. As dependent variables we used *duration*, *duration ratio*, *F1*, and *F2*. As fixed effect, we entered the variables *speaker group*, which distinguishes between monolingual speakers and trilingual speakers, and *repetition* into the full model. As random effects, we had intercepts for *speaker* and *vowel* (or *vowel pair* in the comparison of the duration ratios) as well as by-vowel (pair) random slopes for the effect of *speaker group*. In addition, by-speaker random slopes for the effect of *repetition* were included in the full model. When comparing speaker groups per individual vowel/vowel pair, only random intercepts for *speaker* and by-speaker random slopes for *repetition* were included in the full model as the random effects. A backward elimination of non-significant effects of each full model was performed with the step function of the *lmerTest* package [24]. All *p*-values were calculated using the Satterthwaite approximation in the *lmerTest* package.

3. Results

3.1. Vowel duration

Figure 1 shows the mean vowel duration for the long and short monophthongs of NSG averaged over each speaker group.

Within both speaker groups, a clear separation between long and short vowels is confirmed by a linear mixed effects model with vowel duration as the dependent variable and random intercepts for speaker (MON_NSG long vs. short: β=121.88, SE=3.29, t(522)=37.08, p<.001, TRI_NSG long vs. short: β=83.85, SE=2.75, t(474)=30.48, p<.001). No general effects of *speaker group* were found regarding the subgroups of long and short vowels, i.e. the two speaker groups do not differ in the realization of all long or short categories regarding vowel duration. The comparison of single long vowel categories, however, shows significant durational differences between the two speaker groups for /a:/ (β=43.62, SE=19.37, t(23)=2.25, p<.05) and /e:/ (β=40.56, SE=18.05, t(23)=2.25, p<.05). In both cases the vowel durations are longer for the monolingual speakers. For the short vowels, no effect of *speaker group* on vowel duration was found.

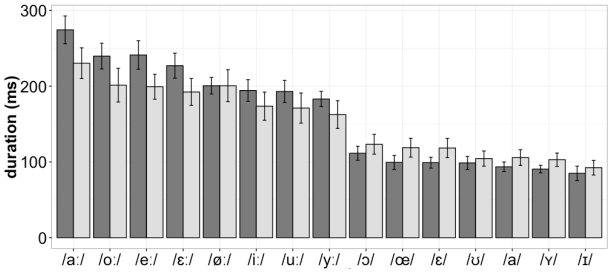


Figure 1: *Mean duration of monophthongs averaged over all speakers per group (monolinguals = dark grey, trilinguals = light grey). Error bars represent 95% confidence intervals of the means.*

[10] found the longest mean durations for the trilinguals' productions in NSG compared to SF and LG but no differences between the SF and LG productions of monophthongs. This effect is most pronounced in the subgroup of long vowels. The mean long vowel duration of the monolinguals exceeds all of the trilingual values, among which the NSG values are the highest.

	ratio monoling.	ratio triling.	β	SE	t(df)	p
a:-a	3.0	2.3	0.75	0.20	3.84(23)	<.001
ɛ:-ɛ	2.3	1.7	0.71	0.16	4.43(21)	<.001
e:-ɛ	2.5	1.8	0.75	0.20	3.73(23)	<.01
ø:-œ	2.2	1.7	0.47	0.16	2.88(23)	<.01
o:-ɔ	2.3	1.7	0.57	0.17	3.32(23)	<.01
i:-ɪ	2.5	1.9	0.56	0.21	2.70(23)	<.05
y:-ʏ	2.1	1.6	0.31	0.10	2.99(23)	<.01
u:-ʊ	2.0	1.7	-	-	-	n.s.
mean	2.4	1.8				

Table 1: *Mean duration ratios measured per vowel pair for each speaker group and the overall mean ratio of each speaker group averaged over all pairs.*

Table 1 illustrates the mean duration ratios for the vowel pairs of short/lax and long/tense monophthongs per group. Averaged over all vowel pair ratios, long vowels are 140% longer than short vowels in the productions of the monolinguals. In the trilinguals, long vowels are only 80% longer. On average, vowel duration differences in phonological short/lax and long/tense oppositions are smaller for the trilingual speaker group. Per vowel pair we compared the duration ratios of the monolinguals with the High German duration ratios of the trilinguals. We found a significant difference for all pairs but /u:-ʊ/. Regarding the pairs /a:-a/ and /e:-ɛ/, the difference between the speaker groups is directly attributable to the longer durations of the long tense vowels in the productions of the monolinguals.

Considering the data from [10], the differences in the duration ratios are similar to the differences observed above for the absolute durations. The trilinguals' average short/lax - long/tense ratios in NSG are higher than the non-distinct SF and LG ratios but smaller than the average ratios of the NSG monolinguals.

3.2. Formant frequencies

Figure 2 shows mean formant values for the 15 measured NSG monophthongs averaged over all subjects per group. No overall effect was found for the relative location of the vowel phonemes within the F1-F2 plane, suggesting that there is no general shift in F1 or F2 between the two speaker groups.

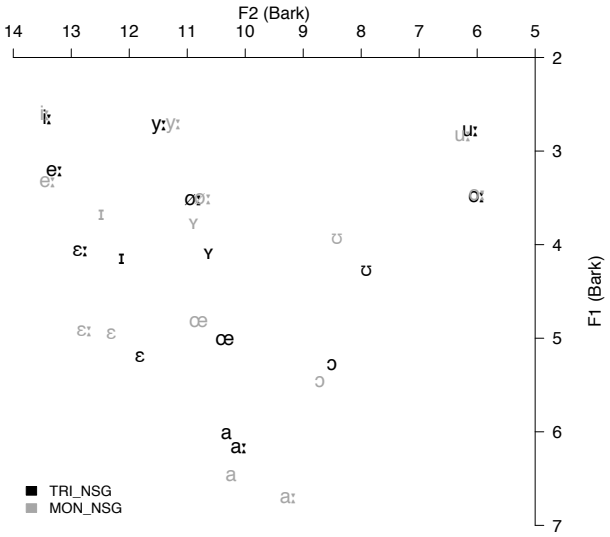


Figure 2: *Mean normalized F1-F2 values of 15 monophthongs measured at vowel midpoint.*

We applied mixed effects models in order to explore the relationships among /ɛ:-ɛ/, /e:-ɛ:/, and /a:-a/ within the speaker groups. The linear mixed effects models were carried out with either *F1* or *F2* as dependent variable, *vowel category* as fixed factor, and random intercepts for *speaker*. The results of the pairwise comparisons for *vowel category* suggest that contrary to prior reports for NSG (e.g. [16, p. 81]) the two speaker groups distinguish between long /ɛ:/ and short /ɛ/ both by duration and by vowel quality (F2 MON_NSG: β=-0.33, SE=0.07, t(54)=-4.46, p<.001; F2 TRI_NSG: β=-1.02, SE=0.06, t(54)=-17.36, p<.001). Long /ɛ:/ and short /ɛ/ are not distinguished by F1 within the monolingual vowel space but within the trilingual vowel space (F1 TRI_NSG: β=1.14, SE=0.04, t(54)=26.01, p<.001). While the /ɛ:/-/ɛ/ opposition is secured by a clearer qualitative distinction in the trilingual vowel space, it is supported by the greater difference in duration in the monolingual vowel space (see 3.1, table 1). Moreover, the results of the pairwise vowel comparisons showed that the monolingual speakers did not neutralize the distinction between /e:/ and /ɛ:/ (F1: β=-1.59, SE=0.08, t(58)=-19.22, p<.001; F2: β=0.78, SE=0.07, t(54)=11.11, p<.001; see footnote 1).

NSG is often reported to distinguish /a/ and /a:/ primarily by duration (cf. [25, p. 59]). As in the study by [16, p. 81], we

found that the /a-aː/ opposition was not only upheld in terms of vowel duration (MON_NSG: β=-180.92, SE=7.72, t(60)=-23.44, p<.001; TRI_NSG: β=-124.64, SE=8.12, t(55)=-15.35, p<.001)) but also in terms of acoustic quality (F1 MON_NSG: β=-0.24, SE=0.07, t(60)=-3.45, p<.01; F2 MON_NSG: β=0.97, SE=0.09, t(60)=11.00, p<.001; F1 TRI_NSG: β=-0.15, SE=0.07, t(55)=-2.14, p<.05; F2 TRI_NSG: β=0.20, SE=0.07, t(55)=2.99, p<.01).

Significant differences were found for the comparison of subgroups of short lax vowels: the trilingual speakers produced /ɪ ʏ ʊ/ with higher F1 values (β=-0.38, SE=0.09, t(23)=-4.16, p<.001) and lower F2 values (β=0.37, SE=0.17, t(22)=2.15, p<.05). Within the trilingual vowel space, the qualitative difference between the phonemically close short/lax and long/tense pairs /iː-ɪ/, /yː-ʏ/, and /uː-ʊ/ is enlarged by lowering of the short lax vowels. Secondly, in comparison to the monolingual productions, short lax /ɛ/ and /œ/ are also lowered and retracted in the trilingual vowel space (/œ/ F1: β=-0.27, SE=0.11, t(23)=-2.56, p<.05; /œ/ F2: β=0.44, SE=0.16, t(23)=2.82, p<.01; /ɛ/ F1: β=-0.24, SE=0.09, t(23)=-2.66, p<.05; /ɛ/ F2: β=0.49, SE=0.19, t(23)=2.59, p<.05). The larger separation of /ɛː/ and /ɛ/ in the trilingual vowel space as compared to in the monolingual vowel space is thus due to both a more close position of /ɛː/ (β=0.85, SE=0.12, t(21)=7.10, p<.001) as well as the lowering of /ɛ/.

Moreover, as with the oppositional pairs described above, the respective lowering of /ɛ/ and /œ/ in the trilingual vowel space results in an enhanced qualitative difference for the oppositions of long tense /eː øː/ with short lax /ɛ œ/. Thirdly, the trilinguals show a more close production of both open vowels and a less retracted realization of the long open category in comparison to the trilinguals (/a/ F1: β=0.44, SE=0.16, t(23)=2.81, p<.05; /aː/ F1: β=0.53, SE=0.17, t(23)=3.06, p<.01; /aː/ F2: β=-0.85, SE=0.13, t(23)=-6.31, p<.001).

All in all, the results for mid-vowel formant frequencies suggest a different internal organization of the two vowel spaces but they also show similarities in the production of shared categories. /iː yː uː/, /eː øː oː/, and /ɔ/ do not differ in phonetic quality between the monolinguals and trilinguals. Moreover, while no phonological opposition was found to be purely quantitative or qualitative in nature for either speaker group. The qualitative difference in the vowel space of the trilinguals is enlarged relative to the qualitative difference in the monolingual vowel space for the close oppositional pairs (/iː-ɪ/, /yː-ʏ/, /uː-ʊ/), the phonemically close-mid long/tense and short/lax oppositions /eː-ɛ/ and /øː-œ/, and the opposition /ɛː-ɛ/.

[10] found significant differences in F1 between the trilinguals' SF and NSG productions of /iː yː uː/, /ɪ ʊ ʏ/, and /ɛː/. The trilinguals produce the long vowels /iː yː uː/ monolingual-like and the short /ɪ ʊ ʏ/ with intermediate F1-values. The position of /ɛː/, however, is neither monolingual-like nor intermediate but instead more close than the monolinguals' realizations of NSG as well as the trilinguals' realizations of SF.

4. Discussion

The majority of the trilinguals' NSG category realizations (10 of 15 categories) approach, or are similar to, the acoustic properties of the monolingual productions. Among these vowels, /oː/, /øː/, and /ɔ/ do not differ in F1 or F2 among the trilinguals' three languages [10]. Similar to [3], we argue that these category realizations present a case of contact-induced phonetic convergence between the two local languages and the standard language in a situation of regional trilingualism and extensive language contact. /ɪ ʏ ʊ/ follow the pattern observed by [1] and the respective hypothesis of the Speech Learning Model (hypothesis #6 in [7]) that language-specific categories are shifted and realized with intermediate values. The observed bidirectional interaction of the trilinguals' three vowel systems further supports the claim that all vowel categories are organized in a common phonological vowel space.

Among the remaining five[2] categories, three (/ɛ œ a/) were produced with values equal to the SF and LG merged categories [10]. The last vowel (/ɛː/) was produced more close than the LG/SF merged category and the productions of the monolinguals. A possible explanation for the high position of NSG /ɛː/ in the phonetic space of the trilinguals is a pull effect: /ɛː/ is produced with the same tongue height as /ɪ ʏ ʊ/ in the trilingual vowel space in SF and LG. Because /ɪ ʏ ʊ/ are produced more close as they approach the F1-F2 values of the monolingual speakers, the maintenance of the internal structure, i.e. the same degree of openness for /ɛː ɪ ʏ ʊ/ in all of the trilinguals' subsystems, would explain the more close position of NSG /ɛː/ as compared to LG and SF.

Our findings differ from [1] and [3] who both report all or the majority of the early bilingual vowel productions in the language with the supraregional communicative range and larger speech community to resemble those of monolingual speakers. In contrast to [1] our data need to be explained with recourse to the crowded trilingual vowel space and the preservation of phonemic contrasts. [2] even report (nearly) monolingual-like vowel productions in both of the bilinguals' languages. However, [2] studied only four of the 15 vowel phonemes distinguished within the bilinguals' language inventories. It is unclear whether further monolingual-like productions would be found in a comparison of all shared categories.

Acoustic distances between non-open short and open long positions are enlarged in the vowel system of the trilingual speakers with the exception of /ɔ-oː/. Primarily due to the relatively lowered short/lax vowels, the perceived vowel length distinction between oppositional pairs is likely to be enlarged by listener compensation (cf. [26]). The comparison of duration ratios revealed larger ratios for the monolingual speaker group. This finding suggests that duration is a more important cue for distinguishing long tense and short lax vowels in the monolingual speaker group, whereas the larger qualitative differences between long tense and short lax vowels in the trilingual speakers – and the entailed larger difference in perceived vowel length – is more important for the preservation of vowel contrasts in the trilingual system.

5. Acknowledgements

We thank all speakers for their participation and Darja Appelganz, Romina Bergmann, Dorothee Lenartz, Michaela Ballin, and Nicole Lommel for labeling the recordings in PRAAT. The research reported in this paper has been funded by the Deutsche Forschungsgemeinschaft, grant number PE 793/2-1.

6. References

[1] S. G. Guion, "The vowel systems of Quichua-Spanish bilinguals: An investigation into age of acquisition effects on the mutual

influence of the first and second languages," *Phonetica,* vol. 60, pp. 98–128, 2003.

[2] A. A. N. MacLeod, C. Stoel-Gammona, and A. B. Wassink, "Production of high vowels in Canadian English and Canadian French: A comparison of early bilingual and monolingual speakers," *Journal of Phonetics,* vol. 37, no. 4, pp. 374–87, 2009.

[3] R. Mayr and S. Montanari, "Differentiation and interaction in the vowel productions of trilingual children," *Proceedings of the 18th International Conference of Phonetic Sciences,* Paper number 1041.1-9, 2015.

[4] W. Baker and P. Trofimovich "Interaction of Native- and Second-Language Vowel System(s) in Early and Late Bilinguals," *Language and Speech*, vol. 48, pp. 1–27, 2005.

[5] J. E. Flege, C. Schirru, and I. R. A. MacKay, „Interaction between the native and second language phonetic subsystems," *Speech Communication,* vol. 40, pp. 467–491, 2003.

[6] Z. S. Bond, V. Stockmal, and D. Markus, "Sixty years of bilingualism affects the pronunciation of Latvian vowels," *Language Variation and Change,* vol. 18, pp. 165–177, 2006.

[7] J. E. Flege, "Second language speech learning: theory, findings, problems," in W. Strange (ed.), *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press, pp. 233-277, 1995.

[8] F. Grosjean, "Neurolinguists, beware! The bilingual is not two monolinguals in one person," *Brain and Language,* vol. 36, pp. 3–15, 1989.

[9] B. Bullock and C. Gerfen, "Phonological convergence in a contracting language variety," *Bilingualism: Language and Cognition,* vol. 7, pp. 95–104, 2004.

[10] W. Heeringa, H. Schoormann, and J. Peters, "Cross-linguistic vowel variation in Saterland: Saterland Frisian, Low German, and High German," *Proceedings of the 18th International Conference of Phonetic Sciences,* Paper number 1041.1-9, 2015.

[11] M. Sundara and L. Polka, "Discrimination of coronal stops by bilingual adults: The timing and nature of language interaction," *Cognition,* vol. 106, pp. 234–258, 2008.

[12] O.-S. Bohn and J. E. Flege, "The production of new and similar vowels by adult German learners of English," *Studies in Second Language Acquisition*, vol. 14, pp. 131–158, 1992.

[13] N. Fuhrhop and J. Peters, *Einführung in die Phonologie und Graphematik*. Stuttgart: Metzler, 2013.

[14] H.-P. Jørgensen, "Die gespannten und ungespannten Vokale in der norddeutschen Hochsprache mit einer spezifischen Untersuchung der Struktur ihrer Formantfrequenzen," *Phonetica,* vol. 19, pp. 217–245, 1969.

[15] K. J. Kohler, *Einführung in die Phonetik des Deutschen*. Berlin: Erich Schmidt Verlag, 1995.

[16] A. K. Steinlen, *The influence of consonants on native and non-native vowel production. A cross-linguistic study*. Tübingen: Gunter Narr, 2005.

[17] O.-S. Bohn, "How to organize a fairly large vowel inventory: The vowels of Fering (North Frisian)," *Journal of the International Phonetic Association*, vol. 34, pp. 161–173, 2004.

[18] R. Mayr, and H. Davies, "A cross-dialectal acoustic study of the monophthongs and diphthongs of Welsh," *Journal of the International Phonetic Association,* vol. 41, pp. 1–25, 2011.

[19] P. Boersma, and D. Weenink, *Praat: doing phonetics by computer*. (Version 6.0.19), <http://www.praat.org/>, 2016.

[20] H. Schoormann, H., W. Heeringa, and J. Peters, "Regional variation of Saterland Frisian vowels," *Proceedings of the 18th International Conference of Phonetic Sciences,* Paper number 1041.1-9, 2015.

[21] W. Strange and O.-S. Bohn, "Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies," *Journal of the Acoustical Society of America,* vol. 104, pp. 488–504, 1998.

[22] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software,* vol. 67, no 1, pp. 1–48, 2015.

[23] R Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, (Version 3.3.0), <http://www.R-project.org/>, 2016.

[24] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, *lmerTest: Tests in Linear Mixed Effects Models*. (R package version 2.0-32), <http://CRAN.R-project.org/package=lmerTest >, 2016.

[25] A. P. Simpson, *Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 33*. 1998.

[26] C. Gussenhoven, "A vowel height split explained: Compensatory Listening and Speaker Control," in J. Cole and J. I. Hualde (eds.), *Laboratory Phonology 9*. Berlin: Mouton de Gruyter, pp. 145–172, 2007.

---

[1] Two monolinguals did not produce a clear distinction between /eː/ and /ɛː/ but a merger of the two with /eː/-like formant values. Their productions of /ɛː/ were therefore excluded from the analysis.

[2] Due to the lack of Saterland Frisian elicitations of /aː/ from the Scharrel trilinguals no conclusion can be drawn for the long open category in this variety.

Untersuchung des Kompensationsverhaltens bei Echtzeitmanipulation der Zeitstruktur des auditorischen Feedbacks

Laura Sichlinger & Philip Hoole

Institut für Phonetik, LMU München

laura.sichlinger@campus.lmu.de

Abstract

Auditorisches Feedback ist ein elementarer Bestandteil der Sprachproduktion. Studien mit Echtzeitmanipulation zeigten, dass die Integration der auditorischen Rückmeldung in ein Kontrollsystem für Sprechbewegungen ausschlaggebend für deren Planung ist. Derartige Manipulation resultierte in artikulatorischen Kompensationsmechanismen. Frühere Untersuchungen beschränkten sich allerdings auf spatiale Manipulationen des Sprachsignals. Da die zeitliche Organisation des Sprechens aber eine beträchtliche Auswirkung auf die Produktion von Sprache hat, ist es unumgänglich diese Komponente in Verbindung mit der auditorischen Rückmeldung und Integration von dieser zu untersuchen. Daher wurde ein Experiment durchgeführt, in dem komplexe Stimuli zeitlich perturbiert und die artikulatorische Reaktion der Sprecher analysiert wurde. Es wurde zudem überlegt, dass aufgrund von engeren zeitlichen Beziehungen innerhalb des Onsets einer Silbe im Vergleich zur Coda eindeutigere Anpassungsmechanismen in der artikulatorischen Reaktion des Sprechers bei Onset-manipulierten Sprachsignalen entstehen. Dafür wurden zwei Stimuli erstellt, von denen jeder jeweils einer Gruppe präsentiert wurde. Sowohl Onset- als auch Coda-manipulierte akustische Signale zogen artikulatorische Kompensationsmechanismen nach sich. Außerdem wurden Sprechbewegungen stärker in Reaktion auf Onset-Manipulationen verändert, was darauf hindeutet, dass zeitliche Kontrollmechanismen in der Sprachproduktion der Perzeption unterliegen und dass Lautstrukturen innerhalb des Onsets im Vergleich zur Coda abhängiger von der mentalen Vorhersage der Zeitstruktur eines Lauts sind.

Schlüsselbegriffe: auditorisches Feedback, Produktion, Perzeption, zeitliche Organisation der Sprache

1. Einleitung

Fehlerfreies und flüssiges Sprechen ist ein hochdynamischer Prozess, der die Produktion und Perzeption von Sprache vereint. Eines der einflussreichsten Modelle in der Sprachproduktionsforschung, das DIVA-Modell [1], berücksichtigt diese Verbindung, um die Komponenten der Sprachproduktion zu beschreiben. Das Modell schlägt vor, dass um flüssige Rede zu garantieren, eine enge Kollaboration von zwei Subsystemen, dem Feedforward- und Feedback-System, ausschlaggebend ist. Im Feedback-System werden akustische und somatosensorische Signale des Sprachlauts mit dessen mentaler Repräsentation verglichen. Das heißt, der eigentliche Laut wird der Vorhersage, wie er sich anfühlen oder anhören soll, gegenüber gestellt. Das Ergebnis dieser Information wird daraufhin in das Feedforward-System eingebettet, welches motorische Befehle in echte Bewegungen übersetzt. Variiert also der Laut von seiner mentalen Repräsentation beträchtlich, werden die motorischen Befehle im Feedforward-System angepasst. Eine elementare Variable für diesen Prozess ist das auditorische Feedback; der Sprecher nimmt seinen produzierten Laut wahr und kann daraufhin (durch die Prozesse im Feedforward- und Feedback-System) abwägen ob eine Abänderung der Sprechbewegungen notwendig ist oder nicht. Dieses Modell wird empirisch von aktuellen Studien aus der Sprachproduktionsforschung und Neurowissenschaft unterstützt; Perturbationen des auditorischen Feedbacks während der Sprachproduktion führen zu Abänderungen der motorischen Prozesse, um die Manipulation zu kompensieren und das geplante Ziel zu erreichen [2] [3]. Diese Forschung beschäftigt sich hauptsächlich mit der Manipulation von Formant- und Grundfrequenzen. Allerdings ist die temporale Organisation von Lauten eine elementare Komponente in der Sprachproduktion, um flüssige Rede zu garantieren. Daher ist es wahrscheinlich, dass die Vorhersagen des Sprechers auch temporale Informationen des Lautes enthalten [4]. Es ist also unumgänglich zeitliche Prozesse zu berücksichtigen, um die Verbindung von Perzeption und Produktion und die daraus resultierende flüssige Rede gänzlich zu verstehen. Aus dieser Überlegung ergibt sich die erste von zwei Fragestellungen: Führt die Echtzeitmanipulation der Zeitstruktur des auditorischen Feedbacks einer Äußerung zu einem kompensierenden Verhalten in der artikulatorischen Reaktion des Sprechers?

Der folgende experimentelle Aufbau zur Beantwortung dieser Fragestellung stützt sich auf zuvor genannte Perturbationsexperimente [2] [3]. Um voraussagendes Verhalten in Kombination mit zeitlichen Verarbeitungsprozessen zu untersuchen, wurde das auditorische Feedback des Sprachsignals der Sprecher während der Produktion temporal manipuliert. Vor dem Hintergrund, dass die Echtzeitmanipulation von Frequenzen zu Kompensationsmechanismen führt, ist es wahrscheinlich, dass der Sprecher auch bei zeitlicher Manipulation seine artikulatorischen Bewegungen anpasst, damit das auditorische Feedback mit dem geplanten Ziel des Lautes übereinstimmt. Eine ähnliche Untersuchung findet sich bei Cai et al. [5]; auch hier wird die artikulatorische Reaktion des Sprechers auf Echtzeitmanipulation der Zeitstruktur einer Äußerung analysiert. Das Studiendesign bedient sich einer Stimulusreihe, die ausschließlich aus Vokalen und Halbvokalen besteht, da die Manipulation auf einer zeitlichen Verschiebung des F2-Minimums in der Äußerung basiert. Somit wurde erreicht, dass das auditorische Feedback während des Sprechens entweder beschleunigt oder verzögert wurde. Tatsächlich passten die Probanden ihre Artikulation bei einer auditorischen Verzögerung an; der betroffene Laut wurde gelängt und die darauffolgende Silbe später initiiert. Eine Beschleunigung resultierte in keinen signifikanten artikulatorischen

Veränderungen. Diese Befunde zeigen zwar, dass zeitliche Prozesse während des Sprechens dynamisch abgeändert werden können, geben aber noch wenig Aufschluss darüber, wie genau sie in die Feedforward-Kontrolle eingegliedert werden. Es ist z.B. noch unklar, ob auch eine temporale Echtzeitmanipulation artikulatorische Kompensationsmechanismen nach sich zieht, da der Versuchsaufbau eher eine Reaktion als eine Kompensation der Probanden einforderte. Das folgende Experiment soll daher überprüfen, ob eine direkte zeitliche Streckung bzw. Längung einzelner Bestandteile der Äußerung in einer kompensierenden Reaktion des Sprechers resultiert.
Zudem sollen durch komplexe Stimuli die Unterschiede in den zeitlichen Mechanismen bezüglich der Silbenstruktur untersucht werden. Studien haben gezeigt, dass CV-Transitionen innerhalb des Onsets einer Silbe eine engere zeitliche Beziehung vorweisen, als innerhalb der Coda [6]. Die zweite Fragestellung formuliert sich also folgendermaßen: Falls eine Kompensation in Folge einer Echtzeitmanipulation auftritt, variiert diese bezüglich der Position der Manipulation in der Silbe?
Die Probanden wurden in zwei Gruppen eingeteilt, um diese Frage zu beantworten. Das auditorische Feedback der ersten Gruppe wurde innerhalb des Onsets der Silbe manipuliert, während das der zweiten Gruppe innerhalb der Coda perturbiert wurde. Die temporale Asymmetrie der verschiedenen Bestandteile der Silbe, veranlasst zu der Vermutung, dass Onset- im Vergleich zu Coda-Manipulationen zu eindeutigeren zeitlichen Veränderungen in der artikulatorischen Reaktion führen.

2. Experiment

2.1. Probanden

An dem Experiment nahmen 23 Versuchspersonen teil, 15 davon waren weiblich. Das Durchschnittsalter war 22.3 Jahre und erstreckte sich von 19 bis 25 Jahre. Alle Probanden waren Studenten und wurden erst nach dem Ablauf des Experiments über dessen Zielsetzung aufgeklärt. Vier Teilnehmer wurden bilingual erzogen, gaben aber deutsch als ihre dominante Sprache an, die verbleibenden 19 sind monolingual-deutsche Sprecher. Alle bestätigten, dass keine Hör-, Sprech-, oder Sprachstörung vorliegt.

2.2. Equipment und Echtzeitmanipulation

Das Sprachsignal der Probanden wurde mittels eines Headset-Mikrofons mit niedriger Latenz aufgenommen. Die manipulierte Version des Signals (Abtastrate 16kHz) wurde den Probanden über Insert-Ohrhörer mit einer Latenz von 25 ms präsentiert. Mittels dem Software Paket Audapter [7] wurde die temporale Echtzeitmanipulation auf das auditorische Feedback angewandt. Die Perturbation spezifischer Sprachlaute wird durch eine Funktion namens Online Status Tracking (OST) ermöglicht. Dies verwendet heuristische Regeln (bspw. anhand.von Schwellwerten im Schallpegelverlauf), um die Perturbation an der gewünschten Stelle innerhalb der Äußerung auszulösen.

Neben der OST musste eine sogenannte Perturbation Configuration File (PCF) erstellt werden, um die temporale Manipulation des auditorischen Feedbacks festzulegen.

Beide hier angewandten Dateien, OST und PCF, wurden individuell für jeden Probanden erstellt hinsichtlich der durchschnittlichen Länge ihrer gesprochenen Laute, um die bestmögliche Perturbation zu ermöglichen.

2.3. Experimentelles Design und Stimuli

Es wurden zwei Stimulusreihen erstellt, um die Reaktionen der Probanden auf die Manipulation von verschiedenen Positionen innerhalb der Silbe gegenüberzustellen. In der Phrase „Mehr **Schn**ecken“ wurde das Onset-Cluster der ersten Silbe in „Schnecken“ gelängt und der darauffolgende Vokal gekürzt. In „**Mensch**en auch“ dagegen wurde der erste Vokal in „Menschen“ gelängt und das Coda-Cluster gekürzt. (Die perturbierten Intervalle werden hier fettgedruckt.)

Die Äußerung, also “Mehr Schnecken “oder “Menschen auch”, wurde insgesamt 80 mal wiederholt. Auf eine sogenannten „baseline“-Phase von 10 Wiederholungen, in denen das Signal nicht manipuliert wurde, folgt eine „ramp“-Phase, in der über 50 Stimuli hinweg die Perturbation graduell ansteigt. Das heißt, dass zum Beispiel die Längung des Onset-Clusters von „Mehr Schnecken“ von 0 auf das 1.5-fache erhöht wurde. Die Probanden konnten aber die Veränderungen von einer Wiederholung zur nächsten nicht bewusst wahrnehmen. Folglich sollte ein Lernmechanismus initiiert werden, durch welchen der akustische Output vorhergesagt werden kann, aufgrund der mental gespeicherten Information der temporalen Struktur eines Lautes. Es wurde also erwartet, dass der Proband das perturbierte auditorische Feedback insofern kompensiert, als dass er z.B. ein kürzeres Onset-Cluster und einen längeren Vokal artikuliert im Vergleich zur „baseline“-Phase. Nach der „ramp“-Phase verharren zehn Wiederholungen auf dem maximalen Grad der Perturbation, gefolgt von wiederum zehn nicht-manipulierten Wiederholung in der „aftereffect“-Phase.

2.4. Vorgehensweise

Weil das experimentelle Design ausschließlich die Manipulation von entweder Coda- oder Onset-Cluster pro Session zulässt, wurden die Probanden in zwei Gruppen eingeteilt, um Trainings-Auswirkungen zu neutralisieren. Elf Probanden produzierten die Onset-manipulierte Sequenz „Mehr Schnecken“ und zwölf artikulierten das coda-manipulierte „Menschen auch“. Die Phrasen wurden von einem Computerbildschirm abgelesen.

In einem Follow-up Interview berichteten drei der Sprecher, deren Onset manipuliert wurde, dass sie die Manipulation bemerkt haben. Die restlichen Probanden gaben an, keine Veränderung ihres auditorischen Feedbacks bewusst wahrgenommen zu haben, obwohl das experimentelle Design nach Vollendung des Experiments ausführlich erklärt wurde,

2.5. Ergebnisse

Um zu überprüfen ob und in welcher Form Kompensationsmechanismen in Reaktion auf die temporale Manipulation des auditorischen Feedbacks auftraten, wurden die Dauern des Onsets und medialen Vokals in „Schnecken“ und die des medialen Vokals und /ʃ/ in „Menschen“ hinsichtlich der folgenden drei Phasen verglichen: nicht-perturbierte „baseline“-Phase, maximal-perturbierte „hold“-Phase und nicht-perturbierte „aftereffect“-Phase.

Eine Auswertung der Daten mittels Varianzanalyse (ANOVA) zeigte, dass in der Sequenz „Mehr Schnecken“ die Dauer des Vokals signifikant von der Phase beeinflusst wurde ($F[1.5, 13.3] = 35.5$, $p < 0.001$). Bei näherer Betrachtung fiel

auf, dass nicht nur die „hold“-Phase ($F[1, 9] = 58.1$, $p < 0.001$), sondern auch die „aftereffect“-Phase ($F[1, 9] = 25.7$, $p < 0.001$) einen signifikanten Einfluss hatte. Diese Ergebnisse werden in Abbildung 1 veranschaulicht. Außerdem legte ein Vergleich der Längen der Vokale dar, dass das Schwa durchschnittlich um 42,3 ms länger in der „hold“-Phase und um 16.55 ms länger in der „aftereffect“-Phase war verglichen zur „baseline“-Phase war. Das heißt, dass die akustische Kürzung des Vokals in einem kompensierenden Längen in der artikulatorischen Reaktion resultierte, was zu einem gewissen Grad aufrecht erhalten blieb, als die Manipulation deaktiviert wurde. Diese Befunde weisen also daraufhin, dass kompensierendes Verhalten eintritt, falls das auditorische Feedback manipuliert wird und dass diese Kompensation nach einer Manipulation adaptiert wird. Allerdings hatte die Perturbation des Onsets keinen signifikanten Einfluss auf die Sprachproduktion ($F[1, 9] = 1.83$, $p = 0.21$).

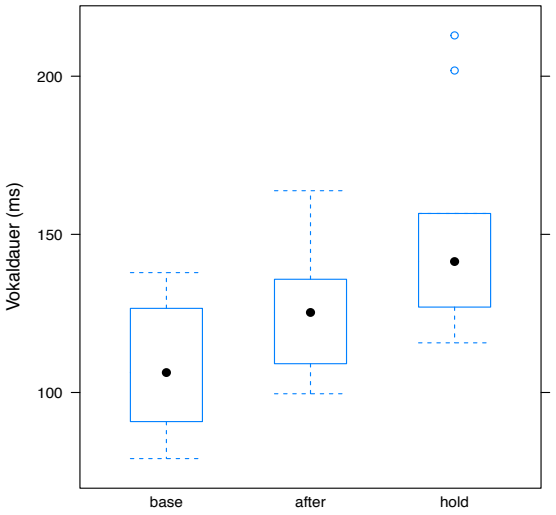


Abbildung 1: Die Verteilung der Dauern des ersten Vokals in „Schnecken“ in Relation zu der nicht-perturbierten „baseline“-Phase, der maximal-perturbierten „hold“-Phase und der nicht-perturbierten „aftereffect“-Phase.

Die Stauchung des /ʃ/ in „Menschen auch“ beeinflusste die artikulatorische Reaktion des Sprechers signifikant ($F[1.3, 9.2] = 10.5$, $p < 0.01$). Präziser gesprochen, führte die maximale Perturbation in der „hold“-Phase zu einer signifikanten Veränderung ($F[1, 7] = 59.1$, $p < 0.001$) in der Produktion des Konsonanten verglichen zur nicht-manipulierten „baseline“-Phase. Das /ʃ/ war durchschnittlich 16.6 ms länger in der „hold“- als in der „baseline“-Phase. Allerdings konnte in diesem Fall kein signifikanter „aftereffect“ ($F[1, 7] = 0.66$, $p = 0.44$) ermittelt werden. Das heißt, dass die Probanden die Kürzung des /ʃ/ offensichtlich kompensierten, indem sie es längten, aber dieses kompensierende Verhalten nach der Manipulation nicht beibehielten. Diese Befunde werden in Abbildung 2 illustriert. Die Längung des Vokals hatte keinen signifikanten Einfluss ($F[1.2, 8.7] = 0.08$, $p = 0.92$) auf die Artikulation des Sprechers.

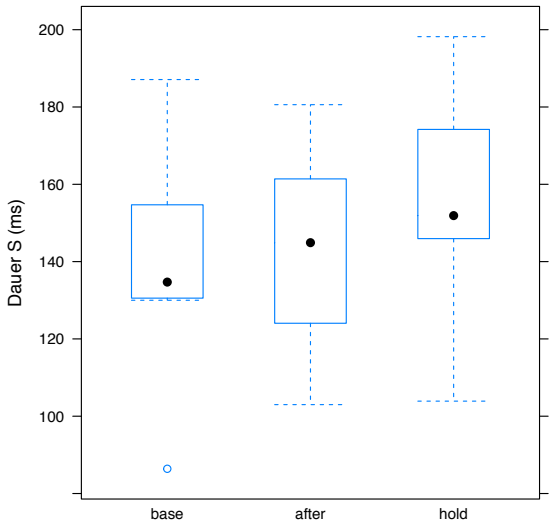


Abbildung 2: Die Verteilung der Dauern des /ʃ/ in „Menschen“ in Relation zu der nicht-perturbierten „baseline“-Phase, der maximal-perturbierten „hold“-Phase und der nicht-perturbierten „aftereffect“-Phase.

3. Diskussion

Beide Hypothesen werden von den Ergebnissen des Experiments unterstützt. Zum einen führte eine temporale Manipulation des auditorischen Feedbacks zu kompensierenden Mechanismen in der artikulatorischen Reaktion. Der Sprecher reagierte entgegengesetzt der Manipulation. Das heißt, dass eine akustische Kürzung des Sprachlauts in einer artikulatorischen Längung resultierte. Diese Ergebnisse unterstützen die Annahme aus früheren Studien, dass temporale Mechanismen in der Artikulation während des Redeflusses adjustiert werden können [5]. Eine neue Erkenntnis ist, dass diese zeitlichen Veränderungen nicht nur auf Äußerungen mit einfachen Lautstrukturen zutreffen, sondern auch auf Artikulationen, die komplexe Strukturen enthalten. Zudem sind die Ergebnisse mit dem DIVA Modell zu vereinbaren und schlagen zusätzlich eine temporale Komponente vor: Augenscheinlich wurde das zeitlich manipulierte auditorische Feedback in das Feedforward-System integriert, was zu einer Anpassung der Artikulatoren führte, so dass die mentale Vorhersage des Sprachlauts mit dem auditorischen Feedback übereinstimmte. Interessanterweise, spiegelt sich in den Daten ein asymmetrisches Muster wider: Während nämlich die Kürzung eines Lauts in einer signifikanten artikulatorischen Längung resultierte, beeinflusste die Längung eines Lautes die Reaktion des Sprechers nicht signifikant.

Zum anderen zogen Perturbationen des Onsets eindeutigere zeitliche Kompensationsmechanismen nach sich als die der Coda. Obwohl in beiden Fällen die Manipulation signifikante artikulatorische Anpassungen nach sich zog, wurden Onset-manipulierte Äußerungen in der Artikulation mehr kompensiert. Dies gibt Aufschluss darüber, inwiefern die mentale Vorhersage von zeitlichen Strukturen im Redefluss mit der Silbenstruktur korreliert; mentale Vorhersagen, die zeitliche Strukturen betreffen scheinen einen größeren Einfluss auf CV-Transitionen innerhalb des Onsets im Vergleich zur

Coda zu haben. Diese Erkenntnis suggeriert, dass die temporalen Beziehungen innerhalb des Onsets einer Silbe eine engere Verbindung eingehen als die innerhalb der Coda.

Allgemein gesprochen demonstrieren diese Ergebnisse die Verbindung von Sprachproduktion und Perzeption in Hinblick auf die temporale Kontrolle und Organisation der Sprechbewegungen. Mittels des auditorischen Feedbacks werden artikulatorische Prozesse angepasst, um geplante auditorische Ziele bestmöglich zu erfüllen. Außerdem pflegen mentale Vorhersagen der zeitlichen Organisation von Sprache eine enge Verbindung zu der Silbenstruktur. Diese Erkenntnis tragen zu unserem Verständnis bei, wie komplexe Sprachlaute produziert und im Akt des Sprechens organisiert werden. Dennoch ist noch viel Forschung zu leisten, um genau zu verstehen wie das Integrieren des auditorischen Feedbacks in das Feedforward-Kontrollsystem zum flüssigen Sprechen beiträgt. In einem nächsten Schritt soll die Reaktion des Sprechers auf feinere und spezifischere Perturbationen der Stimuli analysiert werden; zum Beispiel auf das Längen des C1 und das Kürzen des C2 im Onset von „Schnecken“. Dadurch soll untersucht werden, ob komplexere Zeitstrukturen abhängiger von mentalen Vorhersagen und ob daher eine Manipulation in beträchtlicheren zeitlichen Kompensationsmechanismen resultiert.

4. Bibliographie

[1] J. A. Tourville and F. H. Guenther, “The DIVA model: A neural theory of speech acquistion and production,“ *Lang Cogn Process* vol. 25, pp. 952-981, 2011.

[2] K. G. Munhall, E. N. MacDonald, S. K. Byrne and I. Johnsrude, “Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate,” *J. Acoust. Soc. Am* vol. 125, pp. 384-390, 2008.

[3] J. A. Jones and K. G. Munhall, “Perceptual calibration of F0 production: Evidence from feedback perturbation,” *J. Acoust. Soc. Am* vol. 108, pp. 1246–1251, 2000.

[4] J. Debrabant, F. Gheysen, G. Vingerhoets and H. van Waelvelde, “Age-related differences in predictive response timing in children: evidence from regularly relative to irregularly paced reaction time performance,” *Hum Mov Sci*, vol. 31, no. 4, pp. 801-810, 2012.

[5] S. Cai, S. S. Ghosh, F. H. Guenther and J. Perkell, “Focal Manipulations of Formant Trajectories Reveal a Role of Auditory Feedback in the Online Control of Both Within-Syllable and Between-Syllable Speech Timing,” *J. Neurosci.*, vol. 31, no. 45, pp. 16483-16490, 2011.

[6] P. Hoole and M. Pouplier, “Interarticulatory coordination – Speech Sounds.“ in *The Handbook of Speech Production,* pp. 133 -157, edited by M. Redford, Malden, MA: Wiley Blackwell, 2015.

[7] J. A. Tourville, S. Cai and F. H. Guenther, “Exploring auditory-motor interactions in normal and disordered speech,” *Proceedings of Meeting on Acoustics,* vol. 19, 060180, 2013.

Prosodic variation in conceptual distance and proximity: Self-repairs in French

Johanna Stahnke

Bergische Universität Wuppertal, Germany
stahnke@uni-wuppertal.de

Abstract

The present study addresses the prosodic contextualization of self-repair in conceptual distance and conceptual proximity in spoken French. According to the literature, paraphrases establish semantic equivalence and are intonationally deaccented; corrections constitute semantic difference and show overaccentuation. Repairs are expected to be more frequent in spontaneous proximity than in pre-planned distance. Apart from quantities, it is hypothesized that self-repair is subject to qualitative differences: As a consequence of contextual involvement in conceptual proximity, speakers should more extensively use prosodic contextualization cues than in context-detached distance. Rather than being contextualized for repair type, however, the results of an empirical analysis of two conceptually contrastive corpora suggest a general tendency of deaccentuation in proximity and of overaccentuation in distance, independently of the repair type. The interpretation of these unexpected findings is proposed in a communicative model of language variation and change, in which speakers in conceptual proximity strategically routinize deaccented structures in order to cope with conversationally undesired disruptive corrections. Possibilities for prosodic change are discussed in the light of lexicalization of discourse markers induced by conceptual proximity.

Index Terms: prosody, conceptual variation, self-repair, French

1. Introduction

Prosody, especially intonation, represents an essential contextualization cue of linguistic repair in various languages which use fundamental frequency (f0) as a means to express pragmatic purposes, as has been shown for English (e.g. [1] and [2]) and German (e.g. [3] and [4]), while research of prosodic contextualization cues is sparse for French. Among others, self-repairs include paraphrases and corrections (cf. [5]) and obligatorily consist of a repairable, i.e. a problematic element which needs to be repaired (boldface), and a repair which actually finalizes the paraphrase or the correction (underlined):

(1) Paraphrase
*La croissance et **la place de la France**, cinquième puissance économique du monde, elle passe notamment par notre capacité à nous appuyer sur les acteurs français qui sont aujourd'hui à l'étranger*

'France's [economic] growth and position, fifth economic power in the world, it [the economic growth] is notably due to our capacity to rely on the French who are currently abroad'

(2) Correction
***La résistance**, si je peux dire, la durée de Monsieur Bachar al-Assad était plus longue qu'anticipée*

'The resistance, if I may say, the duration of Mister Bashar al-Assad was longer than anticipated'

While in paraphrases the relation between repairable and repair is semantic equivalence, semantic difference is established in corrections: In example (1), the paraphrastic repair (*cinquième puissance économique du monde*) is a precision of the repairable (*la place de la France*) on a syntagmatic axis with both components being referentially identical; in example (2), by contrast, the correcting repair (*la durée*) is a replacement of the repairable (*la résistance*) with both components excluding each other in paradigmatic opposition.

On a continuum of semantic contrast between repairable and repair, paraphrases are minimally contrastive and corrections maximally contrastive. With respect to continuous discourse flow as one of the basic principles of conversation (cf. [6]), corrections are therefore more disruptive than paraphrases.

A second problem which is related to the issue of discourse coherence in conversation is turn-taking, especially when a high degree of dialogicity prevails (cf. [6], [7]). Despite the preference for self-repair over other-repair (cf. [8]), relatively severe discourse disruptions caused by corrections present an appropriate opportunity for turn-taking in comparison with minor disruptions produced by paraphrases.

As for the occurrences of self-repair in language use, the variation between conceptual distance and proximity offers a systematic account. Repair is a notorious feature of proximity which is characterized by a high degree of dialogicity, situational involvement, cooperation and spontaneity. These extra-linguistic communicative conditions give rise to increased chances of linguistic repair to occur (cf. [9]).

2. Prosodic structure of self-repair

French intonation substantially differs from English, German or other Romance languages' systems in that accentuation is not lexical but phrasal. Within the Autosegmental-Metrical framework of intonation (cf. [10]), a pitch accent H* is assigned to the final full syllable of an Accentual Phrase (AP), and an optional initial accent Hi (i: *initial*) marks the left edge of that phrase in French. One or several APs form an Intonational Phrase (IP; cf. [11]). Despite the differences in prosodic phrasing in French and other languages, the global structures of repair types resemble each other when holistic units are considered from a systemic-tonetic approach, which

is typically used in research on conversational structures in English and German.

The functional differences of the repair types are reflected in their prosodic realizations. For paraphrases, the intonational structure has been shown to be flat and deaccented compared to the surrounding discourse (cf. e.g. [12], [13], [14]). Deaccentuation is not to be understood as a complete deletion of f0 variation, but rather as a minimization of tonal contrasts with a downstep of final accents and a reduction or elision of initial accents (cf. [15]). Figure 1 illustrates the intonational contour of the paraphrase *cinquième puissance économique du monde* (example 1), in which successive downstep of final accents as well as standard accentuation of the preceding repairable (*la place de la France*) and the following discourse (*elle passe notamment*) can be observed:

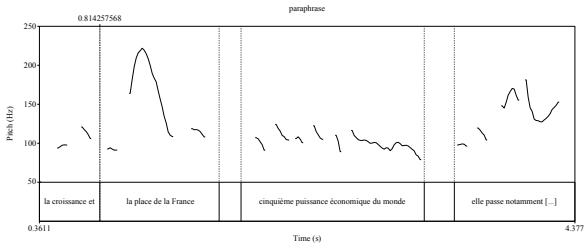


Figure 1: *Prosodic structure of a paraphrase.*

Conversely, correcting repair is prosodically overaccented (cf. e.g. [13]) and possibly followed by a deaccented structure until completion of the prosodic unit. Figure 2 displays the correction *la durée* (example 2) with two overaccentuations on each syllable of the lexical item (*du-*, *-rée*). The remaining lexical material of the IP (*de Monsieur Bachar al-Assad*) is deaccented, and the preceding repairable (*la résistance*) as well as the metalinguistic repair marker (*si je peux dire*) exemplify standard accentuation:

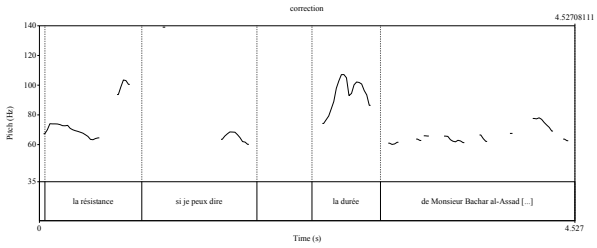


Figure 1: *Prosodic structure of a correction.*

The correcting accent is comparable to the focal accent Hf (f: *focus*) identified by [11] for early focus expressions in French and is considered as a sub-category of (correcting) focus here.

3. Methodology

Since in conceptual proximity various kinds of (linguistic, para- and extra-linguistic) contexts are accessible while conceptual distance is largely decontextualized (cf. [9]), the hypothesis of the present study is as follows:

Hypothesis: For the contextualization of self-repair, speakers in conceptual proximity more extensively use prosodic contextualization cues than in conceptual distance by

a) deaccented structures in paraphrases,
b) overaccented structures in corrections.

Studies on repairing contextualization implicitly use data of spontaneous conceptual proximity without, however, systematically comparing them to conceptual distance. In order to fill this research gap, two different corpora were established – political interviews as instances of conceptual distance and private conversations as instances of conceptual proximity. The interviews were taken from the daily broadcast *L'invité du matin* of the radio station *Radio France*; the conversations were recorded between family members or friends in private and familiar surroundings. The proximity speakers were asked to talk about a topic of their choice as freely and spontaneously as possible. All speakers agreed upon their conversations being recorded and analyzed for linguistic purposes. Speakers were not informed about the research object until after the recordings. The duration of each interaction is about ten minutes and the conversational structure dialogical, thus controlling for extra-linguistic factors and ensuring comparability between the proximity and distance corpora. The entire corpus is made up of 35 subjects and consists of around 55,200 words and 3.5 hours of conversational material, respectively.

A total of 379 self-repairs was extracted from the corpus, of which 300 are paraphrases and 79 corrections; 206 are distance repairs and 173 proximity repairs. All repairs were prosodically transcribed with IVTS (*Intonational Variation Transcription System*, cf. [16]) and coded for prosodic form. Criteria for the categorization of accentuation variants include global measures, e.g. the relative pitch range of the repair compared to that of the repairable, as well as local measures, e.g. the overall pitch movement and the prosodic phrasing of the repair.

The data were analyzed in a multivariate approach using *Varbrul* (cf. [17]) with the accentuation variants (deaccentuation, overaccentuation, standard accentuation) representing the dependent variables. The independent variables are constituted by the functional repair types (paraphrases, corrections) and by conceptual variation (distance, proximity) as factor groups conditioning the outcome of the linguistic variables.

4. Results

Table 1 summarizes the distribution of the prosodic variants of paraphrases in conceptual distance and proximity (with absolute frequencies in parentheses):

Table 1. *Prosodic structure of paraphrases in conceptual distance and proximity.*

Paraphrases	Conceptual distance	Conceptual proximity
Standard accentuation	51% (93)	47% (55)
Deaccentuation	24% (45)	49% (56)
Overaccentuation	25% (46)	4% (5)

In both corpora around half of all paraphrases are contextualized by standard accentuation (51% in conceptual distance, 47% in conceptual proximity). In conceptual proximity, the remaining paraphrases are – as expected –

generally deaccented (49%), and only a minor proportion is overaccented (4%). In contrast, de- and overaccentuation in conceptual distance are evenly distributed (24% and 25%, respectively).

Deaccentuation is influenced by the conceptual factor group at a statistically significant level (p<.001), with proximity as the favoring factor (factor weight: 0.654). The functional factor group does not have a significant effect on deaccentuation (p=.229).

The results of the correcting variants are shown in Table (2):

Table 2. *Prosodic structure of corrections in conceptual distance and proximity.*

Corrections	Conceptual distance	Conceptual proximity
Standard accentuation	18% (4)	42% (24)
Deaccentuation	5% (1)	35% (20)
Overaccentuation	77% (17)	23% (13)

The distance corpus broadly confirms the findings in the literature: The majority of corrections is overaccented (77%), while few cases show standard (18%) or deaccentuation (5%; n=1). In conceptual proximity, however, the expected overaccented variant represents the smallest proportion (23%), while the remaining corrections are either standardly accented (42%) or even deaccented (35%).

The overaccented variant is significantly influenced by conceptual and functional variation (both p<.001). Within factor groups, overaccentuation is favored by conceptual distance (factor weight: 0.730) and by corrections (factor weight: 0.842). In statistically comparing the factor groups by relative strength, the conceptual variation (range: 496) outweighs the functional variation (range: 450).

5. Discussion and conclusions

The hypothesized prosodic contextualization of self-repairs in conceptual proximity can only be maintained for paraphrases (Hypothesis a), but not for corrections (Hypothesis b). Quite on the contrary, the contextualization of corrections does not only falsify the hypothesis, but appears to be prosodically paraphrastic due to a significant number of deaccentuations (35%). On the other hand, an important proportion of paraphrases in conceptual distance is overaccented (25%). Instead of a repair-type-specific contextualization in proximity, the results point to a conception-specific preference of deaccentuation in proximity and overaccentuation in distance. The statistical analyses corroborate this generalization: In paraphrases, the functional factor group is not significant and in corrections, where it does play a role, the conceptual factor group shows to have the higher relative strength. Moreover, deaccentuation is favored by conceptual proximity, whereas overaccentuation is favored by conceptual distance.

These results can be accounted for in a model of routinization. Routines are speaker-strategic solutions to frequently occurring problems in discourse (cf. [18]). As we have seen, corrections present a serious problem in communication as they entail semantic incompatibility and therefore are more disturbing than paraphrases with regard to continuous discourse flow and turn-maintaining. These goals are specifically related to conceptual proximity where dialogicity and turn competition are massively underway (cf. [6]). Speakers possibly develop linguistic techniques to handle undesired corrections and encode them as conversationally less problematic paraphrases by means of prosodic deaccentuation. In doing so, they violate conversational maxims (cf. [19]), but secure discourse coherence and the right to speak.

A similar phenomenon has been described by [2] for the contextualization of interactive repair in English. The study convincingly argues that acoustic repairs can be masked as content-related repairs and vice versa for conversational reasons like face-saving by manipulating the prosodic parameters of speech rhythm and rate (‘prosodic camouflage’). Even though systematic studies addressing prosodic routinization of repairing cues are missing with respect to German, it is highly likely to assume that it functions in a very similar fashion as English and French since the communicative conditions of proximity as well as general conversational motives can be regarded as universal (cf. [9]).

Turning to lexical elements, research on metalinguistic discourse markers in various languages also suggests routinization by flouting the cooperative principle due to pressure of dialogicity, as [6] illustrate on the basis of the illegitimate use of imperatives such as Italian *guarda* (‘look’) for turn-taking purposes, which can also be observed in a number of other languages, cf. English *look* and German *guck/schau (mal)*. Expressivity as another typical condition of proximity (cf. [20]) may also trigger routinization. The metaphorical use of the temporal meaning of the adverb *enfin* ‘finally’ for the reformulative function of the discourse marker *(en)fin* is a case in point (cf. [21], [22]).

When a routine is taken over into the historical level of a linguistic system, variation may lead to change, as has been attested for the lexicalization of discourse markers. Although the topic has recently gained importance in Romance phonology (cf. e.g. [23], [24], [25], [26]), the diachronic development of prosody is largely underrepresented. As far as the prosodic variation of proximity corrections is concerned, the data provide evidence of a possible condition for prosodic change in French.

Considering the fact that there is interaction between factor groups, the high proportions of standard accentuation as the medium form between de- and overaccentuation may be explained by a conflict between the correcting (overaccented) structure favored by conceptual distance and the paraphrastic (deaccented) structure favored by conceptual proximity. Secondly, prosody is gradual and continuous, so that the trichotomy of standard, de- and overaccentuation is a methodologically necessary simplification which hardly represents the complex facts in reality.

Finally, overaccented distance paraphrases strongly challenge the traditional view of the contextualization of repairs. Most obviously, overaccentuation in conceptual distance can be attributed to a generalization of the stable *accent d’emphase* ‘emphatic accent’ of French distance speakers (cf. e.g. [27]). This type of accent is used throughout the entire distance corpus, i.e. also outside of repairs. Deaccentuation in conceptual proximity, though, is only observed in specific contexts like repair which indirectly supports this interpretation.

6. References

[1] W. J. M. Levelt and A. Cutler, „Prosodic marking in speech repair“, *Journal of Semantics* 2, pp. 205-218, 1983.

[2] E. Couper-Kuhlen, „Contextualizing discourse: The prosody of interactive repair“, *KontRI – Kontextualisierung durch Rhythmus und Intonation* 9, pp. 1-35, 1989.
[3] M. Selting, „The role of intonation in the organization of repair and problem handling sequences in conversation”, *Journal of Pragmatics* 12, pp. 293-322, 1988.
[4] P. Bergmann, „The prosodic design of parentheses in spontaneous speech“, P. Bergmann, J. Brenning, M. Pfeiffer, and E. Reber (eds.), *Prosody and embodiment in interactional grammar*, Berlin: de Gruyter, pp. 103-141, 2012.
[5] E. Gülich and T. Kotschi, „Textherstellungsverfahren in mündlicher Kommunikation. Ein Beitrag am Beispiel des Französischen”, W. Motsch (ed.), *Ebenen der Textstruktur*, Tübingen: Niemeyer, pp. 37-80, 1996.
[6] U. Detges and R. Waltereit, „Turn-taking as a trigger for language change“, S. Dessì Schmid, U. Detges, P. Gévaudan, W. Mihatsch, and R. Waltereit (eds.), *Rahmen des Sprechens. Beiträge zu Valenztheorie, Varietätenlinguistik, Kreolistik, Kognitiver und Historischer Semantik. Peter Koch zum 60. Geburtstag*, Tübingen: Narr, pp. 175-189, 2011.
[7] H. Sacks, E. A. Schegloff, and G. Jefferson, „A simplest systematics for the organization of turn-taking for conversation”, *Language* 50, pp. 696-735, 1974.
[8] E. A. Schegloff, G. Jefferson, and H. Sacks, „The preference for selfcorrection in the organization of repair in conversation”, *Language* 53, pp. 361-382, 1977.
[9] P. Koch and W. Oesterreicher, „Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte“, *Romanistisches Jahrbuch* 36, pp. 15-43, 1985.
[10] J. B. Pierrehumbert, *The phonology and phonetics of English intonation*, PhD thesis, http://faculty.wcas.northwestern.edu/~jbp/publications/Pierrehumbert_PhD.pdf, 1980.
[11] S.-A. Jun and C. Fougeron, „A phonological model of French intonation”, A. Botinis (ed.), *Intonation: Analysis, Modeling and Technology,* Dordrecht: Kluwer, pp. 209-242, 2000.
[12] P. Delattre, „Les dix intonations de base du français”, *French Review* 40, pp. 1-14, 1966.
[13] M.-A. Morel and L. Danon-Boileau, *Grammaire de l'intonation*, Gap: Ophrys, 1998.
[14] A.-C. Simon, *La structuration prosodique du discours en français*, Bern: Lang, 2004.
[15] A. Cristo and L. Jankowski, „Prosodic organisation and phrasing after focus in French”, *Proceedings of the XIVth International Congress of Phonetic Sciences*, *San Francisco*, vol. 2, pp. 1565-1568, 1999.
[16] B. Post, E. Delais-Roussarie, and A.-C. Simon, „IVTS, un système de transcription pour la variation prosodique”, J. Durand, B. Laks, and C. Lyche (eds.), *Bulletin PFC* 6, pp. 51-68, 2006.
[17] S. A. Tagliamonte, *Analysing sociolinguistic variation*, Cambridge: Cambridge University Press, 2006.
[18] U. Detges and R. Waltereit, „Grammaticalization and pragmaticalization“, S. Fischer and C. Gabriel (eds.), *Manual of Grammatical Interfaces in Romance*, Berlin: de Gruyter, to appear.
[19] P. H. Grice, „Logic and conversation”, P. Cole and J. L. Morgan (eds.), *Syntax and semantics. Speech acts*, vol. 3, New York: Academic Press, pp. 41-58, 1975.
[20] P. Koch and W. Oesterreicher, „Sprachwandel und expressive Mündlichkeit“, *Zeitschrift für Literaturwissenschaft und Linguistik* 102, pp. 64-96, 1996.
[21] M.-B. M. Hansen, *The function of discourse particles. A study with special reference to spoken standard French*, Amsterdam: Benjamins, 1998.
[22] M.-B. M. Hansen, „From prepositional phrase to hesitation marker: the semantic and pragmatic evolution of French *enfin*”, *Journal of Historical Pragmatics* 6, pp. 37-68, 2005.
[23] J. I. Hualde, „Remarks on the diachronic reconstruction of intonational patterns in Romance with special attention to Occitan as a bridge language”, *Catalan Journal of Linguistics* 2, pp. 181-205, 2003.
[24] J. I. Hualde, „Romance intonation from a comparative and diachronic perspective. Possibilities and limitations”, J. Auger, J. C. Clements, and B. Vance (eds.), *Contemporary Approaches to Romance Linguistics*, Amsterdam: Benjamins, pp. 217-237, 2004.
[25] A. Pešková, I. Feldhausen, E. Kireva, and C. Gabriel, „Diachronic prosody of a contact variety: Analyzing Porteño Spanish spontaneous speech”, K. Braunmüller and C. Gabriel (eds.), *Multilingual individuals and multilingual societies*, Amsterdam: Benjamins, pp. 365-389, 2012.
[26] C. Gabriel, „Emphase, Sprachkontakt und prosodischer Wandel: Überlegungen zum tritonalen Tonhöhenakzent des *Porteño*-Spanischen“, E. Pustka and S. Goldschmitt (eds.), *Emotionen, Expressivität, Emphase*, Berlin: Schmidt, pp. 197-214, 2014.
[27] F. Carton, D. Hirst, A. Marchal, and A. Séguinot, *L'accent d'insistance*, Montréal: Didier, 1976.

Vowel confusions in noise by German listeners: A study of oral and nasalized vowels

Kim Strütjen, Ruben van de Vijver

Institut für Sprache und Information, Heinrich-Heine-Universität Düsseldorf, Deutschland

Kim.Struetjen@hhu.de, Ruben.Vijver@hhu.de

Abstract

Low oral and nasalized vowels are acoustically more similar than non-low oral and nasalized vowels [1], [2]. We hypothesize that low oral and nasalized vowels are more likely to be confused with each other than non-low oral and nasalized vowels, even in languages without contrastive nasalization or allophonic nasalization, such as German [3, pp. 12-13, 32].

In a forced-choice identification task 28 native German adults listened to vowels in the presence of a masking noise and were forced to identify the vowels as one of these six vowels: [a], [ã], [ɛ], [ɛ̃], [i] or [ĩ]. The results confirm the hypothesis and show that native Germans confuse [a] more often with [ã] than [ɛ] with [ɛ̃] or [i] with [ĩ]. These results are in line with previous studies indicating the tendency that high nasalized vowels are easily distinguished from high oral vowels by native speakers of American English [4], [5]. However, native American adults are familiar with allophonically nasalized vowels [6] whereas our participants are not. Our study, therefore, shows that the acoustic modification due to the nasalization is perceived better in non-low vowels than in low vowels even by speakers who are not familiar with vowel nasalization.

Index Terms: vowel nasalization, phonetics, phonology, perception, German

1. Introduction

Previous research on the perception of vowel nasalization has primarily focused on the perceptual correlates of nasalization [7], [8], [9] and the perception of nasalized vowels by speakers who are familiar with vowel nasalization [10], [11], [12]. Native speakers of American English perceive nasalization of vowels varying in vowel height differently. Bond's [4] results show that high and mid nasalized vowels were identified better than low nasalized vowels by native speakers of American English. This is in line with the results of House and Stevens [5] who studied the perception of nasalization with American native speakers using synthesized stimuli. According to them low vowels require a greater velum lowering in comparison to high vowels before they are identified as nasalized.

A possible explanation for the asymmetric perception of vowel nasalization is provided by the acoustic modifications due to the nasalization of vowels. When the velum is lowered to produce nasalized vowels, the oral and the nasal cavity resonate together, which has acoustic consequences [1]: In comparison to the oral spectrum the spectrum of a nasalized vowel has decreased amplitude and additional nasal resonances, which originate in the resonances of the nasal cavity [1]. The most important differences are found in the vicinity of the first formant which has a smaller amplitude and a larger bandwidth in comparison to the oral first formant [1], [5].

These spectral effects have consequences for the perception of nasalized vowels. High and low vowels differ in the degree of acoustic modifications due to the nasalization [1]. In high vowels there is a large difference between the first and the second formant so that the modifications of the first formant cause greater distortions than in the low vowels [2]. In low vowels the first and the second formant are close to each other so that the spectrum becomes only weaker due to the nasalization [2]. That means that the acoustic modifications due to the nasalization seem to be more noticeable in high vowels than in low vowels [1].

If the reason for the different degree of perceived nasalization in high and low vowels is in the acoustics, we expect that speakers who are not familiar with nasalized vowels should show the same asymmetric pattern, because the physics and physiology that cause the effects are the same for all human beings. We therefore hypothesize that low oral and nasalized vowels are more likely to be confused with each other than non-low oral and nasalized vowels by speakers whose native language does neither have contrastive nor allophonic nasalization. To investigate this we conducted a vowel identification experiment with native speakers of German. Nasalized vowels are not part of the German phoneme inventory and allophonic nasalization in e.g. French loan words is not common [3, pp. 12-13, 32]. Our participants live in the Rhineland where nasal vowels in French loan words are pronounced as an oral vowel followed by [ŋ]: The French loan word *Croissant* is thus pronounced as [krosɔŋ] by people from the Rhineland (sometimes even written as *Crossong*). Notwithstanding the absence of lexical or allophonic nasalization of vowels, according to our hypothesis native speakers of German should confuse oral and nasalized low vowels more often with each other than oral and nasalized non-low vowels.

2. Experiment

To test the hypothesis we tested native speakers of German on their ability to identify oral and nasalized vowels. The vowels differed in vowel height and were presented in noise.

2.1. Stimuli

We used the three oral vowels [a], [ɛ] and [i] and their three corresponding nasalized counterparts [ã], [ɛ̃] and [ĩ] as stimuli. The oral vowels were cut out of CV-syllables whereas the nasalized vowels were cut out of CV[m]-syllables. Three oral and three nasalized vowels preceding [l] served as fillers.

The syllables were recorded by a bilingual native speaker of Portuguese (and German) because all these vowels are part of the Portuguese phoneme inventory [13]. Recording took place in an anechoic booth in the phonetics laboratory of the Heinrich-Heine-University Düsseldorf. The sampling rate was 48 KHz. Their intensity was scaled to 70 dB using Praat [14]. The vowels were presented with noise at a signal-to-noise ratio of 20 dB.

2.2. Procedure

The participants were tested with a forced-choice identification experiment using Praat [14]. The experiment took place in an anechoic booth in the phonetics laboratory of the Heinrich-Heine-University Düsseldorf.

Participants listened via headphones to vowels overlaid with a masking noise and were forced to identify each vowel as one of the following six vowels: [a], [ã], [ɛ], [ɛ̃], [i] or [ĩ]. After they made their choice the next vowel was presented. Responses were given by pressing on one of the six vowels on the screen. The stimuli were presented in random order with 6 repetitions so that every participant had to made 72 decisions ((6 experimental items + 6 fillers) x 6 repetitions).

At the beginning of the experiment there was a short introductory phase which aimed at familiarizing the participants with the experiment and its setting. In this introductory phase participants listened to the four vowels [ɔ], [ɔ̃], [u] and [ũ] while seeing on the screen the transcription of the vowels used in the experiment. Doing so, they learned that an oral vowel is represented by an orthographic symbol of this vowel, e.g. [ɔ] was transcribed as <o>. The nasalized vowels were transcribed with a tilde above the corresponding symbols of the oral vowels, e.g. <õ>. In the subsequent test phase the same transcriptions with or without tilde were used. [a] was transcribed as <a>, [ɛ] as <e>, [i] as <i>, [ã] as <ã>, [ɛ̃] as <ẽ> and [ĩ] as <ĩ>. The experiment took approximately 10 minutes.

2.3. Participants

We tested 28 adult native speakers of German (21 women, 7 men, mean age: 20.89 years, Range: 19-23). None of them had any knowledge of a language which makes use of nasalized vowels, e.g. French, Portuguese or Polish, as ascertained by a questionnaire. All participants were undergraduate students of Linguistics at Heinrich-Heine-University Düsseldorf and received course credit for their participation. They had normal or corrected vision and no hearing problems.

2.4. Results

A confusion matrix (see Table 1) was created with which we measured perceptual similarity and perceptual distance for the oral-nasalized vowel pairs based on the formula proposed by Shepard [15].

According to Shepard [15] the similarity S_{ij} between the two sounds *i* and *j* is measured by summing the proportions of confused responses of the sound pair (proportion of *i* recognized as *j* plus proportion of *j* recognized as *i*) and dividing this number by the sum of the proportions of the correct responses to the two sounds (proportion of *i* recognized as *i* plus proportion of *j* recognized as *j*) (see Formula 1).

	[ã]	[a]	[ɛ̃]	[ɛ]	[ĩ]	[i]
<ã>	14 (0.08)	17 (0.10)	17 (0.10)	8 (0.05)	6 (0.04)	5 (0.03)
<a>	86 (0.51)	84 (0.50)	11 (0.07)	8 (0.05)	10 (0.06)	3 (0.02)
<ẽ>	9 (0.05)	14 (0.08)	58 (0.35)	37 (0.22)	23 (0.15)	6 (0.40)
<e>	16 (0.10)	16 (0.10)	41 (0.24)	69 (0.41)	50 (0.30)	24 (0.14)
<ĩ>	18 (0.11)	13 (0.08)	22 (0.13)	18 (0.11)	18 (0.11)	21 (0.13)
<i>	25 (0.15)	24 (0.14)	19 (0.11)	28 (0.17)	61 (0.36)	109 (0.65)

Table 1: *Confusion matrix of oral and nasalized vowels in German. Each column in the matrix corresponds to one of the auditory stimuli and each row corresponds to one of the available responses. Proportions are given in brackets.*

$$S_{ij} = \frac{p_{ij}+p_{ji}}{p_{ii}+p_{jj}} \quad (1)$$

With the following Formula 2 the perceptual similarity S_{ii} between [i] and [ĩ] is calculated.

$$S_{i\tilde{i}} = \frac{0.13+0.36}{0.65+0.11} = 0.64 \quad (2)$$

The higher the perceptual similarity number, the more similar the sounds are. The perceptual distance is calculated by taking the negative log of the perceptual similarity (see Formula 3).

$$d_{ij} = -\ln(S_{ij}) \quad (3)$$

The results for our vowel pairs are summed up in Table 2.

vowel pairs	number and proportion of confusions	perceptual similarity	perceptual distance
[i] & [ĩ]	82 (0.24)	0.64	0.446
[ɛ] & [ɛ̃]	78 (0.23)	0.61	0.494
[a] & [ã]	103 (0.31)	1.05	-0.049

Table 2: *Confusion values, similarities and distance among oral and nasalized vowel pairs in German, based on the confusion matrix in Table 1.*

The comparison of the values for the different vowel pairs shows that oral and nasalized [i] and oral and nasalized [ɛ] are less often confused with each other than oral and nasalized [a]. As a result, the oral and nasalized non-low vowels are perceptual less similar than oral and nasalized low vowels.

Their perceptual distance is illustrated in the perceptual map in Figure 1. This perceptual map is determined by multidimensional scaling (MDS) and is based on a hierarchical cluster analysis using maximum distance measure in R [16].

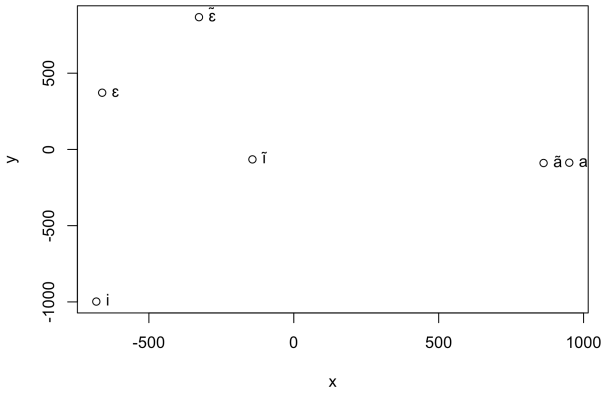


Figure 1: *Perceptual map of oral and nasalized vowels in German.*

Concerning the oral-nasalized vowel pairs the perceptual map shows that [a] and [ã] are perceptually more close to each other than [i] and [ĩ] and [ɛ] and [ɛ̃] respectively.

3. Discussion

The results confirm our hypothesis that low oral and nasalized vowels are more likely to be confused with each other than non-low oral and nasalized vowels. They show that native speakers of German confuse [a] more often with [ã] than [ɛ] with [ɛ̃] or [i] with [ĩ]. These results are in line with previous studies indicating the tendency that high nasalized vowels are easily distinguished from high oral vowels by native speakers of American English [4], [5].

The reason for this asymmetry concerning the perception of nasalization of vowels differing in vowel height is provided by the phonetic nature of nasalization. Non-low oral and nasalized vowels are acoustically more different from each other than low oral and nasalized vowels. This acoustic property is directly linked to the perception of these vowels. The more acoustically similar the two oral and nasalized vowels are, the more perceptually similar they are. This is shown by the calculation of the perceptual similarity and the perceptual distance between the vowel pairs based on Shepard [15], which closely matches our experimental results.

This asymmetry could also explain why nasalized vowels are asymmetrically distributed in the phoneme inventory of many languages. According to Hajek [17] in languages like Chamorro, Valaisan Franco-Provençal and Picard high vowels are the preferred vowels to be nasalized. As the difference between oral and nasalized vowels is recognized better in high vowels than in low vowels, having no low nasalized vowels in a language means an improvement of communication. However, further research is necessary to check whether the different degree of acoustic modification alone can explain the asymmetrical distribution of vowel nasalization in all languages.

The perceptual effects that we found could lend support to the explanation of the typological preference for nasalization in high vowels over nasalization in low vowels as a historical evolutionary effect [18]. However, if the German native speakers are more inclined to learn a nasalization contrast in high vowels than a nasalization contrast in low vowels, it suggests a role for phonetics in synchronic phonology [19].

4. Conclusion

The perceptual vowel identification experiment using oral and nasalized vowels as stimuli showed that native speakers of German who do not have contrastive or allophonic nasalization in their native language perceive the difference between oral and nasalized vowels asymmetrically. Low oral and nasalized vowels are more likely to be confused with each other than non-low oral and nasalized vowels. These experimental results are explained with the acoustic and perceptual similarity between oral and nasalized low vowels, which makes it difficult for the hearer to recognize nasalization. Non-low oral and nasalized vowels differ acoustically and thus also perceptually more with each other than low oral and nasalized vowels, which reduced their probability to be confused with each other. Thus, the acoustic modification due to the nasalization is perceived better in non-low vowels than in low vowels. This is true not only in languages with contrastive or allophonic vowel nasalization but also in language without it.

5. References

[1] J. J. Ohala, “Phonetic Explanations for Nasal Sound Patterns,” in *Nasálfest: Papers from a Symposium on Nasals and Nasalization*, C. A. Ferguson, L. M. Hyman, and J. J. Ohala, Eds. Stanford: Language Universals Project, 1975, pp. 289-316.

[2] M. F. Schwartz, “The Acoustics of Normal and Nasal Vowel Production,” *The Cleft Palate Journal*, vol. 5, no. 2, pp. 125-140, 1968.

[3] R. Wiese, *The Phonology of German*. Oxford: Oxford University Press, 1996.

[4] Z. S. Bond, “Identification of Vowels Excerpted from Neutral and Nasal Contexts,” *Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1229-1232, 1976.

[5] A. S. House and K. N. Stevens, “Analog Studies of the Nasalization of Vowels,” *Journal of Speech and Hearing Disorders*, vol. 21, no. 2, pp. 218-232, 1956.

[6] B. Hayes, *Introductory Phonology*. Chichester: Wiley-Blackwell, 2009.

[7] P. Delattre, “Les attributs acoustiques de la nasalité vocalique et consonantique,” *Studia Linguistica*, vol. 8, no. 2, pp. 103-109, 1954.

[8] S. Maeda, “Acoustics of Vowel Nasalization and Articulatory Shifts in French Nasal Vowels,” in *Nasals, Nasalization, and the Velum. Phonetics and Phonology*, Volume 5, M. K. Huffman, and R. A. Krakow, Eds. San Diego: Academic Press, 1993, pp. 147-167.

[9] D. Whalen and P.S. Beddor, “Connections between Nasality and Vowel Duration and Height: Elucidation of the Eastern Algonquian Intrusive Nasal,” *Language*, vol. 65, no. 3, pp. 457-486, 1989.

[10] L. Lintz and D. Sherman, “Phonetic Elements and Perception of Nasality,” *Journal of Speech, Language, and Hearing Research*, vol. 4, no. 4, pp. 381-396, 1961.

[11] A. P. Benguerel and A. Lafargue, “Perception of Vowel Nasalization in French,” *Journal of Phonetics*, vol. 9, pp. 309-321, 1981.

[12] W. Styler, “On the Acoustical and Perceptual Features of Vowel Nasality, “ Ph.D. dissertation, Depart. Ling., Univ. Colorado, Boulder, CO, 2015.

[13] M. M. Azevedo, *Portuguese: A Linguistic Introduction*. Cambridge: Cambridge University Press, 2005.

[14] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer [Computer program]. Version 5.4.04, retrieved from http://www.praat.org/, 2015.

[15] R. N. Shepard, “Psychological Representation of Speech Sounds,” in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds. New York: McGraw-Hill, 1972, pp. 67-113.

[16] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, URL: http://www.Rproject.org/, 2015.
[17] J. Hajek, *Universals of Sound Change in Nasalization*. Oxford: Blackwell Publishers, 1997.
[18] J. Blevins, *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press, 2004.
[19] B. Hayes and D. Steriade, "Introduction: The Phonetic Bases of Phonological Markedness, " in *Phonetically Based Phonology*, B. Hayes, D. Steriade, and R. M. Kirchner, Eds. Cambridge: Cambridge University Press, 2004, pp. 1-33.

Then, what is charisma? The role of audio-visual prosody in L1 and L2 political speeches

Hiroyuki Tanaka[1], *Tamara Rathcke*[2]

[1]Kotobaax, The Hague, Netherlands
[2] English Language and Linguistics, University of Kent, UK
kotobaax.hiro@gmail.com, t.v.rathcke@kent.ac.uk

Abstract

Charisma plays a significant role in political speeches, and determines the ability of a politician to carry an audience. While acoustic features of charisma have received some empirical attention, the contribution of visual prosody has been mostly neglected in studies focusing on features of a charismatic appearance. Unknown are also the audio-visual cues to charisma in non-native speakers. This small-scale study investigated speeches delivered by Donald Trump (L1 American English) and Arnold Schwarzenegger (L1 Austrian German, L2 American English). Video and audio recordings of their political speeches (around 25 min per speaker) and the transcripts were used. The use of pitch range, speech rate, emphatic stress and hand gestures was analysed. In order to establish the core means of the speakers' persuasive influence on their audiences, within-speaker comparisons were conducted for phrases with and without cheering from the audiences. The results showed some differences in the use of the audio-visual prosodic features between the L1 and L2 speaker as well as some similarities, and suggest that charisma is not easily attributable to a fixed set of prosodic means but may be best understood as a skillful modulation of audio-visual prosody in social interaction.

Index Terms: political speech, charisma, speech prosody, visual prosody, beat gesture, rhythm

1. Introduction

The term "charisma" originates from the Greek χάρισμα, meaning "a gift". The term has been traditionally used to refer to the ability of some persons to exert a strong influence on others, to make them believe in high personal competency and extraordinary powers of the speakers ([1]); these speakers are able to attract and retain large audiences ([2]). Although charisma is difficult to define precisely, listeners usually find it easy to identify if a speaker is charismatic or not.

One of the ways to observe if an audience regards a speaker as charismatic or not is to study the speaker's persuasiveness, and the link between persuasion and charisma has been previously discussed in the literature ([3]). Persuasiveness of a speaker plays a particular role in political speeches that might determine the rise or fall of a political party, make audiences take up required actions.

Although multimodality of charismatic appearance has been previously noted ([4]), most studies to date have concentrated exclusively on acoustic-prosodic features ([4, 5]). The main goal of the present study was to combine the auditory and the visual channels of charisma in political speeches, and to estimate the relative contribution of the two by taking into account an appreciative audience response such as cheering, applause and whistling. The second research question concerned the expression of charisma in non-native, as compared to native, speech. A non-native accent was expected to have its imprint in prosody ([6]), and to impact upon the use of prosodic means to charisma in contrast to non-verbal cues which were expected to be comparably used in both native and non-native politicians.

2. Method

2.1. Choice of speakers

The speakers selected for the present study were Donald Trump (L1 American English; hereafter DT) and Arnold Schwarzenegger (L1 Austrian German, L2 American English; hereafter AS). Both speakers are known to be popular and well-established public figures of American political scene, who have given various speeches for their electoral campaigns (presidential or governmental). Both can be described as charismatic as far as the scope of their persuasive popularity is concerned.

2.2. Speeches and sampling

Three recordings were selected, resulting in approximately 25 min material per speaker. Two shorter speeches were delivered by Trump (DT-1: a victory speech to his audience in Nevada after becoming the Republican nominee and DT-2: an electoral campaign speech on his economic policy in Pennsylvania). One longer speech was delivered by Schwarzenegger in the Republican convention in 2004 in his role as the governor of California. The videos stem from YouTube.

Appreciative audience responses during the speeches (such as applause, whistling, screaming) were identified and marked in time (negative reactions such as booing were not included, cf. [7]). Such audience responses are often considered a significant indicator of speaker persuasiveness ([7]). Syntactically complete phrases overlaid with cheers constituted the group of target phrases of this study (each phrase duration was between 1.2 and 6.8 sec). Control phrases comprised of several consecutive phrases with the overall duration of more than 6.8 sec, and did not involve any simultaneous cheering or disapproving noises. Four sets of control phrases were chosen from DT-1 and DT-2 recordings (with the total duration of 190 sec) and three sets were selected from AS (with the total duration of 125 sec; for more detail on this procedure see [8]).

2.3. Prosodic features

Previous research has established that an increased pitch range, moderately fast speaking rate, and emphatic stress are the core features of prosodic importance for the creation of charisma, at least in American English – the variety under investigation in the present study ([2,5]). Accordingly, these three prosodic features were isolated in target and control phrases, analysed and transcribed using Praat.

2.4. Annotation of gestures

This study concentrated on speech-accompanying hand gestures only. A gesture was defined as a movement or series of continuous movements of one or two hands simultaneously during speech. Table 1 summarises previous accounts of such gestures, in comparison to the account taken up in the present study. Occasional cases of doubts where a hand movement could be interpreted as two different gestures, both gesture types were annotated for the same phrase. Importantly, the duration of gesturing during speech was not measured, only the gesture types in target vs. control phrases were annotated.

Table 1. *A summary of gestural classifications in previous accounts (account-1 ([9,10]); account-2 ([11]); account-3 ([12]) and the present study.*

Account-1	*Account-2*	*Account-3*	*Present annotation*
Baton-like/ Ideographic	-	Beats	Rhythmic (**RG**)
Deictic	Demonstra-tive	Deictic	Indexical (**IG**)
Symbolic/ Emblematic	Connotative	Metaphoric	Connotative (**CG**)
Illustrating/ Icono-graphic	Mimic/ Symbolic	Iconic	Denotative (**DG**)

3. Results and discussion

3.1. Speaker comparison

Table 2 summarises and compares values of the relevant prosodic features, Figure 2 displays the overall distribution of the four gesture types measured for the two speakers. Overall, prosodic measurements support our expectation and show a discrepancy primarily in the pitch range between the two speakers: Although the physiological pitch range of the two speakers is comparable (ranging between 75 and 380 Hz), the functional pitch range is substantially smaller in the L2-speaker AS (due to his more reluctant use of high pitch), which is in line with the findings on L1-German speakers of English ([6]). However, in contrast to our expectation, the two speakers also showed substantially different personal preferences for their speech accompanying gestures, with DT having a strong preference for rhythmic gestures and AS preferring mainly the indexical gestures. Moreover, DT produced emphatic stresses much more frequently than AS.

3.2. Comparison of target vs. control phrases

Table 3 shows the distribution of prosodic features across target vs. control phrases of the two speakers. Emphatic stress occurred exclusively in the target phrases (i.e. accompanied by cheers from the audience) in case of AS; similarly, only one emphatic stress was observed outside target phrases in DT. It was likely to be carried by pitch accents with an extended pitch range. Target phrases also displayed a slight increase of speech rate. In sum, prosodic features previously identified as cueing charismatic speech ([2,5]) were observed in target rather than in control phrases.

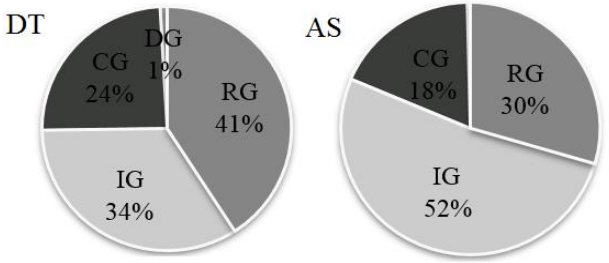
Table 2. *Speech rate and frequency for the two speakers.*

Measurement	*DT*	*AS*
Mean speech rate	4.0 syll/sec	4.0 syll/sec
Min F0 produced	74 Hz	75 Hz
Mean min F0	106 Hz	115 Hz
Max F0 produced	387 Hz	376 Hz
Mean max F0	297 Hz	241 Hz
Mean pitch range	18 st	13 st
Emphatic stress use	3.92 times/min	2.59 times/min

Table 3. *Percentage of emphatic stress occurrences and mean values for speech rate (syll/sec) and pitch range (st) measured in target vs. control phrases of DT vs. AS.*

	DT		*AS*	
Measurement	*target*	*control*	*target*	*control*
Emphatic stress	99 %	1 %	100 %	0 %
Speech rate	4.2	3.8	4.1	3.9
Pitch range	28	18	24	13

Figure 1: *Distribution of the gesture types across all phrases produced by DT and AS.*



4. Conclusions

These results further suggest that charisma is not easily attributable to a fixed set of prosodic means but may be best understood as a skillful modulation of audio-visual prosodic means in social interaction. More fine-grained analyses of gesture timing, inclusion of disapproving responses, cross-cultural comparisons and a larger database of political speeches will help to shed brighter light on how charisma arises in different socio-cultural contexts.

References

[1] Awamleh, R. & Gardner, W. L. (1999). Perceptions of Leader Charisma and Effectiveness: The Effects of Vision Content, Delivery, and Organisational Performance. Leadership Quarterly, 10(3), 345-373.

[2] Rosenberg, A. & Hirschberg, J. (2009). Charisma Perception from Text and Speech. *Speech Communication*, 51, 640-655.

[3] Dewan, T., Humphreys, M., & Rubenson, D. (2013). The Elements of Political Persuasion: Content, Charisma and Cue. The Economic Journal, 124, 257-292.

[4] Signorello, R., D'Errico, F., Poggi, I., Demolin, D., & Mairano, P. (2012). Charisma Perception in Political Speech: A Case Study. In: Mello, H., Pettorino, M., & Raso, T., eds. Proceedings of the 7th GSCP International Conference: Speech and Corpora, February 2012. Firenze University Press, pp. 343-348.

[5] Niebuhr, O., Brem, A., Novák-Tót, E., & Voβe, J. (2016). Charisma in Business Speeches: A Contrastive Acoustic-Prosodic

Analysis of Steve Jobs and Mark Zuckerberg. In: *Proceedings of the 8th International Conference of Speech Prosody, Boston*.

[6] Mennen, I., Schaeffler, F., Docherty, G. (2012). Cross-language differences in fundamental frequency range: A comparison of English and German. Journal of Acoustical Society of America 131(3), 2249.

[7] Guerini, M., Strapparava, C., & Stock, O. (2008). CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing. Journal of Information Technology & Politics, 5(1), 19-32.

[8] Tanaka, H. (2016). *Acoustic and Gestural Analyses of Political Speeches for Effective Public Speaking Training in TESOL*. Unpublished MA thesis, University of Kent.

[9] Ekman, P. & Friesen, W. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage and Coding. Semiotica, 1, 49-98.

[10] Efron, D. (1972). Gesture, Race and Culture. The Hague: Mouton & Co.

[11] Wundt, W. (1973). The Language of Gestures. The Hague: Mouton & Co.

[12] McNeill, D. (1992). Hand and Mind. Chicago: University of Chicago Press.

Coordination Deficits in Essential Tremor Patients with Deep Brain Stimulation

Tabea Thies, Anne Hermes & Doris Mücke

I*f*L Phonetik, University of Cologne, Germany
{tabeathies;annehermes;doris.muecke}@uni-koeln.de

Abstract

It has been shown elsewhere that speech is deteriorated in Essential Tremor patients treated with Deep Brain Stimulation [2]. However, those studies were restricted to fast syllable repetition tasks. The present study is a pilot study investigating the coordination of oral gestures in normal sentence production using Electromagnetic Articulography (EMA). It focuses on intergestural coordination patterns of the labial and lingual system. We test the interplay of these patterns in syllables with different complexity within the framework of Articulatory Phonology. When comparing simple (CV) and complex syllables (CCV), it is assumed that the rightmost C of a consonant cluster shifts towards the vowel whereas the leftmost C shifts away from it to make room for the added C. However, we did not find this expected shift of the rightward C in the patients' production, indicating an articulatory mistiming within a prosodic constituent such as the syllable.

Index Terms: Deep Brain Stimulation, syllable structure, onset coordination, gestural coordination patterns, Essential Tremor patients, speech motor system, competitive coupling structure

1. Introduction

Deep Brain Stimulation (DBS) is an essential treatment for patients with medication resistant Essential Tremor (ET). A frequent side effect of this treatment is that ET patients report on detrimental effects on their speech under stimulation. Acoustic studies based on fast syllable repetition tasks for patients with multiple sclerosis [1] and essential tremor [2] showed a deterioration of speech in the acoustic domain in fast syllable repetition tasks such as /tatata/, /kakaka/ or /papapa/. They found frication during intended voiceless closures in /ta/, /ka/ and /pa/ as well as an increase in voicing during the entire syllable cycle. Both, incomplete oral closures and a reduced glottis control indicate a deterioration in speech motor control of the oral and glottal system under stimulation.

In Articulatory Phonology, the physical dimension of speech is examined, where speech production is divided into articulatory gestures. These gestures have occupied coordination patterns which are determined within the coupled oscillator model [3]. Each gesture is associated with a temporal trigger or clock (modelled by oscillators that are coupled with each other in a pairwise fashion). There are two intrinsic modes: gestures are either activated simultaneously (in-phase) or sequentially (anti-phase). This temporal trigger has been applied to syllable coordination. According to this model, in simple syllable onsets (CV) consonants and vowels should be coupled in-phase, assuming that the consonantal and vocalic gestures are initiated at the same time. In contrast to this, a complex syllable onset pattern (CCV) require a combination of these coupling modes (competitive coupling structure): (i) both consonants belong to the same syllable onset, both are coupled in-phase with the vowel and (ii) for perceptual recoverability, both consonants are at the same time coupled in an anti-phase mode with each other, leading to a leftward and a rightward shift of both consonants. The rightmost C shifts towards the vowel, whereas the leftmost C shifts further apart from ([4], [5], referred to as C-Center effect). Thus, these shifts are used as a diagnosis for phonological syllable constituency.

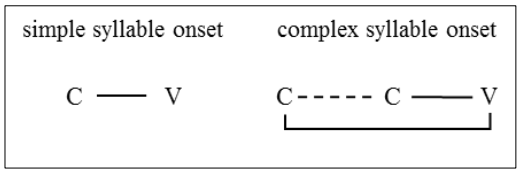


Figure 1: Coupling graphs for simple and complex syllable onsets. The solid line represents in-phase coupling, dotted line anti-phase coupling.

The present study investigates the effects of DBS in ET patients in normal sentence production directly in the kinematic domain by using Electromagnetic Articulography. It focuses on the intergestural coordination patterns of the labial and lingual system. We test coordination patterns in syllables with different complexity within the framework of Articulatory Phonology [3] [6]. These different complexities require different demands in the speech motor system. More specifically, this study investigates the timing relation of competitive and non-competitive coupling structures in simple syllable onsets (CV) and in complex syllable onsets (CCV) to shed light on the question of timing deficits in the patients' production of prosodic constituents such as the syllable.

2. Method

2.1. Speakers and Recordings

So far, we analyzed 5 ET patients with VIM-DBS (bilateral implanted). Patients were recorded with stimulation-on (DBS-on) and stimulation-off (DBS-off). The articulatory data was recorded with a 3D Electromagnetic Articulograph (AG 501) with sensors placed on the upper and lower lip, the tongue tip and tongue dorsum. The recordings took place at the I*f*L Phonetics lab in Cologne.

2.2. Test material

The speech material, which is a subset of larger corpus, is designed to test gestural coordination in simple and in complex syllable onsets. The words /lima/ and /pina/ have a simple syllable onset (CV), whereas /plina/ has a complex one (CCV). The target words were presented in a carrier sentence such as "Er hat wieder _____ gesagt." *('He said _____ again.')*. The target word bears the nuclear accent and the accented syllable

was the first one. Each target word was presented five times in a pseudo-randomized order. The speakers were instructed to read the sentences at a normal speaking rate. In total 150 tokens were recorded (5 speaker x 3 target words x 5 repetitions x 2 DBS conditions).

2.3. Annotation and Measurements

The articulatory data were annotated manually in the EMU speech database system [7]. We labelled the gestural movements of the lips, the tongue tip and the tongue dorsum. For every gestural segment two different measuring points got chosen. We marked the gestural onset and the gestural target of every vocalic and consonantal gesture of the first syllable of the target word in CV and CCV by identifying zero-crossings in the respective velocity and acceleration traces (see Fig. 2).

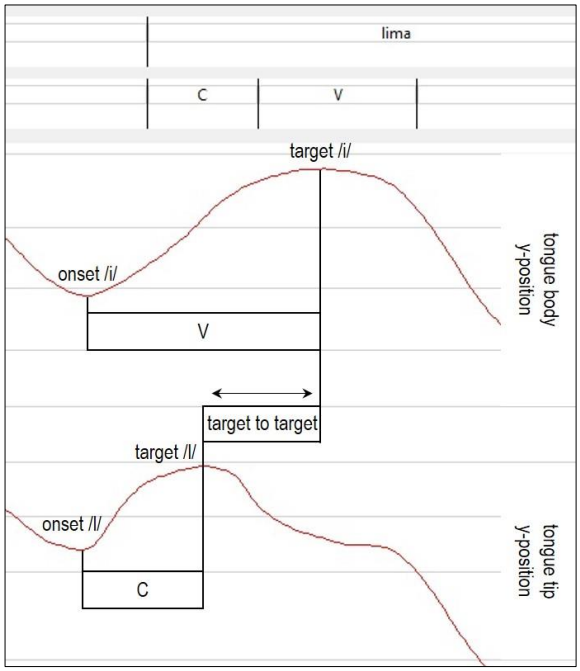


Figure 2: Annotation of the tongue tip and tongue body movement at two levels in EMU. In addition, target to target interval for timing relation of a non-competitive coupling structure in /lima/ (CV).

For the rightmost C shift we measured the interval between the target of the rightmost C (prevocalic) and the target of the following vowel in both CV and CCV. We compared the interval between the target of /l/ with the target of /i/ in /lima/ (CV) with the appropriate interval in /plina/ (CCV). In the latter case the latency should decrease, implying a shift of the rightmost C towards the vowel.

To determine the shift of the leftmost C, we compared the timing interval between the target of the leftmost C, i.e. /p/ in /pina/ with /p/ in /plina/ relative to the target of the vowel. In this case the interval in the complex syllable onset is assumed to be greater since the consonant should shift further apart from the vowel.

2.4. Analysis

The present analysis is restricted to the results for patients in DBS-off condition. However, preliminary results show that there are no differences between DBS-on and DBS-off condition in the intergestural timing patterns, indicating that there is already deterioration in speech due to the tremor itself.

For the analysis of the intergestural coordination patterns we compared the shifts of the consonants in CV vs. CCV. In complex syllable onsets a leftward and a rightward shift of both consonants relative to the following vowel is assumed. To test whether the consonants are adjusted when a consonant is added we determined the interval between the targets of the consonantal gestures to analyze a possible shift (cf. target-to-target in Fig.2).

2.4.1. Rightmost C analysis

Figure 3 shows the temporal interval between the rightmost C and the vowel. Contrary to the assumptions made, the interval *increases* from CV to CCV syllable structure for 4 out of 5 speakers (except S3). The averaged interval for /l/ relative to /i/ is 28 ms *longer* in /plina/ (138 ms) compared to /lima/ (110 ms). This means that the rightmost C does not shift at all to the right but to the left.

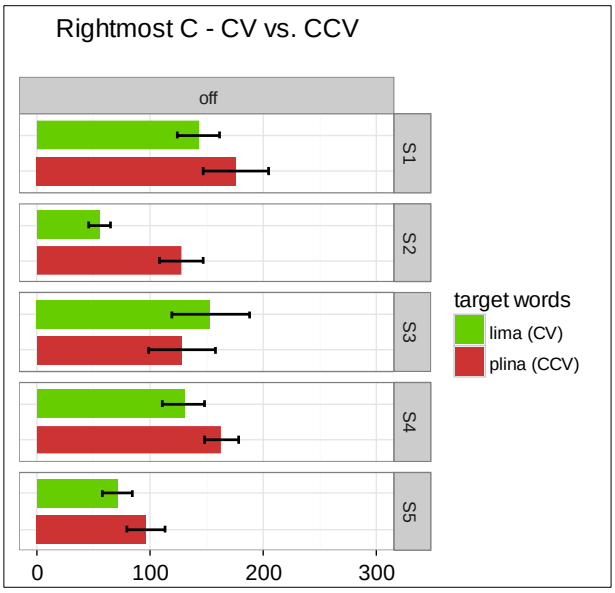


Figure 3: Latencies from the target of the rightmost C relative to the target of vowel (ms) comparing simple CV vs. complex CCV condition in DBS-off.

To test our results, we ran an ANOVA with speaker as specified subject variable and target word as within-speaker variable. Testing the position of the rightmost C for all patients showed that the latencies in CV do not differ from the ones in CCV ($F(1,4) = 3.17$, $p>0.05$). It can be concluded that the rightmost C in the cluster does not shift, thus, not making room for the added /p/.

2.4.2. Leftmost C analysis

Figure 4 illustrates the temporal intervals between the leftmost C and the vowel. The interval increases on average of 169 ms from CV to CCV syllable structure for all speakers. As

expected, the leftwards C (/p/) shifts away from the vowel (/i/). The overall mean interval between the leftmost C and the following vowel increases from 184 ms in /pina/ to 253 ms in /plina/.

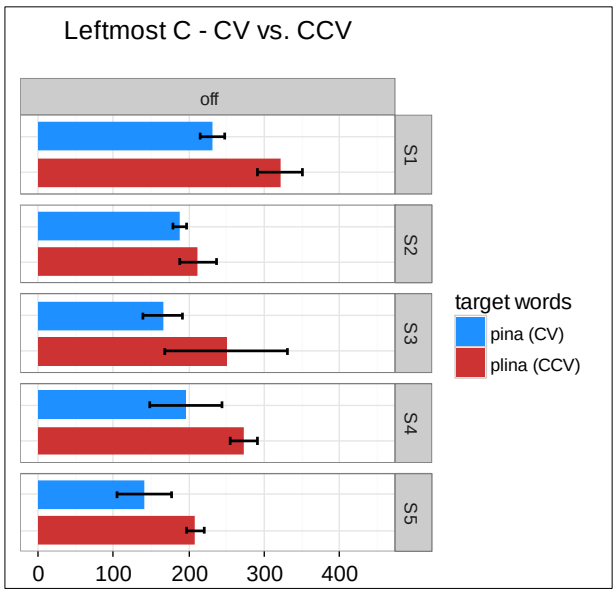


Figure 4: Latencies from the target of the leftmost C relative to the target of vowel (ms) comparing simple CV vs. complex CCV condition in DBS-off.

The ANOVA showed that syllable structure has a significant effect on the leftmost C shift ($F(1,4) = 34.53$, $p<0.05$). The leftmost C in the cluster systematically shifts to the left.

3. Discussion and Conclusion

In this paper we show timing deficits in the production of complex syllable onsets for Essential Tremor patients treated with VIM-DBS. The timing deficits already occur in stimulation-off condition. It is likely that the tremor itself already induces some coordination deficits in CCV syllables, requiring a complex timing pattern. When comparing CV and CCV syllables (/pina/, /lima/ and /plina/), as expected the leftmost C shifts further apart from the vowel. However, the expected rightward shift for the rightmost C in a complex onset does not take place. We interpret the results in the way that Essential Tremor patients have problems with realization of competitive coupling structures in the articulatory domain, resulting in a coordination deficit in syllable onsets.

4. References

[1] Pützer, M., Barry, W. J. & Moringlane, J. R. (2007). Effect of deep brain stimulation on different speech subsystems in patients with multiple sclerosis. *Journal of Voice*, 21(6):741-753.

[2] Mücke, D., Becker, J., Barbe, M. T., Meister, I., Liebhart, L., Roettger, T. B., ... & Grice, M. (2014). The effect of deep brain stimulation on the speech motor system. *Journal of Speech, Language, and Hearing Research*, 57(4), 1206-1218.

[3] Nam, H. & Saltzman, E. (08/2003). A competitive, coupled oscillator model of syllable structure. In: *Proceedings of the 15th international congress of phonetic sciences*, 2253-2256.

[4] Marin, Stefania & Pouplier, Marianne. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. In: *Motor Control* 14(3), 380-407.

[5] Goldstein, L., Chitoran, I., & Selkirk, E. (2007, August). Syllable structure as coupled oscillator modes: evidence from Georgian vs. Tashlhiyt Berber. In: *Proceedings of the XVIth international congress of phonetic sciences* (pp. 241-244).

[6] Browman, C. P. & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. Bulletin De La Communication Parlée, 5, 25–34.

[7] Cassidy, S. & Harrington, J. (2011). Multi-level annotation in the EMU speech database management system. *Speech Communication* 33, 611-677.

Die Wahrnehmung reduzierter Sprache unter Rauschen

Frederike Urke [1], *Prof. Dr. Henning Reetz* [2], *Dr. Gea De Jong-Lendle* [1]

[1] Deutschland, Philipps-Universität Marburg, Institut für Germanistische Sprachwissenschaft
[2] Deutschland, Goethe-Universität Frankfurt, Institut für Empirische Sprachwissenschaft / Phonetik

frederike.urke@staff.uni-marburg.de, reetz@em.uni-frankfurt.de, gea.dejong@staff.uni-marburg.de

Abstract

Diese Studie befasst sich mit der Frage, ob es einen Unterschied beim Verständnis reduzierter gegenüber kanonischer Sprache im Störgeräusch gibt. Das durchgeführte Experiment zeigt, dass reduzierte Sprache unter rosa Rauschen schlechter erkannt wird als die kanonischen Pendants. Grund dafür könnte das Wechselspiel zwischen der Maskierung akustischer Eigenschaften, die für das Verständnis reduzierter Sprache relevant sind, und den Reduzierungen selbst sein.

Schlüsselbegriffe: Sprachwahrnehmung, reduzierte Sprache, Sprache im Störgeräusch

1. Einführung

Sprachproduktion und -wahrnehmung unterliegen komplexen Prozessen, auf physiologischer wie auf psychologischer Ebene. Generell wird angenommen, dass für die sprachliche Kommunikation zwischen zwei Menschen verschiedene Stationen durchlaufen werden: Sprechergehirn, Artikulationsorgane des Sprechers, Medium (z.B. Luft), Perzeptionsorgane des Hörers und Hörergehirn [1]. In und zwischen all diesen Stationen kann es zu Variationen kommen, die unterschiedliche Ursachen haben können. Dabei spielen Aspekte der Planungsäußerung genauso eine Rolle wie die Gesprächssituation oder physiologische Elemente. So ist die Planungsäußerung von der Gesprächssituation (Umgebung, Gesprächspartner, etc.) abhängig, die Begebenheiten der Sprech- und Hörorgane der Gesprächspartner spielen eine Rolle und auch Erwartungen und Erfahrung der Beteiligten können Variationsquellen darstellen.

Sprachproduktion und -wahrnehmung sind eng miteinander verknüpft; das Sprachsignal stellt eine Verbindung zwischen Sprechen und Hören dar [2]. Deshalb erschwert auch Variation in der Sprachproduktion die Modellierung von Sprachverstehensprozessen: Das Problem der Invarianz bezieht sich darauf, dass dasselbe Phonem (kontextabhängig) nicht immer gleich klingt (Koartikulation), während das Problem der Segmentierung sich darauf bezieht, dass gesprochene Laute ineinander übergehen und eine Trennung erschweren (Assimilation) [3]. Die Hypo- und Hyperartikulationstheorie (H&H-Theorie) versucht die Varianz der Sprachproduktion zu erklären, indem von einem Kontinuum zwischen den Endpunkten Hypo- und Hyperartikulation beim Sprechen ausgegangen wird [4]. Dabei versuche der Sprecher einerseits möglichst äußerungsorientiert/verständlich (Hyperartikulation) und andererseits so systemorientiert/ökonomisch wie möglich (Hypoartikulation) zu artikulieren. Der Hörer müsse den sprachlichen Kode von anderen Lexemen diskriminieren können und nutze dazu auch Vorwissen und Erfahrung [4].

Reduzierte Sprache meint insbesondere Assimilationen und Auslassungen ganzer Segmente und tritt speziell in natürlichen Gesprächssituationen auf [5, 6]. Obwohl Sprache mit großer Varianz und verschiedenen (Stärken an) Reduktionen produziert wird, ist ihre Verständlichkeit meist dennoch gegeben [6]. Dass sprachliches Verhalten bei der Sprachproduktion im Störgeräusch angepasst wird (Lombard-Reflex: lauteres, deutlicheres Sprechen im Störgeräusch), ist zwar bereits gezeigt worden [7, 8], doch kann insbesondere bei natürlichen, informellen Gesprächssituationen angenommen werden, dass auch reduzierte Sprache im Störgeräusch produziert wird [6].

Sprachverständlichkeit im Störgeräusch war bisher meist Thema bei klinischen Studien [9, 10], wobei die Sprachverständlichkeitsschwelle eine zentrale Rolle spielte. Diese Studie will die Aspekte reduzierter Sprache und von Sprache im Störgeräusch kombiniert hinsichtlich ihrer Verständlichkeit untersuchen. Motivation für dieses Thema ist nachstehende generelle Beobachtung, die Zimmerer wie folgt formuliert: „Despite these reductions, listeners usually understand what has been said, even in very noisy conditions“ [6]. Daran anknüpfend soll in dieser Studie die folgende Forschungsfrage beantwortet werden: Gibt es einen Unterschied zwischen der Verständlichkeit reduzierter und kanonischer Sprache unter Rauschen?

2. Reduzierte Sprache

Reduzierte Sprache meint von der kanonischen Aussprache durch Reduktion von Segmenten und artikulatorischen Gesten abweichende Sprache [11, 12]. Im Deutschen sind das besonders Elisionen, Assimilationen und Vokalzentralisierung, wobei als Hauptursache die sprachliche Ökonomie angesehen wird [13]. Bisher konnten diverse Faktoren identifiziert werden, die die Produktion reduzierter Sprache beeinflussen: Sprechstil und -tempo [6], Sprechergeschlecht, Dialektregion [11], Wortfrequenz und -länge, Position im Wort, Betontheit [14], Wortklasse [13] und -vorhersagbarkeit, Kollokationshäufigkeit [15] und Dichte phonologischer Nachbarschaft [16].

Auch hinsichtlich Verständlichkeit und Worterkennung wurde reduzierte Sprache bereits untersucht, wobei Effekte bezüglich des Kontextes und der Reduktionsstärke [17] und phonetischer Parameter [18] gefunden wurden. Ernestus et al. können zeigen, dass sowohl der phonetische als auch der semantische/syntaktische Kontext das Erkennen von Wortformen vereinfachen [17]. Für die Studie wurden Hörern verschieden stark reduzierte Wortformen in drei Kontexten (isoliert, eingeschränkt, voller Kontext) präsentiert. Diese Vorgehensweise führte auch zu der Erkenntnis, dass die Worterkennung mit ansteigender Reduktionsstärke und verminderndem Maß an Kontext verringert wird. Zusätzlich kann aufgezeigt werden, dass eine größere phonetische Distanz zwischen redu-

zierter und kanonischer Form die Worterkennung der reduzierten Varianten erschwert [17]. Weiterhin wurde festgestellt, dass phonetischer/akustischer Kontext eine weitaus wichtigere Rolle bei der Worterkennung spielt als die Wortvorhersagbarkeit lediglich durch semantischen/syntaktischen Kontext [19].

Hinsichtlich der phonetischen Faktoren konnte Kohler zeigen, dass durch Reduzierungen auftretende Nasalierung und Glottalisierung disambiguierende Funktionen in der Sprachverständlichkeit einnehmen können [18]. Nasalierung betrifft die Nachbarphone eines ursprünglichen, dann aber getilgten Nasals; Glottalisierung bezieht sich auf die Reduktion eines oralen Plosivs bis zum Glottalverschluss. Das akustische Merkmal der Nasalierung wird meist dazu genutzt, Formen mit und ohne Nasal zu unterscheiden. Gleiches gilt bezüglich der Glottalisierung für die An- oder Abwesenheit eines Plosivs [18].

3. Sprachverständlichkeit im Störgeräusch

Die Verständlichkeit von Sprache in unterschiedlichen Störgeräuschen war, wie reduzierte Sprache, bereits häufiger Gegenstand der Forschung. So konnten auch hier Faktoren identifiziert werden, die die Sprachverständlichkeit im Störgeräusch beeinflussen: So sind längere Worte generell leichter verständlich als kurze und hoch-frequente lassen sich bei einem Signal-Rausch-Verhältnis (S/N) verstehen, das 13,5 dB niedriger (also stärkeres Rauschen, schwächeres Signal) ist als das bei der Verständlichkeit nieder-frequenter Worte [20]. Daraus schlussfolgert Howes, dass Wortlänge und -frequenz für den Großteil der Varianz der Sprachverständlichkeit im Störgeräusch (bei einem breiten Rauschspektrum) verantwortlich seien, während die Sprachverständlichkeit unabhängig von der Verständlichkeit einzelner Phoneme sei [20]. Weiterhin konnte gezeigt werden, dass eine einfache Korrelation zwischen Wortvorhersagbarkeit und -verständlichkeit besteht: Mit steigendem S/N wird ein Wort schneller erkannt, wenn es leichter vorhersagbar ist [21].

Auch bezüglich der Sprachwahrnehmung im Störgeräusch durch mehrsprachige Hörer und hinsichtlich der Wahrnehmung von Dialektvariation und Akzent im Störgeräusch konnten bereits Erkenntnisse erlangt werden. So konnte demonstriert werden, dass monolinguale Hörer einen Vorteil gegenüber zwei- oder dreisprachigen Hörern bei der Erkennung von Sprache im Störgeräusch haben [22]. Interessanterweise kann für Dialektvariation festgestellt werden, dass Unterschiede bei einem niedrigen S/N signifikant sind und sich dieser Effekt mit steigendem S/N abschwächt [23]. Song und Iverson untersuchten die Verständlichkeit verschiedener Akzente im Englischen (südbritisch, finnisch, koreanisch) im Störgeräusch bei Hörern mit unterschiedlichen Muttersprachen (Englisch, Koreanisch) [24]. Es zeigt sich, dass englische Muttersprachler den englischen Akzent im Störgeräusch besser verstehen als den finnischen, der wiederum besser verständlich ist als der koreanische. Bei koreanischen Muttersprachlern zeigt sich dieses Muster nicht [24]. Bezüglich Dialekt und Akzent scheint es also eher unerwartete Ergebnisse zu geben.

Eine relevante Größe bei der Sprachverständlichkeit im Störgeräusch stellt die Sprachverständlichkeitsschwelle dar. Sie wird allgemein als das Niveau definiert, bei dem ein Hörer 50% der präsentierten Sprache korrekt wiedergeben/erkennen kann [25]. Um diese Schwelle ermitteln zu können, wurde bereits 1977 der SPIN-Test (SPeech Intelligibility in Noise) entwickelt, der eine große Anzahl an Faktoren berücksichtigt, die das Satzverständnis beeinflussen können, wie: phonetische und prosodische Aspekte, Satzkontext, Vertrautheit des Wortes, Störgeräusch und hörer-spezifische Faktoren [10]. Sowohl auf internationaler Ebene, als auch konkret für das Deutsche wurden bereits Sprachverständlichkeitsschwellen für Normalhörende ermittelt. So wurde als internationaler Referenzwert ein S/N von -8,14 dB errechnet [26]. Für das Deutsche werden für den Oldenburger und den Göttinger Satztest Referenzwerte von -6,2 dB bzw. -7,1 dB angegeben [27]; d.h. ein deutschsprachiger Normalhörender erkennt ca. 50% der präsentierten Sprache bei einem S/N von -6,2 dB bzw. -7,1 dB.

4. Material und Methoden

Um die aufgeworfene Forschungsfrage zu untersuchen, wurde ein Experiment durchgeführt, bei dem in einen Trägersatz eingebettete Zielphrasen von Hörern erkannt und notiert werden sollten. Die Zielphrasen waren jeweils einmal in kanonischer und einmal in reduzierter Form eingesprochen worden und wurden anschließend mit rosa Rauschen überlagert.

4.1. Zielphrasen

Bei der Auswahl der Zielphrasen mussten verschiedene Aspekte berücksichtigt werden: Die Stimuli mussten reduzierbar sein, weshalb sich für die Zielphrasen auf Funktionswörter beschränkt wurde, da diese stärker und häufiger reduziert werden (können) als Inhaltswörter [13]. Gleichzeitig sollte vermieden werden, dass die Sprachverständlichkeit durch den Satzkontext beeinflusst wird, weshalb als Trägersatz „*Dieser Satz endet mit …*“ festgelegt wurde. Die eigentlichen Zielphrasen wurden aus Kohlers Studie zu artikulatorischen Prosodien [18] adaptiert, da diese Stimuli nach ähnlichen Kriterien (Reduzierbarkeit, Einbettung in Trägersatz) ausgewählt worden waren. Zusätzlich könnten Aspekte wie Nasalierung und Glottalisierung sowie generell phonetischer Kontext innerhalb der Zielphrase die Verständlichkeit der Phrasen erleichtern [17, 18]. Folgende Zielphrasen wurden also in den Trägersatz eingebettet: *sollen wir, sollten wir, soll er, soll sie, sollen sie, sollten sie, wir können ihn, wir könnten ihn, die können uns, die könnten uns, die können wir uns* und *die könnten wir uns*. Diese Stimuli wurden, jeweils reduziert und kanonisch, von einer phonetisch geschulten weiblichen Person eingesprochen. Es wurde sichergestellt, dass sich reduzierte und kanonische Aussprache voneinander unterschieden. Für jede Zielphrase und jede Aussprachevariante (kanonisch, reduziert) wurde der gesamte Trägersatz eingesprochen, so dass der Trägersatz auch der Aussprachevariante der Zielphrase entsprach. Die Stimuli wurden in einer schallisolierten Kabine mittels eines Großmembran-Kondensator-Mikrofons (AKG Perception 220) mit Popschutz aufgenommen. Der Abstand von Mund zu Mikrofon betrug ca. 20cm. Die Aufnahmen wurden mit dem Focusrite Scarlett Solo (Vers. 07/2014) und mit Audacity (Vers. 2.0.4.) erstellt (Abtastrate: 44100 Hz, Abtasttiefe: 24 bit).

4.2. Manipulation

Die jeweiligen Zielphrasen (im Trägersatz) wurden mit rosa Rauschen überlagert, da dieses Rauschen sich besser für die Maskierung von Sprache eignet [28], sodass die Signal-Rausch-Verhältnisse der einzelnen Zielphrasen zwischen -7,0 dB und -8,0 dB lagen. Dieses S/N wurde festgelegt, nachdem ein Vortest mit den für dieses Experiment aufgenommenen Zielphrasen zeigte, dass die mittlere Sprachverständlichkeitsschwelle für die Stimuli zwischen -7,0 dB und -8,0 dB lag (Schwankungen innerhalb dieses Bereichs waren unvermeid-

lich). Es wurde berücksichtigt, ab welchem Wert überhaupt Sprache erkannt worden war und ab welchem Wert die Zielphrasen korrekt erkannt worden waren. Die Berechnung erfolgte mittels der Formel S/N = 10 * LOG (Signal-RMS² / Rausch-RMS²).

4.3. Probanden

20 Personen (12 weiblich, acht männlich, 21-33 Jahre) nahmen freiwillig an dem Hörexperiment teil, wobei keiner an dem Vortest teilgenommen hatte. Drei Probanden waren zweisprachig aufgewachsen, keiner berichtete Hörschäden.

4.4. Durchführung

Die Präsentation der Stimuli erfolgte bei allen Versuchspersonen einzeln über Kopfhörer (Sennheiser HD 449) bei 55 dB Lautstärke (Brüel & Kjœr Präzisions-Schallpegelmesser Typ 2203/1613). Zu Beginn wurden Metadaten der Probanden abgefragt. Die Sätze wurden den Probanden in pseudorandomisierter Reihenfolge vorgespielt: Es wurde immer die reduzierte Aussprachevariante eines Stimulus‘ vor der kanonischen Variante präsentiert, wobei jedoch zwischen den beiden Varianten unterschiedlich viele andere Stimuli lagen. Es erfolgte also keine paarweise Präsentation von reduzierter und kanonischer Zielphrase. Ein Piepton kündigte die Sätze jeweils an, auf diese folgte sodann eine zehn-sekündige Antwortzeit, in der die Probanden notieren sollten, was sie verstanden hatten. Dafür wurde ein offenes Frageparadigma mittels Fragebogen gewählt. Die Antwortzeit war festgesetzt, auch wenn die Zielphrasen unterschiedliche Längen aufwiesen. Den Probanden wurde neben dieser Information auch mitgeteilt, dass es sich bei der Sprache unter dem Rauschen um Worte oder Satzteile handeln konnte, um Effekte durch Erwartungshaltung zu minimieren. Zusätzlich sollten die Teilnehmer die Sinnhaftigkeit der Satzbedeutung nicht in den Vordergrund stellen, sondern schlicht versuchen, die Zielphrasen zu verstehen.

5. Auswertung und Resultate

480 Antworten wurden gegeben und ausgewertet (12 Zielphrasen x 2 Aussprachevarianten x 20 Probanden). Auffällig dabei war, dass die Wortformen *soll* und *sollen* in den Zielphrasen meist hinsichtlich der zweiten Silbe in *soll**en*** differenziert werden konnten und dass die Segmentkombination Nasal-Plosiv in *kö**nnt**en* ebenfalls meistens als solche erkannt wurden.

Um zu prüfen, wie gut eine Zielphrase von den einzelnen Teilnehmern erkannt worden war, wurde die orthografische Repräsentation mit einer groben phonetischen Transkription der jeweiligen Zielphrase abgeglichen. Bei diesem Abgleich spielte auch die Anzahl erkennbarer Segmente eine Rolle (z.B. ist ein transkribierter Glottalverschluss nicht in der Orthografie zu erwarten, während ein geschriebener Doppelkonsonant nicht einer phonetischen Doppelartikulation entspricht). So wurde beispielsweise die reduzierte Aussprache von *die könnten wir uns* als [dikʰœnʔmaʊnsː] mit 11 zählbaren Segmenten realisiert, die kanonische Variante hingegen mit 14 zählbaren Segmenten als [dikʰœntʰənvɪəʔʊnsː]. Außerdem wurde die relative Position von Segmenten in der Zielphrase (meist auf Silbenbasis) berücksichtigt. Zusätzlich musste beachtet werden, dass die Orthografie reduzierter Sprache kaum Rechnung trägt, so dass dieselbe orthografische Schreibung zweier Stimuli, nämlich die reduzierte und die kanonische Variante einer Zielphrase, vorkommen kann, obwohl sie sich deutlich hinsichtlich ihrer tatsächlichen Aussprache unterschieden.

Für alle Antworten wurde mittels des Abgleichs eine Erkennungsrate berechnet. Dies wurde auf Segmentebene unternommen, wobei dabei jedes transkribierte Segment einem Segment in der Orthografie entsprach (bspw. wurde ein Glottalverschluss vor Vokal mit diesem als ein Segment zusammengefasst, da er orthografisch lediglich durch den Vokal repräsentiert werden konnte). Die Erkennungsraten wurden für alle Antworten berechnet und so ergibt sich für reduzierte Sprache ein Mittelwert von 41,26% (SD = 11,65), und für kanonische ein Mittelwert von 53,23% (SD=16,68). Damit liegen beide nahe der Sprachverständlichkeitsschwelle (50%).

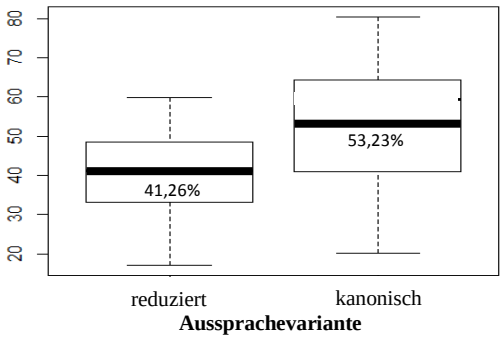


Abbildung 1: *Unterschied der Verteilung der Erkennungsraten (%) nach Aussprachevariante (N=240).*

Abbildung 1 zeigt die Verteilung der Erkennungsraten für beide Aussprachevarianten, inklusive der Mittelwerte. Der Unterschied zwischen den Erkennungsratenmittelwerten ist signifikant (ß=5,25, F(1; 23,4)=9,85, p=0,0045). Reduzierte Zielphrasen werden also signifikant schlechter erkannt als die kanonischen Pendants. Dies wurde mittels eines gemischten Modells berechnet, wobei Zielphrase und Versuchsperson zufällige Faktoren waren und Aussprachevariante (reduziert, kanonisch), Zielphrasenlänge (Anzahl erkennbarer Segmente), S/N und Zielphrase x Aussprachevariante Prädiktoren. Die Berechnungen ergaben, dass der Interaktionseffekt zwischen Zielphrase x Aussprachevariante hoch signifikant ist (F(11; 241,8) =6,27, p<0,0001). Die einzelnen Zielphrasen waren unterschiedlich stark reduzierbar, was einen Einfluss auf die Erkennungsleistung hat. Interessanterweise wird bei der genaueren Analyse deutlich, dass sich dieser Effekt lediglich in drei Zielphrasen (*soll er*, *wir könn(t)en ihn*) manifestiert. Gründe für diese Tatsache müssten in einer weiteren Studie untersucht werden. Sowohl für Zielphrasenlänge als auch für S/N ergaben sich keine signifikanten Effekte, weshalb sie als systematisch beeinflussende Faktoren ausgeschlossen werden können.

6. Diskussion

Das Resultat der Studie zeigt, dass es einen Unterschied bei der Verständlichkeit reduzierter und kanonischer Sprache gibt: Kanonische Sprache wird unter rosa Rauschen signifikant besser verstanden als reduzierte.

Bezüglich des Kontextes ist zu vermerken, dass den Hörern kaum semantischer/syntaktischer Kontext zur Verfügung stand, obwohl dies die Sprachverständlichkeit erleichtert [17]. Da semantischer/syntaktischer und phonetischer Kontext in den Stimuli stets sehr ähnlich waren, wird impliziert, dass die akustischen Informationen der Zielphrase selbst zu unterschiedlichen Leistungen in der Spracherkennung führen und dass es einen relevanten Unterschied zwischen den akustischen Informationen, die die Spracherkennung erleichtern, in

reduzierten im Gegensatz zu kanonischen Zielphrasen gibt. Es ist also möglich, dass bei reduzierten Zielphrasen genau die akustischen Eigenschaften und Segmente durch das rosa Rauschen maskiert werden, die sonst die Erkennung reduzierter Wortformen erleichtern/ermöglichen.

Diese Interpretation harmoniert mit den Erkenntnissen von Van de Ven et al. [19], die die Relevanz akustischer Eigenschaften bei der Erkennung reduzierter Wortformen herausstellen. Nasalierung und Glottalisierung wurden bereits von Kohler als akustische Eigenschaften identifiziert, die zur Disambiguierung bei der Verständlichkeit reduzierter Sprache genutzt werden können. Dies kann durch diese Arbeit bestätigt werden: Unterschiede zwischen *soll* und *sollen* (Nasalierung/Nasal) und die Anwesenheit der Nasal-Plosiv-Kombination in *könnten* (Glottalisierung) wurden meist erkannt. Zusätzlich kann davon ausgegangen werden, dass rosa Rauschen verschiedene Laute unterschiedlich stark maskiert [10] und dass ein Wechselspiel von Reduktionen und Maskierungseffekten die schlechtere Erkennung von reduzierter Sprache hervorruft, was erneut mit der obigen Interpretation einhergeht.

Eine andere Studie [17] zeigt zwar, dass zwischen Erkennungsleistung und Reduktionsstärke ein simpler Korrelationseffekt besteht, doch konnte dieser hier nicht zweifellos bestätigt werden. Der aufgetretene Interaktionseffekt war lediglich bei drei Zielphrasen (hoch) signifikant, weshalb hier weitere Untersuchungen sinnvoll wären.

Howes‘ Annahme, dass Effekte von Wortfrequenz und -länge allein einen Großteil der Sprachverständlichkeit im Störgeräusch erklären könnten [20], widersprechen die Ergebnisse dieser Arbeit. So wurde einerseits kein Effekt für Zielphrasenlänge gefunden und andererseits wäre nach Howes kein Unterschied in den Erkennungsleistungen von reduzierter und kanonischer Sprache zu erwarten gewesen, da die Wörter der Zielphrasen dieselben bei beiden Aussprachevarianten waren und somit auch die Wortfrequenz dieselbe war.

Weiterführende Forschung könnte die Segmente und akustischen Eigenschaften identifizieren, die durch rosa Rauschen maskiert werden und gleichzeitig für die Worterkennung reduzierter Sprache eine relevante Rolle spielen. Auch wäre wietere Forschung hinsichtlich der Mehrsprachigkeit (und auch bezüglich Akzent und Dialekt) von Hörern im Bereich reduzierter Sprache im Störgeräusch bereichernd.

7. Referenzen

[1] R. Wiese, Phonetik und Phonologie, Paderborn: Wilhelm Fink, 2011.

[2] J. R. Sawusch, „Acoustic Analysis and Synthesis of Speech,“ in *The Handbook of Speech Peception*, Malden, MA: Blackwell Publishing Ltd, 2005, pp. 7-27.

[3] T. A. Harley, The Psychology of Language: From Data to Theory, 4th Ed., London & New York: Psychology Press, 2014.

[4] B. Lindblom, „Explaining Phonetic Variation: A Sketch of the H&H Theory,“ in *Speech Production and Speech Modelling*, Kluwer Academic Publishers, 1990, pp. 403-439.

[5] Z. S. Bond, „Slips of the Ear,“ in *The Handbook of Speech Perception*, Malden, MA: Blackwell Publishing Ltd, 2005, pp. 290-310.

[6] F. Zimmerer, Reduction in Natural Speech, 2009.

[7] R. M. Uchanski, „Clear Speech,“ in *The Handbook of Speech Perception*, Malden, Massachusetts: Balckwell Publishing Ltd, 2005, pp. 207-235.

[8] J.-C. Junqua, „The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex,“ *Speech Communication*, Nr. 20, pp. 13-22, 1996.

[9] A. W. Bronkhorst und R. Plomp, „A Clinical Test for the Assessment of Binaural Speech Perception in Noise,“ *Audiology*, Nr. 29, pp. 275-285, 1990.

[10] D. N. Kalikow, K. N. Stevens und L. L. Elliott, „Development of a test of speech intelligibility in noise using sentence materials with controlled predictability,“ *Journal of the Acoustical Society of America*, Nr. 61 (5), pp. 1337-1351, 1977.

[11] D. Byrd, „Relations of sex and dialect to reduction,“ *Speech Communication*, Nr. 15, pp. 39-54, 1994.

[12] M. Ernestus, Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface, Utrecht: LOT, 2000.

[13] K. J. Kohler, „Segmental reduction in connected speech in German: phonological facts and phonetic explanations,“ in *Speech Production and Speech Modelling*, Kluwer Academic Publishers, 1990, pp. 69-92.

[14] H. Mitterer, „How are words reduced in spontaneous speech?,“ in *Proceedings of the 2nd ISCA Workshop on Experimental Linguistics*, University of Athens, 2008.

[15] D. Jurafsky, A. Bell, M. Gregory und W. D. Raymond, „Probabilistic Relations between Words: Evidence from Reduction in Lexical Production,“ in *Frequency and the emergence of linguistic structure*, Amsterdam, John Benjamins, 2000, pp. 229-254.

[16] B. Munson und N. P. Solomon, „The Effect of Phonological Neighborhood Density on Vowel Articulation,“ *Journal of Speech Language and Hearing Research*, Nr. 47 (5), p. 1048–1058, October 2004.

[17] M. Ernestus, R. H. Baayen und R. Schreuder, „The Recognition of Reduced Word Forms,“ *Brain and Language*, Nr. 81, pp. 162-173, 2002.

[18] K. J. Kohler, „Articulatory prosodies in German reduced speech,“ in *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco, 1999.

[19] M. van de Ven, M. Ernestus und R. Schreuder, „Predicting Acoustically Reduced words in Spontaneous speech: The role of Semantic/Syntactic and Acoustic cues in context,“ 2012. [Online]. Available: http://pubman.mpdl.mpg.de/pubman/item/escidoc:1562845/component/escidoc:1562844/VandeVen_Lab_Phon_2012.pdf. [Zugriff Oktober 2015].

[20] D. Howes, „On the Relation between the Intelligibility and Frequency of Occurence of English Words,“ *Journal of the Acoustical Society of America*, Nr. 29 (2), pp. 296-305, 1957.

[21] H. Rubenstein und I. Pollack, „Word Predictability and Intelligibility,“ *Journal of Verbal Learning and Verbal Behavior*, Nr. 2, pp. 147-158, 1963.

[22] D. Tabri, K. M. Smith Abou Chacra und T. Pring, „Speech Perception in Noise by Monolingual, Bilingual and Trilingual Listeners,“ *International Journal of Language & Communication Disorders*, pp. 411-422, 2010.

[23] C. G. Clopper und A. R. Bradlow, „Perception of Dialect Variation in Noise: Intelligibility and Classification,“ *Language and Speech*, Nr. 51 (3), pp. 175-198, 2008.

[24] J. Song und P. Iverson, „Measuring Speech-in-Noise Intelligibility for Spontaneous Speech: The Effect of Native and Non-Native accents,“ in *Proceedings of the XVIIIth International Congress of Phonetic Sciences*, Glasgow, 2015.

[25] M. T. O'Toole, Hrsg., „Speech Reception Threshold,“ in *Mosby's Medical Dictionary*, 8 Hrsg., Elsevier Inc., 2009, http:// medical-dictionary.thefreedictionary.com/speech+reception +threshold.

[26] M. O. Henriques, E. Crestani de Miranda und M. J. Costa, „Speech Recognition Thresholds in Noisy Areas: Reference Values for Normal Hearing Adults,“ *Brazilian Journal of Otorhinolaryngology*, Nr. 74 (2), pp. 188-192, 2008.

[27] HörTech GmbH, „Oldenburger Sprachtests - Überblick und praktische Durchführung,“ 2004. [Online]. Available: http:// www.hoertech.de/web/produkte/audiotests.shtml. [Zugriff 2015]

[28] T. Saeki, T. Tamesue, S. Yamaguchi und K. Sunada, „Selection of meaningsless steady noise for masking speech,“ *Applied Acoustics*, Nr. 65, pp. 203-210, 2004.

Beat it! – Gesture-based Prominence Annotation as a Window to Individual Prosody Processing Strategies

Petra Wagner[1,2], *Aleksandra Ćwiek*[1], *Barbara Samlowski*[3]

[1]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
[2]Cluster of Excellence Cognitive Interaction Technology (CITEC)
[3] Amazon Development Center, Aachen

petra.wagner@uni-bielefeld.de, aleksandra.cwiek@uni-bielefeld.de

Abstract

In recent work [1], we have suggested a novel approach for fine-grained and fast prominence annotation by naïve listeners. Our approach relies on annotators' "drummed" replications of a perceived utterance, modulating their drumming velocity in accordance with the perceptual prominence of consecutive linguistic units (syllables, words). The drumming velocity is then used as a fine-grained operationalization of prosodic prominence. Due to its speed and ease, it allows for the rapid annotation of large amounts of data and yields results that are comparable to fine-grained expert annotations of prominence.

In the present study, we evaluated our method further by (1) comparing the intra-sentential prosodic variation as measured with traditional annotations and the drumming method. Our results show that "drummed" prominences capture speaking-style related variability similarly to conventional annotation methods. Additionally (2), we examined whether individual processing strategies can be identified with the help of Random Forests. This method allows for estimating the individual impact of established prominence correlates on prominence impressions. Our analyses unveil individual listener strategies for blending and integrating top-down, bottom-up and context cues into impressions of prosodic prominence.

1. Introduction

Recently, [1], we introduced a novel approach for fine-grained and fast prominence annotation by naïve listeners, relying on annotators' "drummed" replications of a perceived utterance. Instead of training annotators to make fine-grained judgements of prosodic strength in a time-consuming and cumbersome way, this approach asks listeners to "repeat" a previously heard utterance in a drumming task and to modulate their drumming velocity in accordance with the perceptual prominence of consecutive linguistic units (syllables, words). Drumming velocity (measured by the MIDI output of an electronic drum pad) is used as a fine-grained operationalization of prosodic prominence. This intuitive method exploits the established link between prominence and speech-accompanying gestures [2, 3]. The method has been shown to work with naïve annotators after a very short training phase (10 sentences) and can be used to assess impressions on the level of syllables and words. Due to its speed (close to real-time) and ease, it allows for the annotation of large amounts of data. Another interesting finding was that the prominence patterns yielded by the drumming task show a high correspondence with experts' fine-grained prominence impressions, when averaged across several naïve annotators. Individual naïve "drumming annotators" typically correlate only moderately among each other, especially when drumming "syllable prominences" (cf. Figure 1). This finding points to a lot of individual variation of listeners' strategies in blending linguistic and acoustic prominence cues. Obviously, expert annotations are not helpful to comprehend these individual strategies, as these are typically based on our existing knowledge of how prominences are signaled and strive for a high degree of inter-annotator agreement. An alternative method for rapid prominence annotations by naïve listeners relies on binary impressions of prominence which are later cumulated into a fine-grained prominence profiles [4]. A binary score is too coarse to provide a detailed picture of the individual listener's cue blending strategy.

In this study, we set out to examine "drummed prominence impressions" as a window to individual prominence processing strategies. To this day, a lot of research has shown a myriad of cues to be influential in prominence perception, e.g. acoustic cues such as fundamental frequency excursion and shape [5, 6], duration, intensity, [7], linguistic cues such as word order or lexical class [8, 9] and context cues such as metrical priming or the presence of a nearby pitch accent [10, 11]. It is also known that both acoustic "bottom-up" and linguistic "top-down" cues are somehow integrated when processing prominence [4, 12, 13]. However, we still know little about the presence of individual processing strategies when weighing the many prominence correlates that have been identified.

Our analysis consists of two steps: First, we evaluate our drumming method further to find out whether speaking-style related prosodic production variability is identified similarly in conventional (auditory) and drummed prominence annotations. This should shed light on the question of whether the blending of top-down and bottom-up processing works in a comparable way in both approaches. Second, we build Random Forest regression models predicting individual listeners' "prominence drumming behavior". We use these to assess individual prosody perception strategies by weighing the individual importance of well-established prominence correlates (acoustic, linguistic, contextual) in the prediction models.

2. Is prosodic production variation reflected similarly in expert annotations and drumming?

If the same sentence is uttered in a prosodically different way across speakers or styles, e.g. due to a different prosodic focus or rhythmic pattern, this variation ought to be reflected

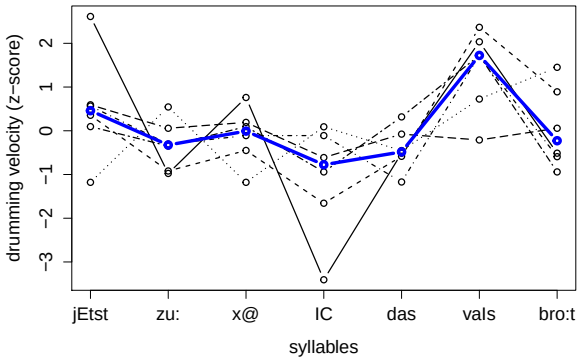


Figure 1: Syllable-based drummed prominence patterns of 6 annotators for the same sentence. The thick blue line illustrates the median drumming velocities ("average drummer")

in prominence impressions and consequently yield different prominence annotations as well. Due to the influence of top-down expectations on prominence perception (cf. section 1), such fluctuations in prosodic structure and style may be somewhat neutralized in perception: to some extent, we "perceive what we expect". It is possible that certain annotation methods cause a stronger or less strong reliance on such top-down expectations than others, e.g. as they induce listeners more to rely "on their inner voice". In order to test whether drummed or conventional prominence annotations behave differently or similarly in this respect, we compared the extent to which annotations varied across identical sentences produced by different speakers. If the balance of top-down expectation and bottom-up processing is similar in both annotation procedures, then a set of (orthographically) identical sentences perceived as prosodically rather different with a conventional annotation method, should also yield in a stronger perceptual variation in a drumming task and vice versa.

2.1. Methods

The data used in the drumming task contained a set of 20 German sentences, each of which was produced by three different native speakers, i.e. 20 sentence triples. Within each triple, the individual productions are likely to differ to some extent, e.g. due to different reading styles or different linguistic interpretations of the read material. These variations ought to be reflected in the prominence annotations. To test whether the two compared annotation methods (drumming and conventional auditory prominence perception) indeed measure a very similar quality of prominence, we calculated the intra-sentential ICC-variability of "drummed" and "perceived" prominence patterns within the three productions in each triple. The "drummed prominences" are based on median velocities of 6 individual drummers ("average drummer"), the perceived prominence was based on the fine-grained (31 levels) median prominence annotations of 3 prosodic experts ("average expert listener"). As the conventional (expert) annotations were only available on syllable level, we also used syllable level drummed annotations for the comparison. All analyses were carried out with the help of the irr-package available for the statistical software package "R" [14].

2.2. Results

When plotting the intra-sentential ICC statistics based on variability in both perception and drumming, it becomes evident that the perceived variability is indeed similar across perceptions in both modalities, albeit not perfectly aligned (cf. Figure 2). A correlation analysis confirms this visual impression ($cc = 0.62, df = 18, p < 0.01$). The ICCs are significantly higher ($t = 4.6$, $df = 37.2$, $p < 0.0001$) for the conventional annotation method ($M = 0.78$, $SD = 0.15$) compared to the drumming method ($M = 0.55$, $SD = 0.17$).

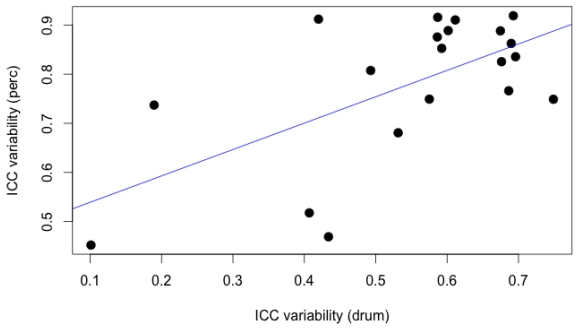


Figure 2: Relationship between conventionally perceived (y-axis) and drummed (x-axis) intra-sentential prosodic variability (ICC)

2.3. Discussion

The analyses reveal that if the same sentence uttered by different speakers receives similar prominence annotations in the drumming task, this is likely to be the case in the prominence annotation task as well and vice versa. This supports the assumption that prosodic variation on the is taken into account by both approaches similarly. However, compared to the conventional annotation, the drumming task yields overall a stronger variability across orthographically identical sentences. This may be interpreted tentatively in such a way that drumming is guided less strongly by top-down expectations, as its annotations are comparatively less uniform across linguistically identical sentences. However, a likely alternative explanation for this finding may be that the stronger overall variability in drumming is caused by the naïve annotaters who may be less stable and following more individual strategies than trained experts. At this point, further conclusions are difficult to make, especially as the analyses rely on comparatively few data points.

3. Identifying listener strategies in "prominence drumming"

In our previous study [1] we found that the drumming approach to prominence annotation reveals much individual variation. We stated above that this individual variation may provide a window to unveil individual strategies of prosody processing. In this section, we want to find out whether individual listener profiles can be estimated based on the drummed annotations. As these strategies may differ depending on the level of linguistic prominence annotation, we will analyze both syllable-based and word-based prominence drumming. These analyses will reveal

whether there are listeners who are guided more by bottom-up or top-down strategies than others. Also, it may reveal the influence of contextual cues to prominence perception.

3.1. Methods

60 German sentences produced by 3 speakers and annotated by 12 listeners (6 syllable drummers, 6 word drummers) serve as material for our study. For each annotator, we trained a Random Forest regression tree to model the prominence annotations (= drumming velocities) based on a set of acoustic, linguistic and contextual prediction variables that have been shown to influence the impression of prosodic prominence (cf. Table 3.1). For both the syllable and the word drumming, an identical set of predictor variables was chosen, with the following exceptions: The syllable's stressability was not used to predict word drumming velocities, as each German word contains at least one stressable syllable, making this feature redundant. The predictor variable "Clash" refers to the accentuation status of the previous syllable in the syllable-based annotations, but to the previous word in the word-based annotations. Training and subsequent analysis was carried out with the *randomForest* R-package [14] using the standard settings and 3000 training cyles. In order to weigh the impact of the individual parameters on the drumming velocities, the importance for all predictor variables was computed as part of the training procedure. This importance measure captures the mean decrease in classification accuracy (MSE) after the predictor variable has been permuted across all trees.We use importance measure (z-score normalized) to weigh all predictor variables' influence on the dependent variable "drumming velocity".

prominence correlate	**description**	**type**
F0	a normalized value between 0 (F0-min) and 1 (F0-max))	acoustic
SyllDur	syllable duration	acoustic
POS	part-of-speech, lexical class	linguistic
Schwa	phonological stressability status of syllable	linguistic
Clash	pitch accentedness of previous syllable/word	contextual
AccentDist	distance to previously pitch accented syllable (in syllables)	contextual

Table 1: Overview of predictor variables the RandomForests are trained on. In the word drumming task, the syllable-based measurements relate to the lexically stressed syllable.

3.2. Results for Syllable Drumming Task

When comparing the importance of the various predictor variables across all drumming annotators, it becomes evident that they relied predominantly on the F0 excursion when modulating their drumming velocity. However, the remaining prominence cues were used to different degrees. Two out of five annotators used syllable duration as the second most important cue to modulate drumming strength, while for two others, POS-based information was the second-best predictor. The more "duration oriented" drummers can be regarded as being slightly more guided by bottom-up cues, the "POS-oriented" drummers as more guided by top-down cues. One of the "POS-drummers" used POS information to an equal degree as F0-based information, thus relying rather heavily on linguistic information. Contextual (Clash, AccentDist) and phonological (Schwa) information was used somewhat less by most annotators. However, one drummer relied mostly on these contextual cues in combination with local F0 height and but paid little attention to duration or POS information. The overall prediction accuracy for the individual drumming behavior based on the chosen predictor variables differs vastly, and explaines practically none of the behavioral drumming variance for two anntotators (2 and 5), while explaing more than a third of the drumming variation in annotator 4. When pooling the individual annotators' impressions to an "average annotator" (cf. Section 2), the RandomForest model is able to account for 47% of the variance in the velocity data. Not surprisingly, the limited set of variables taken into account in our study is obviously not entirely sufficient to account for the collected velocity data, especially when based on a small data set of 60 sentences. An overview of the results is presented in Table 3.2

annotator	*F0*	*SyllDur*	*POS*	*Schwa*	*Clash*	*Accent-Dist*	*% Variance Explained*
1	**72**	19	11	**37**	27	**29**	14
2	**84**	**46**	**50**	18	23	26	26
3	**33**	**21**	**20**	17	10	10	0
4	**83**	**47**	**55**	41	37	27	34
5	**39**	**26**	**23**	2	13	4	1
6	**68**	**34**	**67**	11	17	24	21
average annotator	**100**	**71**	**77**	34	33	39	47

Table 2: Importance (%) of the various predictor variables per annotator in the syllable drumming task. The three most important predictor variables per annotator are shown in boldface.

3.3. Results for Word Drumming Task

The results show comparatively more individual variation than the syllable drumming. Some of the annotators rely predominantly on fundamental frequency excursion, others more on duration, linguistic cues such as lexical class or contextual cues such as the distance to the previous accented syllables. Annotators 2, 3 and 5 relied heavily on a combination of duration, lexical class and context, annotators 1, 4 and 6 on the combination likewise favored by the syllable annotators: F0 excursion, duration and lexical class. As for syllable drumming, the overall prediction accuracy for the individual drumming behavior based on the chosen predictor variables differs vastly, and explaines practically none of the behavioral drumming variance for two anntotators (4 and 6), while explaing up to two thirds of the drumming variation in the remaining annotators. When pooling the individual annotators' impressions to an "average annotator" (cf. Section 2), the RandomForest model is able to account for 67% of the variance in the velocity data, which is considerably more than what was achieved for the syllable data. Interestingly, for the average annotator, the predictor variable "clash" reaches very high importance, while it plays little role for the individual annotators. However, in this condition all predictor variables have a stable and important influence on the drumming velocity result. Not surprisingly, the limited set of variables taken into account in our study is obviously not entirely sufficient to account for the collected velocity data, especially when based on a small data set of 60 sentences only. An overview of the results is presented in Table 3.3

annotator	*F0*	*SyllDur*	*POS*	*Clash*	*Accent-Dist*	*% Variance Explained*
1	**59**	**48**	**62**	22	22	44
2	39	**61**	**60**	8	**55**	58
3	37	**49**	**64**	8	**39**	36
4	**14**	**14**	**12**	11	10	0
5	29	**41**	**46**	24	**54**	29
6	**34**	**12**	**15**	0	4	7
average annotator	**97**	77	**91**	**87**	80	76

Table 3: Importance (%) of the various predictor variables per annotator in the word drumming task. The three most important predictor variables per annotator are shown in boldface.

3.4. Discussion

In line with the previous analyses, the syllable drumming shows higher inter-annotator variation and a less clear correspondence with well-established top-down, bottom-up or contextual correlates of prosodic prominence. In both word- and syllable drumming, listeners blend both top-down and bottom-up cues when modulating drumming strength. F0 appears to be the strongest predictor of drumming velocity in the syllable drumming task across annotators, while lexical class and syllable duration appear are reliable cues across annotators in the word drumming task. When drumming syllables, context cues seem less influential for most annotators compared to word drumming. Other than syllable drumming performance, word drumming can be explained quite well based on a very small data set and a limited set of acoustic, linguistic and contextual predictor variables. This may be partly due to the fact that out of 6 word drumming annotators, 2 (annotators 1 and 2) had prosodic training. The non-experts' performance variability is explained in a similar range as the syllable drumming performance. Interestingly, those annotators whose performance was explained least by the set of predictor variables were also the ones with the least inter-annotator agreement (annotators 3 and 5 for syllable drumming, annotator 4 for word drumming, [1]). This seems to support the fact that inter-annotator agreement can be traced to well-established cues of prosodic prominence. Interestingly, the word drumming task shows considerably more individual strategies, which may be a consequence of the stronger cognitive processing necessary for this task [1]: In word drumming, annotators deliberately choose to rely on an individual set of cues in order to fulfill the task, while the more intuitive syllable drumming is guided by similar cues across annotators, despite them showing more individual variability.

4. General Discussion

We found that drumming-based annotation method reflects prosodic variability present on the signal level similarly as more conventional prominence annotations. This is encouraging as it indicates a comparability of research results gathered with two rather different methods. With respect to the investigation of listener strategies, we feel that our method appears to be fruitful and could verify the importance of already well-established cues to prosodic prominence. Also, the models show that most listeners rely on a blend of top-down and bottom-up cues in their prominence interpretation. Interestingly, the word level prominence drumming revealed more individual strategies compared to the syllable-based method, perhaps pointing to a higher degree of linguistic awareness and less intuition. It is difficult to say at this point which method (intuitive syllable drumming, linguistically informed word drumming) is most adequate to get to the core of prominence processing in everyday communicative interaction. For a fuller understanding of individual listening strategies, further established prominence correlates (e.g. information structure, predictability, phrasal position, F0 shape, intensity, spectral emphasis) have to be included in the models as a next step, and more data needs to be annotated.

5. Bibliography

[1] B. Samlowski and P. Wagner, "Promdrum — exploiting the prosody-gesture link for intuitive, fast and fine-grained prominence annotation," in *Proceedings of Speech Prosody 2016*, 2016, p. p5.06.

[2] P. Wagner, Z. Malisz, and S. Kopp, "Speech and gesture in interaction: an overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[3] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, "Spatiotemporal coupling between speech and manual motor actions," *Journal of Phonetics*, vol. 42, pp. 1—11, 2014.

[4] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation based factors in the perception of prosodic prominence," in *Journal of Laboratory Phonology*, vol. 1, 2010, pp. 425–452.

[5] J. Terken, "Fundamental frequency and perceived prominence," *Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[6] S. Baumann and C. T. Röhr, "The perceptual prominence of pitch accent types in german," in *Proceedings of ICPhS 2015*, Glasgow, Scotland, 2015.

[7] K. de Jong, "The supraglottal articulation of prominence in english: Linguistic stress as localized hyperarticulation," *Journal of the Acoustical Society of America*, vol. 97, pp. 491–504, 1995.

[8] M. Vainio and J. Järvikivi, "Tonal features, intensity, and word order in the perception of prominence," *Journal of Phonetics*, vol. 34, no. 3, pp. 319–342, 2006.

[9] C. Widera, T. Portele, and M. Wolters, "Prediction of word prominence," in *Proceedings of Eurospeech*, vol. 2, Rhodes, Greece, 1997, pp. 999–1002.

[10] C. Gussenhoven and A. Rietveld, "Fundamental frequency declination in dutch: testing three hypotheses," *Journal of Phonetics*, vol. 16, pp. 355–369, 1988.

[11] D. Arnold, P. Wagner, and H. Baayen, "Using generelized additive models and random forests to model prosodic prominence in german," in *Proceedings of Interspeech 2013*, 2013, pp. 272–276.

[12] A. Eriksson, G. Thunberg, and H. Traunmüller, "Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing," in *Proceedings of EUROSPEECH*, Aalborg, Denmark, 2001, pp. 399–402.

[13] P. Wagner, "Great Expectations - Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates," in *Interspeech 2005, September, 4-8, Lisbon, Portugal*, 2005, pp. 2381–2384.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[15] P. Wagner, A. Origlia, C. Avesani, G. Christodoulides, F. Cutugno, M. D'Imperio, D. Escudero Mancebo, B. Gili Fivela, A. Lacheret, B. Ludusan, H. Moniz, A. Ní Chasaide, O. Niebuhr, L. Rousier-Vercruyssen, A. C. Simon, J. Simko, F. Tesser, and M. Vainio, "Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland, 2015.

Teasing apart lexical stress and sentence accent in Hungarian and German

Ádám Szalontai[1], *Petra Wagner*[2], *Katalin Mády*[1], *Andreas Windmann*[2]

[1]Research Institute for Linguistics, Hungarian Academy of Sciences
[2]University of Bielefeld
{szalontai.adam|mady.katalin@mta.nytud.hu}, {petra.wagner|awindmann2}@uni-bielefeld.de

Abstract

This study compares the strategies to mark lexical stress and sentence-level accent in Hungarian and in German by employing two production experiments of comparable designs. The experimental conditions elicited target segments in +/– stressed and +/– accented conditions. The results indicated that while German, a language with variable lexical stress placement, clearly marks both stress and accent with a number of phonetic parameters, Hungarian, a language with fixed word-level stress placement marks accents, but not stress.

Index Terms: lexical stress, accent, Hungarian, German, prominence marking

1. Introduction

Lexical stress and sentence accent are marked on the same segments (syllables), however they are associated with different levels of prosodic structure. The parameters (e.g. intensity, duration) that are available for a language for prominence marking apply to both prominence levels, therefore it can be assumed that there is some degree of difference in the quantity of these parameters when they are employed to mark lexical stress and sentence accent. This difference might be linked to the prosodic system of a given language, where redundancies and marking necessities might play an important role in influencing what categories receive marking and to what degree. The present study aims to answer this question by comparing acoustic cues of stress and accent in Hungarian and German, two languages with different prosodic systems.

1.1. Hungarian and German stress and accent systems

Word-level stress in Hungarian is highly predictable: stress is always assigned to the initial syllable of a prosodic word. There is no evidence for secondary stress [1]. Stress has been shown to effect intensity [2], but it has not been shown to have a considerable effect on lengthening [2, 3, 4]. The lack of a lengthening effect might be due to the presence of a vowel quantity distinction in the language. While word-level stress shift is possible, it is linked to very specific conditions: the segment where stress is shifted to needs to be contrasted.

German behaves differently from Hungarian in this respect. Lexical stress is not strictly assigned to a given syllable, instead, it is restricted to a 3-syllable window (e.g. **Li**.ba.non 'Libanon', Ba.**na**.ne 'banana', E.le.**fant** 'elephant') [5]. Word-level stress placement is marginally contrastive, a property that was exploited in the creation of the target stimuli as shown in (1) to (4), showing how stress placement differentiates between the name ***Au****gust* and the month *Au****gust***.

Prominence is associated with sentence-level (i. e. pitch) accents in Hungarian broad focus sentences on each content word or syntactic constituent, while in sentences containing a narrow focus, the focused item receives the highest prominence while the following items are deaccented. The focus occurs in a specific syntactic position. The degree to which the accent manifested on the focus is more prominent than non-focus accents in broad focus sentences has been debated with some evidence suggesting larger f0 range [6], with other studies not finding significant differences [7].

German has a large variety of pitch accents as well as a nuclear accent [8], where the focused item usually co-occurs with the nuclear accent. In these cases there is post-focal compression. Different accent shapes might be associated with different levels of prominence [9].

1.2. Motivation and goals

The present study aims to differentiate lexical stress and sentence accent marking in Hungarian and German. By working with comparable experimental paradigms in the two languages, it is also our aim to show what, if any, differences exist in the prominence marking of these languages with different prosodic systems.

2. Methods and materials

A production experiment was conducted to elicit target syllables that varied in their assignment of lexical stress and accent. In German the following four combinations of +/– stress and +/– accent conditions were created:

+lexical stress, +sentence accent

(1) Um den Garten wird sich der alte ***Au***gust kümmern.
'The garden will be taken care of by old August.'

+lexical stress, –sentence accent

(2) Um den Hund wird sich **nicht** der alte *Au*gust kümmern.
'The dog will not be taken care of by old August'

–lexical stress, +sentence accent

(3) Zurück werde ich wohl Mitte **Au*gust*** kommen.
'I will probably come back by Mid-August.'

–lexical stress, –sentence accent

(4) Vielleicht werde ich aber auch erst **Ende** Au*gust* kommen.
'Perhaps I will come back no earlier than end of August.'

In the above example the target syllable is underlined, the position of lexical stress is indicated by *italics* and the position of sentence-level accent is in **bold**. The +/- stress conditions were created by making use of different stress placements on the words *August*, the name, and *August*, the month, to modify the placement of lexical stress. In the case of sentence accents, the -conditions were created by placing the target syllable under the scope of negation as in (2), or by shifting focus to a preceding word as in (4).

In Hungarian not all possible variations of factors is possible due to syntactic reasons, thus, only three conditions could be created:

+lexical stress, +sentence accent

(5) **Jól** ***lo***csold meg a **mus**kátlit.
'Water the geraniums well.'

+lexical stress, –sentence accent

(6) **Sem**miképp **ne** *lo*csold meg a **kak**tuszt.
'In no way should you water the cactus.'

–lexical stress, +sentence accent

(7) **Ne**hogy **meglocsold** az **or**chideát.
'Don't water the orchid ever.'

The target syllable and the placement of stress and accent are indicated as above. The target syllable was always the first syllable of a verb. In Hungarian neutral clauses the verb is most often the first element of a syntactic unit which coincides with a prosodic phrase, its first syllable therefore receives a pitch accent by default. Other pitch accents that are present in the sentence are also indicated. These are usually assigned to syntactic phrases. In the sentence (7) a verbal modifier is placed in front of the verb, forming one prosodic unit with it. In this configuration lexical stress is assigned to the verbal modifier as it now contains the first syllable. The presence of sentence-level accents was manipulated by placing the verb under the scope of negation as in (6). In this sentence the pitch accent is shifted from the verb to the negation word.

The experimental paradigm aimed to introduce a degree of communication between the two participants. They were asked to imagine a scenario where one of them is going on holiday, and the other will stay in the apartment. The task involved giving instructions to the friend on what to do and what not to do in the flat. Participants were presented with slides containing images and unconjugated/morphologically unmarked words that formed the target sentences as in Figure 1 with two examples from the German stimuli. Participants were then asked to say aloud the instructions, while the dialogue partner was asked to remember as many of the instructions as possible.

Recordings were made in sound-proof rooms in Budapest and in Bielefeld, using SpeechRecorder [10] and head-mounted microphones. For both languages there were 7 target words with 2 factors (accent and stress). Stimuli were presented in randomized order together with filler sentences. There were two repetitions for each item. 30 German and 12 Hungarian native speakers participated in the experiment.

The following parameters based on the target vowel were investigated: duration, intensity, spectral balance, and f0 maximum.

Results were analyzed using linear mixed effect models with stress and accent as fixed effects and speaker and item as random effects. In order to account for the unbalanced design of the Hungarian data, tests were run separately for all conditions, i.e. the comparison between stressed and unstressed syllables in accented words was separated from the comparison between stressed and unstressed syllables in deaccented words.

Figure 1: *Stimuli to elicit the sentences given in (1) (top) and (3) (bottom).*

3. Results

3.1. Duration

Duration was measured on the vowel of the target syllable fore each language. The results are presented in Figures 2 and 3.

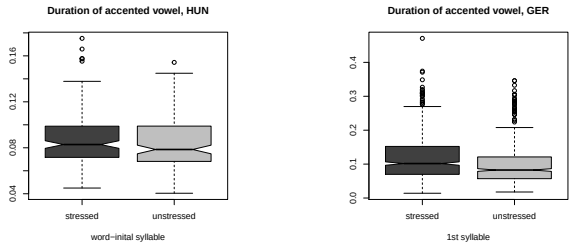


Figure 2: *The effect of stress on the duration of the accented vowel in Hungarian (left) and in German (right).*

The statistical analysis revealed that duration was significantly effected by word-level stress placement and by accent placement on the target syllable in both languages.

3.2. Intensity and spectral balance

The analysis of intensity was done on maximum intensity values extracted from the vowel in the target syllable. The results are presented in Figures 4 and 5.

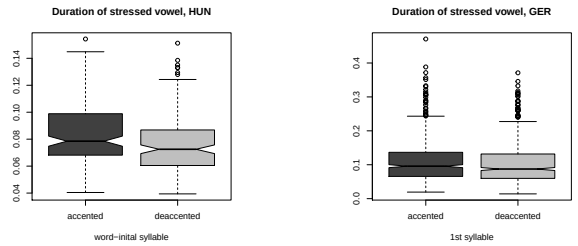


Figure 3: *The effect of stress on the duration of the stressed syllable in Hungarian (left) and in German (right).*

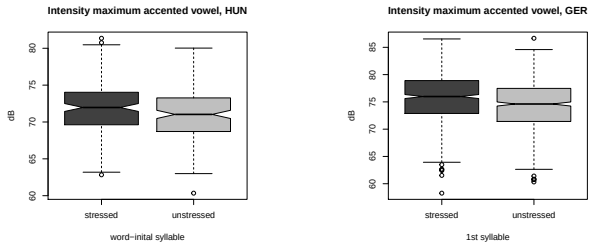


Figure 4: *The effect of lexical stress on the intensity maximum on the accented syllable in Hungarian (left) and in German (right).*

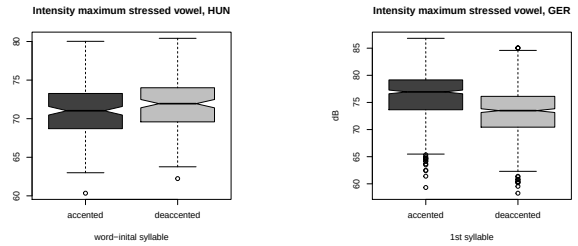


Figure 5: *The effect of accent on the intensity maximum on stressed syllable in Hungarian (left) and in German (right).*

Statistical analysis revealed that intensity was significantly different in the case of German for both stress and accent factors. In the case of Hungarian only accent lead to significantly different intensity levels, but not stress.

Another parameter investigated was spectral balance. Spectral balance (SPLH–SPL) is calculated by subtracting the sound pressure level (SPL) from the sound pressure level at high frequencies (SPLH) of the same segment. This parameter has been shown to be a good indicator of vocal quality and prominence [11]. The plots below show the results of the SPLH-SPL data in Figures 6 and 7. s

Statistical analysis revealed that differences were significant for German both for the effects of stress and accent. However, in Hungarian only the effect of accent showed significant differences, lexical stress did not, as in the case of intensity maximum.

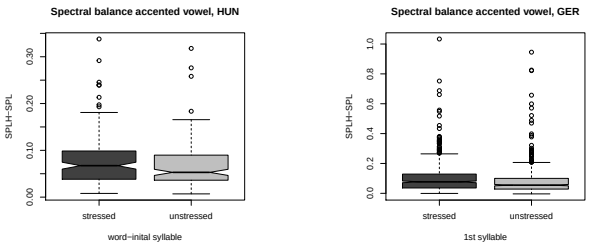


Figure 6: *The effect of lexical stress on spectral balance on accented syllable in Hungarian (left) and in German (right).*

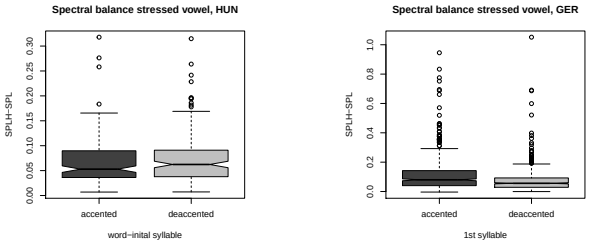


Figure 7: *The effect of accent on spectral balance on stressed syllable in Hungarian (left) and in German (right).*

3.3. F0 maximum

F0 maximum Hz values were extracted from the sound files, and values where converted to semitones using speaker specific median values as baselines. The results are presented in Figures 8 and 9.

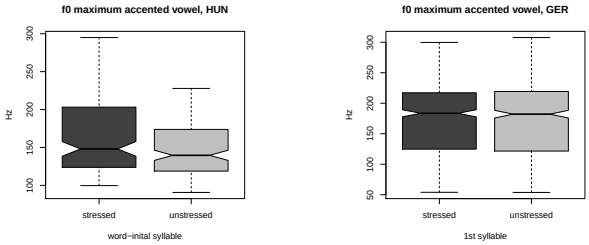


Figure 8: *The effect of stress on the f0 maximum on accented syllable in Hungarian (left) and in German (right).*

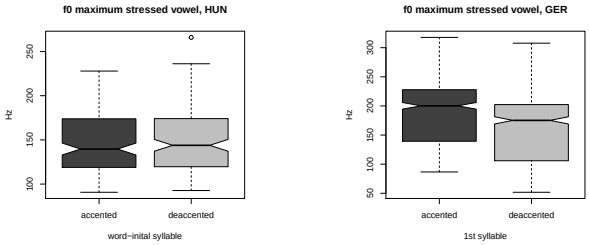


Figure 9: *the effect of accent on the f0 maximum on stressed syllable in Hungarian (left) and in German (right).*

As in the case of intensity maximum and spectral balance, German showed significant effects for both stress and accent while Hungarian only showed them for accent. It should be noted that for 1st syllable target syllables in German the factor lexical stress did not show a significant difference for f0, however it did for syllables in the second position.

4. Discussion and conclusions

We have shown that German and Hungarian mark lexical stress and sentence accent in different ways. Higher-level prominence marking by pitch accents is present in both languages. German also marks lexical stress across all parameters observed in this study, while Hungarian does not. We did find a difference in duration as an effect of lexical stress, however, this might be due to the necessity of moving the accented syllable from the initial to a word-medial position. We assume that these results originate from the prosodic systems of the two languages, namely that German lexical stress marking is not predictable, while Hungarian stress marking is highly predictable. Therefore marking differences in prominence in German is a necessity on both levels examined in this study, while for Hungarian it is a redundancy on the level of lexical stress but not when it comes to sentence-level accents.

5. Bibliography

[1] S. Blaho and D. Szeredi, "Secondary stress in Hungarian: (morpho)-syntactic, not metrical," in *Proceedings of the 28th West Coast Conference on Formal Linguistics*, M. B. W. et al., Ed. Somerville, MA: Cascadilla Proceedings Project, 2011, pp. 51–59.

[2] I. Fónagy, *A hangsúlyról [On stress]*, ser. Nyelvtudományi Értekezések 18. Budapest: Akadémiai Kiadó, 1958.

[3] I. Kassai, *Időtartam és kvantitás a magyar nyelvben [Duration and quantity in Hungarian]*. Budapest: Akadémiai Kiadó, 1979.

[4] K. Mády, L. Bombien, and U. D. Reichel, "Is Hungarian losing the vowel quantity distinction?" in *Proc. 8th International Seminar on Speech Production, Strasbourg*, 2008, pp. 445–448.

[5] M. Jessen, "German," in *Word prosody system in the languages of Europe*, H. van der Hulst, Ed. Berlin: Mouton de Gruyter, 1999, pp. 515–545.

[6] S. Genzel, S. Ishihara, and B. Surányi, "The prosodic expression of focus, contrast and givenness: A production study of Hungarian," *Lingua*, 2014.

[7] K. Mády, "Prosodic (non-)realisation of broad, narrow and contrastive focus in Hungarian: a production and a perception study," in *Proc. Interspeech 2015*, Dresden, 2015, pp. 948–952.

[8] F. e. a. Kügler, "Dima – annotation guidelines for German intonation," in *Proc. 18. International Congress on Phonetic Sciences*, Glasgow, 2015.

[9] S. Baumann and C. T. Röhr, "The perceptual prominence of pitch accent types in German," in *Proc. 18. International Congress on Phonetic Sciences*, Glasgow, 2015.

[10] C. Draxler and K. Jänsch, "SpeechRecorder – a universal platform independent multi-channel audio recording software," in *Proc. International Conference on Language Resources and Evaluation*, Lisbon, 2004, pp. 559–562.

[11] G. Fant, A. Kruckenberg, and J. Liljencrants, "Prominence correlates in swedish prosody," in *International Conference of Phonetic Science, San Francisco, USA*, vol. 3, 1999, pp. 1749–1752.

Two-stage Decision Trees for Automatic Speaker Likability Classification

Mathias Walther[1], *Oliver Jokisch*[2], *Taïeb Mellouli*[1]

[1]Department of Business Information Systems and Operations Research,
Martin Luther University Halle-Wittenberg, Germany
[2] Institute of Communications Engineering, Leipzig University of Telecommunications, Germany
mathias.walther@wiwi.uni-halle.de

Abstract

This article discusses a two-stage classification system for paralinguistic speaker traits which is part of a prototypical expert system for rhetorical and vocal quality assessment in call center talks. The system is based on pre-trained models for vocal features and outputs comprehensible classification rules so that the agent can improve his rhetorical abilities. The recognition of vocal features is modeled by competing classification systems and combined into a multi-classifier system which is based on decision trees. We compare two decision tree inducers, namely C4.5 and random forest, both in prediction accuracy and their rule sets. The experiments were conducted with the Speaker Likability Database (SLD) and benchmarked against the results of the Interspeech 2012 Speaker Trait Challenge. In terms of accuracy, the proposed two-stage classification performs similar to the baseline results with the advantage of being introspectable.

Index Terms: decision tree, conversation quality, multi-classifier system, speaker likability, call center

1. Introduction

We have been systematically exploring the criteria of conversation quality using authentic corpora since 2006. Our studies show that conversational quality has different aspects, including paralinguistic features and user states or traits, such as emotionality, authenticity and friendliness [1]. With the aim of a substantiated empirically and theoretically based didactic for professionally operated phone calls on an industrial scale, we rely on a combination of methods from qualitative and quantitative linguistic, phonetic and conversational approaches.

Monitoring of conversational quality is an important part of quality management and a critical success factor for call centers [2]. In Germany, there are 6,800 call centers with 520,000 employees, who hold 25 million conversations per day. Because of the huge amount of calls, a manual monitoring can only consider a very small proportion of talks. From a technical point of view, there is no support system for automatically monitoring conversational quality. Although detection and classification of paralinguistic speech features has moved into research focus over the last years. A wide range of classification models was presented in various areas [3, 4, 5]. Besides a good prediction accuracy, which is reached in some fields, e. g. emotion recognition, the models' biggest drawback is the lack of explanatory power that is needed to measure and analyze the complex phenomena of conversational quality. Due to extensive feature vectors and classification algorithms such as support vector machines or artificial neural nets, models are hard to interpret. These models act like a black-box and do not provide the user with classification rules [6]. Nonetheless, the classification of a user state is sufficient in many use cases like human-computer interaction. For example the main purpose of emotion detection in call centers is to differentiate the callers' state between emotional and non-emotional and route the call either to an agent or to an automatic system [7]. Meaningful models are not necessary in this application.

We intend to build a classification system as a monitoring tool for call center conversations with explainable decisions and possible recommendations based on the decisions. The system under development improves the assessment of conversational quality by aggregation and formalization of expert knowledge. The use case is the following: The agent's voice is recorded and classified during a call in real-time and or as a batch run. This process does not need human interaction and can be executed for a large number of calls. Afterwards the systems' output is analyzed by coach and agent. The high explanatory power of the decision model helps to support both agent and trainer in their effort to improve conversational quality. To address these challenges, a two-stage classification technique is proposed, which is based on pre-trained classification models for speech and vocal features. The main goal is the transformation from the numeric representation in feature vectors to a symbolic annotation, which can be understood by an expert.

In a real-world use case scenario, the system has to classify unknown conversations, which is a challenge to any paralinguistic classification system, since the new speech data differs from the original corpus in several aspects – e. g. there are different speakers, the recording quality might be different and in addition to that even the assessment criteria can vary. These influence factors can decrease classification performance significantly [8]. Hence the cross-corpus validation of the proposed modeling technique is crucial to the application in real-world conversations. In this article we demonstrate and validate our systems' performance on unknown speech data. In the experimental part we use the Speaker Likability Database (SLD) and compare our results to findings of the Interspeech 2012 Speaker Trait Challenge [9].

2. Two-stage Classification

2.1. Overview

The two-stage modeling approach is based on the working hypothesis that conversational quality can be described by perceptual features, which can be recognized by automatic classification systems but also understood by human experts. The perceptive features are expressed as 13 criteria of speech and voice presentation, which have been identified in previous studies [10]. The goal of stage 1 is the creation of base classifiers

for the perceptual features, which are the basis for the models on stage 2. The base classifiers are marked bold in Figure 1. The modeling approach consists of the following steps:

Stage 1: Create base classifier for speech and voice presentation

(a) Extract features from signal,

(b) Train classifiers for every criterion,

(c) Select best model to become the base classifier.

Stage 2: Apply base classifiers to a different corpus

(d) Extract features from signal,

(e) Classify each instance with each of the base models and keep the original class of the instance,

(f) Train a decision tree on the new instances based on symbolic representation.

Please note that stage 2 is independent from stage 1. Stage 1 is executed only once, since in stage 2 the pre-trained models classify a new data set. The trained base classifiers can be applied to different corpora for creating stage 2 models. Their decisions are aggregated to a new instance in a symbolic form.

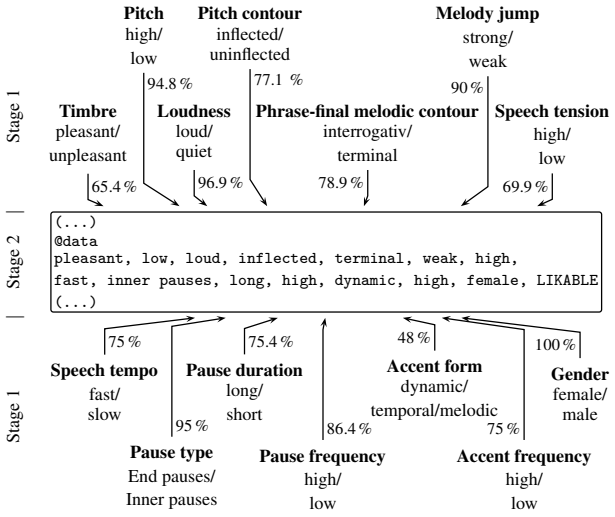


Figure 1: *General structure of the two-stage classification. Stage 1 base classifiers are printed bold.*

2.2. Stage 1 – Base Classifiers

The creation of the base classifiers follows the work flow of paralinguistic speech processing [11]. The models for all criteria are trained on a separate subset of a corpus which was created for research of conversational quality [1]. In the first step (a) features were extracted with openSmile in version 1.0.1 [12]. The feature set is based on the configuration file from the Interspeech 2010 Paralinguistic Challenge [13] and contains 38 low-level descriptors and 21 functionals – including acoustic, spectral and prosodic features. This configuration has been supplemented with five formants and their statistical functionals. In addition, the gender of the agent has been manually labeled. This procedure leads to 2,106 features in total.

Since the accuracy of the whole classification system heavily relies on the accuracy of the base models, the best performing model for each criterion has to be identified. Therefore we tested eight widely used classification algorithms for each criterion of speech and voice presentation in step (b): naive Bayes, Bayes net, logistic model tree, ripper, support vector machine (SVM), Ada boost, C4.5, and multilayer perceptron. The Weka toolkit [14] was used for all experiments. The ranking of the algorithms was conducted via an analysis of variance based on a cross-validation followed by Duncan's test in step (c). A detailed discussion of the statistical methods is given in [15]. The recognition rate (RR) of each base classifier (i. e. of the best model) is shown in Figure 1 next to the arrows. It can be seen that the accuracy of the base models for all dichotomous classes is above 65 %. The best accuracy is achieved on loudness and perceived pitch. Pause type, melody jump and stress frequency show good recognition rates, too. The criterion accent form has an accuracy of 48 % which outreaches 33.3 % – the expected value of randomly guessing three classes. The gender is not recognized by a model, since it is annotated in the corpus. Therefore its RR is 100 %.

2.3. Stage 2 – Combining Models

Stage 2 marks our central concept in the classification system. The high-level model extends the concept of fusion systems by a learning algorithm, which creates a traceable decision tree. The outcome of stage 2 is the combination of the subsystems' decisions. The construction of a classification model for stage 2 starts with any acoustic speech corpus, i. e. with an unknown data set. In the pre-processing step (d), the features are extracted from the speech signal. Since the base models have to be applied on this feature set, it has to be identical to the one used in step (a). The next step (e) refers to the transformation from numeric input (the signal features) to a symbolic representation.

In order to create a symbolic data set from which the second-stage model can be trained, each instance in the data set is classified by every model for speech and voice presentation. The decisions of all 13 base classifiers for a segment in the corpus form an instance of the symbolic data set. An example of the transformed SLD in Weka's data format is shown in the inner box of Figure 1. Both class and gender, *Likable* and *female*, are known from the SLD's meta data. The final data set (including all transformed instances) is used to train the classifier in stage 2.

Due to its structure the described data can be input to all classification algorithms that can handle nominal attributes. Since the main objective of the modeling technique is comprehensibility, rule or tree inducing algorithms are preferred.

The basic concept of decision tree learning is a recursive divide-and-conquer process [6]. At first, an attribute is selected and placed in the root node. From the root node a branch is created for each value of the attribute. The goal of these so called splits is to find the best separation for a given class distribution. The algorithm continues until all instances in a leaf belong to the same class and no further splits are necessary. A full grown tree is likely too complex and trends to overfit the data. To avoid that, different strategies of pruning can be applied. Their goal is to cut and aggregate branches to reduce complexity and improve the ability to generalize [16]. Many different tree learning algorithms have been developed over the years. They share the basic concept but differ in their evaluation criteria for splits and pruning methods. For our two-stage models, two decision tree algorithms are tested on the SLD: C4.5 [17] and random forest (RF) [18].

C4.5 is popular in different areas of machine classification, since it is considered accurate and fast [19]. It uses the entropy

based information gain, which is founded in information theory, for selecting the split attributes and an error-based pruning algorithm [17].

Random forest is composed of several individual unpruned decision trees. In contrast to other tree inducers such as C4.5, the splits on the nodes are constructed randomly. Due to the randomness, a single tree might be less accurate than those constructed by more sophisticated algorithms. To compensate for that, the trees' decisions are combined by a majority vote for the final classification. RF is configurable in three aspects: the number of trees, the number of features that can be used for a tree and the maximum depth of the trees. The following learning algorithms and configurations were tested in the experiments:

C4.5: Standard algorithm and configuration as implemented in Weka (J48) [16].

RF Configuration 1: 10 trees with unlimited depth and at most 4 attributes. This is Weka's standard configuration.

RF Configuration 2: 10 trees, maximum depth as well as number of attributes is limited to 3.

3. Cross-corpus Validation

3.1. Test Database

In the previous work, we have validated our system with standard techniques like cross-validation. The experiments showed good results [20, 10]. However, with the aim of building a flexible classification system, the two-stage classification technique must perform well in different scenarios. To prove its universality, tests with other speech corpora have to be performed. The cross-corpus validation is executed with the Speaker Likability Database (SLD) [21], which was chosen for the following reasons:

- Speaker likability is part of the complex phenomenon conversational quality and thus highly relevant for call center communication.
- SLD is labeled in two classes, which correspond to the proposed modeling approach.
- The relatively large size of the data set is helpful for statistical validation.
- SLD can be used for benchmarking our results with other state-of-the-art classification methods.

The SLD is part of the "Agender Database" [22], which was created for research in automatic detection of age and gender in German phone calls. The corpus was recorded in conventional telephone quality (8 kHz sampling rate) and consists of different utterance types, ranging from single words, e. g. commands, dates or weekdays to sentences [23]. The data is divided into three partitions: training, development and test. Table 1 shows the partitioning of the SLD. All models are built on a combination of training and development. The test partition is kept for calculating the models' performance.

3.2. Baseline Results

Table 2 gives an overview about the classification accuracies in 16 studies referring to SLD. It includes 10 contributions to the "Interspeech 2012 Speaker Trait Challenge" and three other studies [24, 25, 23]. The main performance measurement is the unweighted average (UA), which has been used in several other

Table 1: *Partitions of the SLD.*

Class	**Training**	**Development**	**Test**	$\sum$
Likable (L)	189	92	119	400
Non-Likable (NL)	205	86	109	400
$\sum$	394	178	228	800

Interspeech challenges. The UA is the average recall of the two classes [9]:

$$UA = \frac{\text{Recall}(positive) + \text{Recall}(negative)}{2} \quad (1)$$

Two results were set as baseline for the challenge: An SVM achieved $UA = 0.559$, and a random forest classifier scored $UA = 0.59$ on the test partition. In Table 2, the UA on test partition ranges from 0.518 [26] to 0.687 [23]. The best UA is given by a combined classifier with three contributors' models. For detailed summary of the challenge see [23]. In the scenario of quality detection, the class distribution is regarded as even. Hence there are no mis-classification costs, and no classes should be preferred by the classifier. Therefore we examine – in addition to the UA – the true-positive rate (TPR) and the true-negative rate (TNR), i. e. the recall for both classes.

Table 2: *Comparison of classification performance (UA) on SLD.*

Rank	**Reference**	**Training**	**Test**
1	Schuller et al. [23]	–	0.687
2	Montacié and Caraty [27]	–	0.658
3	Brueckner and Schuller [28]	0.570	0.640
4	Gonzalez and Anguera [25]	0.632	0.622
5	Liu and Hansen [24]	0.616	–
6	Buisman and Postma [29]	0.742	0.614
7	Pohjalainen et al. [30]	0.620	0.613
8	Lu and Sha [31]	0.621	0.601
9	Wu et al. [32]	0.686	0.595
10	Schuller et al. [9]	0.576	0.590
11	Hewlett Sanchez et al. [33]	–	0.582
12	Attabi and Dumouchel [34]	–	0.565
13	Schuller et al. [9]	0.585	0.559
14	Cummins et al. [26]	0.646	0.545
15	Anumanchipalli et al. [35]	0.594	0.540
16	Cummins et al. [26]	0.683	0.518

4. Experimental Results

4.1. Classification

The C4.5 tree achieves an average $UA = 0.575$ on the combined training and development set with ten-fold cross-validation. In contrast, the UA is as low as 0.515 for the test set. Nearly all studies in Table 2 achieve better results on training than on test set. Large differences between training and test can be seen on some contributions to the challenge, e. g. [29, 26, 32]. Those differences may arise from the data structure of the test set. Decision trees are particularly sensitive to changes in the data set [36].

On C4.5 algorithm, we observe both during cross-validation and on the test set a bias to the positive class *Lik-*

able (L) with $TPR = 0.672$. On the other hand, scoring $TNR = 0.358$, the error for *Non-Likable (NL)* is much higher.

The two configurations of random forest show divergent results. Weka's standard configuration with unlimited trees performs in cross-validation significantly worse than C4.5 ($UA = 0.510$). In contrast, UA on the test set is 0.548, which is better than C4.5. Analysis of TPR and TNR shows that class *Likable* is as biased as in C4.5. The second configuration of RF reaches a higher accuracy in training phase ($UA = 0.549$). On test set, the random forest with configuration 2 scores with $UA = 0.556$, which refers to rank 14 in Table 2. In contrast to the other models, both C4.5 and configuration 1, TNR and TPR are above 0.5. For the test set, $TPR = 0.571$ and $TNR = 0.541$ are achieved.

4.2. Decision Tree Analysis

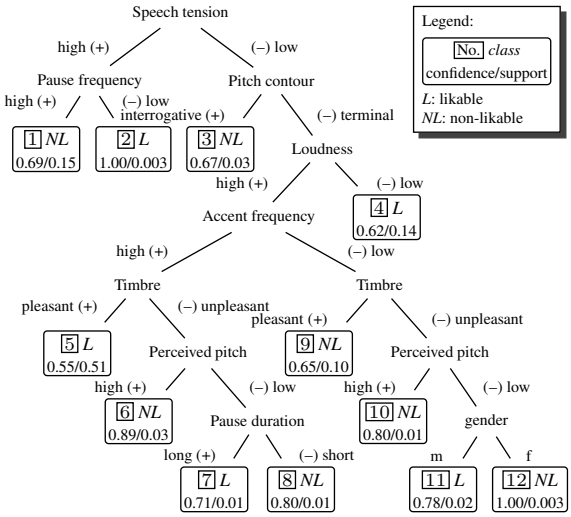


Figure 2: *C4.5 decision tree – learned from training and development partition.*

A benefit of decision trees is the ability to transform the branches and leafs into decision rules [6]. In contrast to other classification techniques, decision trees and thus rules can be analyzed in detail. This characteristic feature of tree inducing algorithms is used in our paper for the model analysis. During transformation the leaves become the target classes, and the branches create decisions upon the attributes' values. Accordingly, the sample tree in Figure 2 can be transformed into 12 decision rules. Figure 3 shows rules 1 and 4 from the tree in a pseudocode annotation. For each rule, individual quality measures can be calculated. Figure 2 shows confidence and support for each leaf, which are concepts from association rule mining. The support is the proportion of all instances that are covered by the antecedent of the rule in the training set, whereas confidence is the number of all instances which are correctly classified by the rule divided by the number of all instance on which the rule can be applied. Hence the confidence equates to the accuracy of the rule [6].

```
1: NL := high speech tension AND high pause
   frequency

4: L := low speech tension AND terminal pitch
   contour AND low loudness
```

Figure 3: *Transformed rules 1 and 4.*

The major finding in our analysis is the high support of the rules 1, 4, 5 and 9. These rules are valid for 90 % of all instances in the training and development partition, which is altogether 572 instances (Table 1). The most universal rule 5 applies to 51 % of the instances. It has a confidence of 55 %, which is the lowest of all rules. The eight remaining rules can be applied to 10 % of the data set. Having a confidence ≥ 0.67, they are more accurate. From our analysis the conclusion can be drawn that most of the rules are covering special cases.

Both, analysis of the C4.5 tree and the RF models suggest that the short rules from small trees are better classifiers than long ones. To emphasize this hypothesis, the relation between tree size and accuracy can be examined. Figure 4 shows the relation between the number of rules per tree and the corresponding average UA on the data from the cross-validation of the C4.5 learner. The best average UA is achieved by trees that have 12 rules. Accordingly, trees with less than 11 rules and more than 15 rules score to a lower UA during the cross-validation. Our results can be interpreted as follows: Smaller trees are to prone to underfitting, whereas complex trees with many rules tend to overfitting [37]. The tree presented in Figure 2 seems to be a good compromise regarding its size, so that the model is likely to reach a good accuracy on new data.

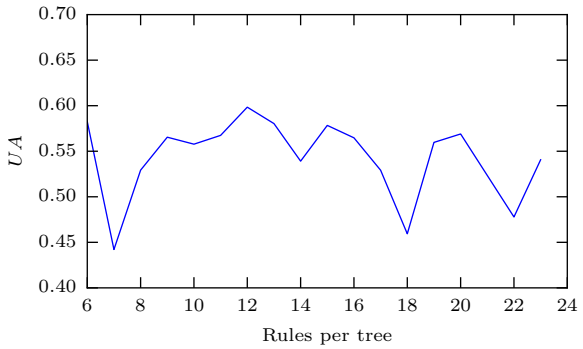


Figure 4: *Relation between tree size and classification accuracy for C4.5.*

5. Conclusion

The analysis and optimization of industrial omni-channel-communication – especially of conversational skills – provide a highly relevant field of application. We introduced a two-stage classifier system for speaker traits that is based on vocal and rhetorical features of the voice. We showed that vocal and rhetorical features, which are important indicators for conversation quality in call centers, can be detected with classification algorithms.

Due to the main goal of the target system – classifying speech data from different sources – we used the well-known Speaker Likability Database for cross-corpus validation. We tested different decision tree inducers and showed that the results are on par with contributions to the Interspeech 2012 Speaker Trait Challenge. The main advantage of the proposed novel system is comprehensibility. Hence a didactic expert, e. g. a call center coach, can understand the rules and is able to examine special cases among the talks/agents and their classification. Based on that analysis the coach can implement training meth-

ods to teach the agent how to improve his speech quality, e. g. perceived likability.

In our future research, we will focus on the build process of the classifiers, both base classifier and second stage systems. Considering the accuracy of the rule set, which mainly depends on the performance of the base classifiers, a further goal is the improvement of their classification accuracy.

6. Bibliography

[1] S. Meißner, J. Pietschmann, M. Walther, and L. Nöbel, "Innovative IT-gestützte Ansätze zur Bewertung der Gesprächsqualität in Telefonverkaufsgesprächen," in *Erforschung und Optimierung der Callcenterkommunikation.*, U. Hirschfeld and B. Neuber, Eds. Berlin: Frank & Timme, 2011, pp. 195–214, in German.

[2] K. Dawson, *The Call Center Handbook: The Complete Guide to Starting, Running, and Improving Your Call Center*, 5th ed., ser. Call Center Handbook. San Francisco: CMP Books, 2003.

[3] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 1st ed. New York: John Wiley & Sons, 2014.

[4] B. Schuller, S. Steidl, and A. Batliner, "Introduction to the special issue on paralinguistics in naturalistic speech and language," *Computer Speech & Language*, vol. 27, no. 1, pp. 1–3, 2013.

[5] B. Schuller and F. Weninger, "Ten recent trends in computational paralinguistics," in *4th COST 2102 International Training School on Cognitive Behavioural Systems*, ser. Lecture Notes on Computer Science (LNCS), A. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds. Berlin: Springer, 2012, pp. 35–49.

[6] F. Gorunescu, *Data Mining: Concepts, models and techniques*. Berlin: Springer, 2011.

[7] J. Pittermann and A. Pittermann, "Integrating emotion recognition into an adaptive spoken language dialogue system," in *Intelligent Environments, 2006. IE 06. 2nd IET International Conference on*, vol. 1. IET, 2006, pp. 197–202.

[8] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments," in *Proc. of the Third International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, pp. 77–82.

[9] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The interspeech 2012 speaker trait challenge," in *INTERSPEECH 2012*, 2012, pp. 254–257.

[10] M. Walther, B. Neuber, O. Jokisch, and T. Mellouli, "Towards a conversational expert system for rhetorical and vocal quality assessment in call center talks," in *SLaTE 2015 – Sixth Workshop on Speech and Language Technology in Education*, 2015, pp. 29–34.

[11] B. Schuller, "Voice and speech analysis in search of states and traits," in *Computer Analysis of Human Behavior*, A. A. Salah and T. Gevers, Eds. Springer, 2011, pp. 227–253.

[12] F. Eyben, M. Wöllmer, and B. Schuller, *openSMILE – the Munich open Speech and Music Interpretation by Large Space Extraction toolkit*, München, 2010, programmdokumentation, Version 1.0.1, 23.05.2010.

[13] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *INTERSPEECH 2010*. ISCA, 2010, pp. 2795–2798.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[15] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[16] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco: Morgan Kaufmann, 2011.

[17] J. R. Quinlan, *C4.5: programs for machine learning*. San Mateo: Morgan Kaufmann, 1993.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and P. S. Yu, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[20] M. Walther, T. Mellouli, and O. Jokisch, "Fusion von Klassifikationsmodellen zur automatischen Erkennung von Stimmeigenschaften in der Qualitätsbewertung von Callcentergesprächen," in *Konferenzband Elektronische Sprachsignalverarbeitung (ESSV) 2015*, ser. Studientexte zur Sprachkommunikation, G. Wirsching, Ed., vol. 78. Dresden: TUDpress, 2015, pp. 188–195, in German.

[21] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, ",would you buy a car from me?' – on the likability of telephone voices," in *INTERSPEECH 2011*. ISCA, 2011, pp. 1557–1560.

[22] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech." in *LREC 2010*, 2010, pp. 1562–1565.

[23] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Computer Speech & Language*, vol. 29, no. 1, pp. 100–131, 2015.

[24] G. Liu and J. H. Hansen, "Supra-segmental feature based speaker trait detection," in *Proc. Odyssey*, 2014.

[25] S. Gonzalez and X. Anguera, "Perceptually inspired features for speaker likability classification." in *ICASSP*, 2013, pp. 8490–8494.

[26] N. Cummins, J. Epps, and J. M. K. Kua, "A comparison of classification paradigms for speaker likeability determination." in *INTERSPEECH 2012*, 2012.

[27] C. Montacié and M.-J. Caraty, "Pitch and intonation contribution to speakers' traits classification," in *INTERSPEECH 2012*. ISCA, 2012, pp. 526–529.

[28] R. Brueckner and B. Schuller, "Likability classification-a not so deep neural network approach." in *INTERSPEECH 2012*. ISCA, 2012, pp. 290–293.

[29] H. Buisman and E. Postma, "The log-gabor method: speech classification using spectrogram image analysis," in *INTERSPEECH 2012*. ISCA, 2012, pp. 518–521.

[30] J. Pohjalainen, S. Kadioglu, and O. Räsänen, "Feature selection for speaker traits," in *INTERSPEECH 2012*. ISCA, 2012, pp. 270–273.

[31] D. Lu and F. Sha, "Predicting likability of speakers with gaussian processes," in *INTERSPEECH 2012*. ISCA, 2012, pp. 286–289.

[32] D. Wu, "Genetic algorithm based feature selection for speaker trait classification," in *INTERSPEECH 2012*. ISCA, 2012, pp. 294–297.

[33] M. Hewlett Sanchez, A. Lawson, D. Vergyri, and H. Bratt, "Multi-system fusion of extended context prosodic and cepstral features for paralinguistic speaker trait classification." in *INTERSPEECH 2012*. ISCA, 2012, pp. 514–517.

[34] Y. Attabi and P. Dumouchel, "Anchor models and wccn normalization for speaker trait classification," in *INTERSPEECH 2012*. ISCA, 2012, pp. 522–525.

[35] G. K. Anumanchipalli, H. Meinedo, M. Bugalho, I. Trancoso, L. C. Oliveira, and A. W. Black, "Text-dependent pathological voice detection," in *INTERSPEECH 2012*. ISCA, 2012, pp. 530–533.

[36] L. Breiman *et al.*, "Heuristics of instability and stabilization in model selection," *The annals of statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

[37] D. T. Larose, *Discovering Knowledge in Data: an Introduction to Data Mining*. Hoboken: John Wiley & Sons, 2005.

Changes in IDS and ADS during parental leave – project sketch and first results of pilot study

Melanie Weirich, Adrian P. Simpson

Institute of German Linguistics, Friedrich Schiller University Jena

Melanie.weirich@uni-jena.de, adrian.simpson@uni-jena.de

Abstract

We are investigating the speech of German and Swedish mothers and fathers during the first year of their first baby. Both infant- and adult-directed speech are analyzed and compared between the sexes but also between different time points during the first year. In addition, the involvement in child care is considered as a potential factor. We are now in the process of finding participants and gathering data (read speech and spontaneous speech) from the first recording before birth of the child. Here, the speech material, our hypotheses and first pilot studies are presented.

keywords: IDS, ADS, sex-specific differences

1. Introduction

There is a growing body of evidence from articulatory and acoustic studies indicating that females speak more clearly than males (e.g. [1, 2, 3, 4, 5]) However, it still remains a matter of speculation as to why this might be the case. In particular, sociolinguistics has suggested that female speakers, often primary caregivers and guardians of the mother tongue, may seek to produce a clearer speaking style [6]. To our knowledge, this has never been empirically tested. In this project we investigate possible predictions arising from this assumption:

1) Male and female speakers produce clearer speech forms during their time as caregivers.
2) The clearer speech forms are not only restricted to infant-directed speech (IDS) but are also a feature of adult-directed speech (ADS)

While there has been a good deal of studies on female IDS (motherese), also comparing it to female ADS (e.g. [7, 8, 9, 10, 11]), similar studies on male subjects are few and far between (but see [12]). Recent socio-political developments in Germany and particularly in Sweden have made it possible to systematically investigate not only the speech of mothers but also of fathers in the role of primary caregiver.

The analysis is oriented on findings regarding sex-specific differences and characteristics of IDS. Both IDS and female speech differ from ADS and male speech, respectively, in temporal patterns and spectral characteristics (for an overview see e.g. [13], [14]). The project comprises the investigation of speech tempo, segment durations, average f0, f0 range, vowel space and sibilant contrast realizations. The speech material consists of a reading task and a picture naming task. The practicability of elicitation methods and recording procedures have been tested in two pilot studies.

[15] compared f0 and vowel space dimensions in IDS and ADS of seven male fathers. A larger vowel space in IDS than in ADS was found only in the reading task, not in the picture task. A higher average f0 and f0 range (SD) in IDS than in ADS was found with a stronger effect in the picture than in the reading task (see Figure 1).

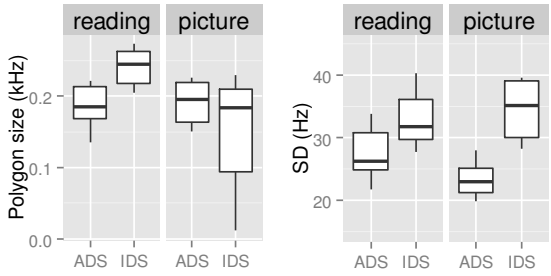


Figure 1: *Polygon size of F1/F2 space measured in /i: ɪ a ʊ u:/ and f0 variation (SD of average f0) of 7 fathers during ADS and IDS in reading and picture naming task*

In a further pilot study IDS and ADS of a single male subject before and after a two-month period of parental leave were investigated. A larger vowel space was found in IDS after the birth of the child, but only *before* parental leave, with a stronger effect in the reading task. *After* parental leave the ADS vowel space was found to be enlarged by comparison with its size before parental leave. Both average f0 and f0 range (SD) were greater in IDS than ADS *before* parental leave, with a stronger effect in reading task. *After* parental leave we find both a greater average f0 and f0 range in ADS by comparison with the recording before parental leave.

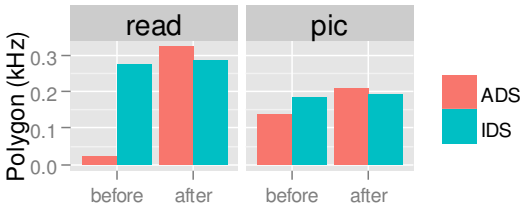


Figure 2: *Polygon size of a father in reading (left) and naming task (right) during ADS (red) and IDS (blue) at two time points (before parental leave and after)*

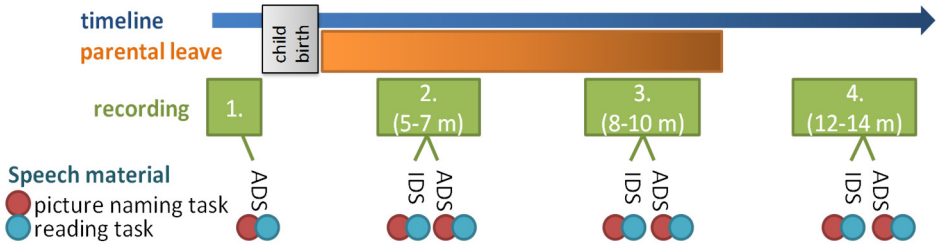


Figure 3: *Timeline of the speech recordings to be made during the longitudinal study*

2. Project sketch

The project investigates the potential relationship between typical phonetic correlates of careful speech and gender-specific roles - here, in the function of a primary caregiver of an infant. The focus of the study is not solely on infant-directed speech, but in particular the potential impact of being the primary caregiver on *adult*-directed speech. The project involves two different locations in Europe: Jena, Germany and Stockholm, Sweden. Sweden is particularly interesting due to its leading position in gender equality and the compatibility of family and work in Europe since the 1970s. Thus, the ongoing change of the conventional gender roles will be reflected in this project and cross-cultural and cross-linguistic analyses can be made.

We investigate both IDS and ADS speech samples of mothers and fathers at four time points during the first year of the child: 1) before birth, 2) and 3) during parental leave and 4) one month after parental leave is finished (see Fig. 3). The age of the infant at the time of recordings has to be comparable between different participants. Thus, the second and third recording will take place when the infant is between 5 and 7 and between 8 and 10 months old. These time periods were chosen since it is possible to find fathers who stay at home while the mothers are working only from around month six.

Figure 4: *Examples of pictures containing some of the target words*

We aim at gathering data from 15 mothers and 15 fathers in parental leave, and 10 working fathers as a control group. We are still in the process of finding participants but until now we have collected German data of the first recording session from 10 mothers (Mo), 2 fathers (Fa) and 9 control fathers (CFa).

The speech material consists of read speech and a picture naming task. The read speech comprises seven modified extracts from Astrid Lindgren stories [16], chosen because they are also very well known in a German context. The texts were modified to contain multiple tokens of peripheral vowel qualities and sibilants. The target sounds are embedded in the names of the children who are repeated frequently in every short story. Samples of spontaneous speech are elicited from a naming task using 15 pictures of animals and objects. Figure 2 shows some of the pictures used to elicit target words containing the vowels /iː, ɛ, a, ɔ, uː/ and sibilants /s, ʃ/ (e.g. *Kuh, Katze, Tiger; Tasse, Tasche*). The target objects were also chosen because they are suitable for eliciting vowels in both German and Swedish (e.g. *ko, kat, tiger; mössa, körsbär*).

Each participant sees each text only once, and the order of the task (reading, picture naming) is randomized over the sessions. Also, the order of speech register (ADS vs. IDS) is randomized for each speaker in each recording session. The interlocutor for the ADS register is always the same (a female student assistant) in each language.

Additional information about the participants is gathered at the different recording sessions by means of several questionnaires. Two important factors are: 1) the general involvement in care-giving and 2) the amount of speech used with the child. The answers to the questions result in numbers for the two factors that can be correlated with parameters in the acoustic analysis.

Socio-psychological information regarding self-reported assessment of masculinity/femininity (TMF, [17]) is also collected together with data on the positive attributes more socially desirable for women (e.g. emotional, helpful) using the GEPAQ-F scale [18].

2.1. Hypotheses

The project seeks to test two general hypotheses:

a) A change is found towards an enhancement of speech clarity in IDS in fathers (and mothers), and it is affected by the duration of parental leave and the involvement in child care.

Figure 5: *Vowel spaces in F1xF2 dimension of two fathers (CFa02, Fa01) and one mother (Mo01)*

One question is whether the amount of involvement in child care by the father correlates with the typical parameters of IDS. Related to this is whether the general involvement in care-giving is less important than the amount of speech used with the child (through singing, reading and talking).

b) A change is found in the fathers' phonetic correlates of clear speech also in ADS resulting in a decrease in sex-specific differences, during, or even after parental leave.

We will consider this against the backdrop of the socio-psychological attributes (TMF, GEPAQ-F).

3. First results

Due to the early state of the project, we will concentrate on data from three German speakers, who are representatives of the three different speaker groups that we are investigating: 1) mothers on parental leave (Mo), 2) fathers on parental leave (Fa) and 3) working "control" fathers (CFa). Until now, we only have data from the first recording (before the birth of the child). We will focus on the initial measurements taken in the read speech and the picture naming task in ADS.

Table 1: *Average f0 and SD of three speakers for the two tasks*

speaker	speech material	f0 (hz)	sd (hz)
Mo01	pic	227.4	37.64
	read	223.91	36.95
CFa02	pic	98.43	28.09
	read	98.6	28.69
Fa01	pic	131.44	32.3
	read	146.15	36.25

Table 1 shows the average f0 and the standard deviation of f0 for the three speakers. While Mo01 and CFa01 show expected values of 225 Hz for the female and 98 Hz for the male, Fa02 reveals a higher mean f0 then expected from the average value of a male speaker.

The two main spectral analysis parameters are 1) the size and dimensions of the vowel space and 2) the sibilant contrast realization. Figure 2 shows the vowel space spanned by F1 and F2 (in Hz) separated by speech material and speaker. The control father reveals the smallest vowel space, while the mother - as expected - has the largest vowel space. Interestingly the father that plans to stay at home and take care of the child for some months lies between the two. Differences between the speech tasks are speaker-specific: CFa02 exhibits a larger space in the picture naming task, Mo01 a larger space in the reading task and Fa01 does not differ much at all. To quantify the dimensions of the vowel space Table 2 shows the EDs in the horizontal and vertical dimension (ia and iu).

Table 2: *EDs between /a/ and /i:/ and between /i:/ and /u:/*

speaker	speech material	ED_ia (hz)	ED_iu (hz)
Mo01	pic	989	1580
	read	1206	1828
CFa02	pic	722	1444
	read	529	1226
Fa01	pic	958	1122
	read	904	973

For the acoustic parameterization of the sibilants, the spectral moments following [19] are measured but also *Discrete Cosine Transformation* (DCT; [20]) is used. The spectral moments consist of 1) the centroid frequency or Center of Gravity (COG), 2) the Standard Deviation (SD) of the COG, 3) the skewness describing the energy distribution over the whole frequency range of the spectrum and expresses if the frequencies are skewed towards the higher or the lower frequencies; and 4) kurtosis which reveals the spectral peakedness of the distribution. Recently, DCT has been shown to distinguish sibilants very well ([21, 22, 23, 24]). DCT decomposes the signal into a set of half-cycle cosine waves whereby the resulting amplitudes of these cosine waves are the DCT coefficients. We will concentrate on three DCT coefficients, which 1) are proportional to the linear slope of the spectrum (DCT1), 2) correspond to its curvature (DCT2), and 3) describe the amplitude of the higher frequencies (DCT3). Figure 6 shows DCT3 plotted as a function of DCT1 separated by speaker, reading task and sibilant. A difference in acoustic contrast can be seen between the speech tasks, with a

clearer contrast in the read speech. Regarding sex-specific differences, a clearer contrast in the female speaker (Mo01) was expected, but is only apparent in comparison to one of the two male speakers (i.e. CFa02).

Figure 6: *Acoustic contrast of sibilants measured in DCT1xDCT3 space*

To quantify the sibilant contrast, the Euclidian distance between the sibilants in DCT1xDCT2xDCT3 space was calculated for each speaker and speech task. The male speaker, CFa01, and the female speaker, Mo01, exhibit contrasts as expected: in the picture task the male's ED is nearly half of the female's ED (5.1 vs. 9.1). In the reading task the contrast is enhanced for both but much more so for the male (m: 10.9. vs. f: 11.3). The second male speaker (Fa01) again exhibits an unexpected pattern (as for f0 and vowel space size): he has the highest contrast in both speech tasks (11.6 and 12.6).

4. Conclusion

Two small pilot studies together with the results of the initial recordings in the main study have proven the practicability of the elicitation methods and recording procedures. The acquisition of further data, especially at different time points, will show whether the initial findings showing the expected tendencies can be substantiated, i.e. with the group of fathers on parental leave having f0, vowel space values and sibilant contrasts between those of the mothers and the working "control" fathers.

5. Acknowledgements

This work is supported by a German Research Council Grant (WE 5757/2-1) awarded to the first author. We would like to thank all participating subjects.

6. References

[1] Byrd, D. (1992) Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America* 92, 593-596.

[2] Hillenbrand, J., Getty, L. A., Clark, M. J. & Wheeler, K. (1995) Acoustic characteristics of American English vowels, *Journal of the Acoustical Society of America* 97, 3099-3111.

[3] Whiteside, S. P. (2001) Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *Journal of the Acoustical Society of America* 110, 464–478.

[4] Simpson, A. P. & Ericsdotter, C. (2003) Sex-specific durational differences in English and Swedish. *Proceedings of the XVth International Congress of Phonetic Sciences*, Barcelona, 1113–1116.

[5] Weirich, M., Fuchs, S., Simpson, A., Winkler, R. & Perrier, P. (forthcoming) Mumbling: macho or morphology. *Journal of Speech, Language and Hearing Research.*

[6] Labov, W. (1990) The intersection of sex and social class in the course of linguistic change. *Language Varia-tion and Change* 2, 205-254.

[7] Fernald, A. & Simon, T. (1984) Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* 20(1), 104-113.

[8] Shute, B. & Wheldall, K. (1989) Pitch alterations in british motherese: Some preliminary acoustic data. *Journal of Child Language* 16(3), 503-512.

[9] Kuhl, P. K., Andruski, J. E., Chistovich, I. A. & Chistovich, L. A. (1997) Cross-language analysis of phonetic units in language addressed to infants. *Science* 277(5326), 684-686.

[10] Liu, H. M., Kuhl, P. K. & Tsao, F. M. (2003) An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science* 6, F1-F10.

[11] Cristia, A. & Seidl, A. (2013) The hyperarticulation hypothesis of infant directed speech. *Journal of Child Language* 40 (2), 1-22.

[12] Sheehan E. A. (2004) *Influence of paternal involvement on fathers' infant-directed speech and infants' brain activity to male and female speech*. PhD dissertation. Emory University.

[13] Simpson, A. P. (2009) Phonetic differences between male and female speech, In *Language and Linguistics Compass*, vol. 3, no. 2, 621-640.

[14] Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review* 27. 501 – 532.

[15] Brosch (2016) *Kindgerichtete und erwachsenengerichtete Sprache im Vergleich bei jungen Vätern*. BA-Thesis. FSU Jena.

[16] Lindgren, A. (1988). *Wir Kinder aus Bullerbü*. 14. Auflage. Oetinger Verlag.

[17] Kachel S., Steffens M. C. & Niedlich C. (2016) Traditional Masculinity and Femininity: Validation of a New Scale Assessing Gender Roles. Frontiers in Psychology 7, 956.

[18] Runge TE, Frey D, Gollwitzer PM, Helmreich RL, Spence JT (1981) Masculine (instrumental) and feminine (expressive) traits. A comparison between students in the United States and West Germany. *J Cross-Cult Psychol* 12:142–162

[19] Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R.N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America* 841, 115-123.

[20] Watson, C. I. & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America* 106, 458-468.

[21] Guzik, K. & Harrington, J. (2007). The quantification of place of articulation assimilation in electropalatographic data using the similarity index (SI). *Advances in Speech Language Pathology* 9 (1), 109-119.,

[22] Weirich, M. & Simpson, A. (2015) Gender-specific differences in sibilant contrast realizations in English and German. *Proceedings of the XVIII. ICPhS*, Glasgow, 1-4,

[23] Bukmaier, V. & Harrington, J. (2016). The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change. *Journal of the International Phonetic Association*, 1-19.

[24] Jannedy, S. & Weirich, M. (2016). The Acoustics of Fricative Contrasts in Two German Dialects. in *PundP 2016– 12. Phonetik und Phonologie Konferenz, 13.-14. Oktober 2016, München.*

Voice Descriptions by Non-Experts: Validation of a Questionnaire

Benjamin Weiss

Technische Universität Berlin, Germany

benjamin.weiss@tu-berlin.de

Abstract

A questionnaire for non-expert German listeners was developed and validated in order to describe unacquainted voices on common perceptual dimensions. This instrument can be used to study acoustic correlates of perceptual features and voice-based attributions. Thirteen male and thirteen female speakers have been rated on a comprehensive questionnaire with 34 bipolar items by samesex listeners (30 women, 31 men). Two stimuli of each speaker have been evaluated. The aim was to finalize the development of the questionnaire by internal validation with factor analysis, and by testing for correlations with meaningful acoustic parameters for each obtained factor. After disregarding only few questionnaire items, the final sets of items show very good performances and reveals five, respectively six factors: Activity, Fluency, Precision, Softness, and Darkness for both, and additionally Tempo for women. Acoustic parameters were extracted automatically from the speech signal. Applying step-wise inclusion of preselected parameter sets, significant linear models were obtained for most of the factors. Only pronunciation Precision and Fluency will require more sophisticated approaches for automatic analysis.

Index Terms: vocal dimensions, speaker differences, semantic differential

1. Introduction

According to Scherer's modified version [1] of the Brunswick Lens Model, proximal cues represent listener's perceptions of speaker externalizations and therefore provide useful information in order to study paralinguistic topics like perception of personality, affect, or social status. Accordingly, instruments have to be developed assessing such non-expert listener perceptions (without expertise in, e.g., Phonetics or Psychology) of unacquainted speakers. With such instruments, which are typically questionnaires applying the same 5-, or 7-point ratings scale on all items, listeners' descriptions of speakers' voices and speaking styles can be studied, and related to acoustic cues of the speakers or to speaker attributions.

Although such questionnaires already exist, most are used only once. Additionally, verbal items have to be validated for each language/culture separately. Still, some have been empirically validated and even revised during development. Among those validated instruments are those for American English [2, 3] and Dutch [4, 5].

Last P&P, a German questionnaire under development was briefly presented. In this paper, a thorough analysis of this instrument is conducted, including the presentation of some acoustic correlates of the obtained perceptual dimensions. In line with the Lens Model, the ultimate goal is to use this instrument along with person attributions to analyze vocal likablity for the speaker database currently compiled at our lab [6].

Apart from the issue to properly translate the items to other languages, some validation studies were conducted only with male speakers [3], or with listeners, which might possess expert knowledge (e.g., 1st–3rd year students of Speech Therapy [4]).

Perceptual dimensions found vary a lot in labels, but can be clustered in phonetically motivated major categories:

- average pitch [4, 7, 2, 8],
- intonation (also called melodiousness or variation) [4, 9],
- rhythm (fluency or regularity) [2],
- tempo (also rate, duration, or animation) [4, 7, 2, 3],
- pronunciation (articulatory quality, precision, or clarity) [4, 2, 3],
- timbre (warmth, sharpness, creakiness, hoarseness, and so on) [4, 2, 3], and
- vocal effort (also harshness, roughness) [7, 8].

2. Procedure

A German questionnaire comprising 34 antonym pairs was evaluated in two same-sex listening tests. 13 male and 13 female speakers were chosen from the German Phondat I database of read sentences [10]. For each of the 26 speakers, two sentences were selected as stimuli.

For the male speakers these are (English translation by the authors): a) *"The knife and the fork are lying besides the plate."* (Messer und Gabel liegen neben dem Teller.) b) *"Now I am looking for white bread."*

The sentence a) was the same for the female speakers, but sentence b) was different, as not every sentence is available for each speaker: b) *"Who has still to do their homework?"* (Wer muss noch Schularbeiten machen?)

As a result, 26 stimuli of male speakers were presented randomly to 31 men (18–65 years, M=31.2, SD=10.48), and 26 stimuli of female speakers were randomly presented to 30 women (18–39 years, M=25.9, SD=5.30). No listener reported a background related to speech communication (Phonetics or the like). For playback, AKG K-601 headphones were used and the participants were paid for their contribution. Each session was planned for one hour, but most participants finished before 55 min.

3. Factor Analysis

In order to evaluate the questionnaire, a statistical analysis was performed separately for female and male data. However, the procedure for both data sets is identical: All items are consistently poled and the positive label of the antonyms used throughout this paper. Then, single items are excluded in three steps:

1. Items lacking rater consistency are excluded based on the Intra-Class-Correlation coefficient $ICC_{(3,2)}$. The level to exclude items was set to an ICC of .68 or lower. Actually, all excluded items have an ICC lower than .38.

2. A factor analysis is conducted and all items are removed which do not meet the following criterion: Item loading on one factor $\geq .5$, and the next strongest cross-loading are smaller than .29.
3. Cronbach's α is calculated for the factors and items are removed if α increases due to the exclusion.

Between step 1 and 2 the Kaiser-Meyer-Olkin index (KMO) is reported on the reduced set of items as measure of factor adequacy, i.e. a factor analysis can be conducted. Based on Horn's parallel analysis, the number of factors is determined. The subsequent factor analysis uses the oblique *oblimin* rotation. When inspecting the factor loadings in step 2, we follow the procedure advised in [11] and only changed the cross-loading criterion from a difference from .30 to .29 in order to include the item *"melodic"* for the male data, with a loading of .55 on *Activity* and .26 on *Softness*. After step 3, each factor is labeled based on the final items assigned to it.

For the male data, the items *"not breathy", "remarkable", "not nasal", "not creaky"*, and *"clear"*[1] are excluded first. The remaining items show sufficient rater consistency, e.g. the average ICC of these items equals .825. The parallel analysis reveals 5 factors (sufficient KMO = .93). These factors are named *Activity* ($\alpha = .82$), *Fluency* ($\alpha = .83$), *Softness* ($\alpha = .70$), *Precision* ($\alpha = .81$), and *Darkness* ($\alpha = .74$).

Table 1: Rotated factor matrix for the male data.

	Factor loadings				
Scale	1	2	3	4	5
powerful	**.70**				
emphasized	**.69**				
varied	**.69**				
sonorous	**.62**				
melodic	**.55**				
sharp	.40				-.38
professional	.39			.38	
fluent		**.70**			
conjoint		**.64**			
smooth		**.64**			
stable		.55		.36	
even		**.55**			
firm		.53		.35	
soft			**.70**		
gentle			**.56**		
warm	.46		.46		
pleasant	.48		.32		
loose	.35		.45		
relaxed		-.32	.45		
quiet	.30		.43		
natural			.36		
precise				**.74**	
articulate				**.72**	
accent-free				.49	
not coarse					-.37
dark					**.77**
low					**.75**
short		-.47			
fast		-.44	-.35		-.32
% Prop. Var	.14	.13	.10	.09	.06
% Cumul. Var	.14	.27	.37	.46	.52

Note: For better readability values smaller than .30 are suppressed and included items are bold.

Confer Table 1 for the loadings of the items on each factor and the variance explained. No additional items had to be removed to increase Cronbach's α as described in step 3. An additional factor analysis conducted with only the bold items would explain 60% of variance.

For the female data, only *"not breathy"*[2] is rated inconsistently (ICC = .33). Some of the other items excluded from the male data set do exhibit low ICCs (*"remarkable"*: .743, *"not nasal"*: .681, *"not creaky"*: .715, and *"clear"*: .696).[3] Nevertheless, these items are included for the female data, as they do not stand out of the body of the items if considering the ICC. The average ICC of all included items is .857.

The parallel analysis reveals 6 factors (sufficient KMO of .92). The factors are labeled *Softness* ($\alpha = .85$), *Fluency* ($\alpha = .89$), *Activity* ($\alpha = .83$), *Precision* ($\alpha = .69$), *Darkness* ($\alpha = .83$), and *Tempo* ($\alpha = .69$). Confer Table 2 for the loadings of items to the factors and the variance explained.

Table 2: Rotated factor matrix for the female data.

	Factor loadings					
Scale	1	2	3	4	5	6
powerful			**.62**			
emphasized			**.55**			
varied			**.70**			
sonorous			**.67**			
remarkable			**.61**			
melodic	.40		.53			
sharp			.49			
professional		.30	.32			
fluent		**.65**				
conjoint		.39				.38
smooth		**.75**				
stable		**.77**				
even		.46				
firm		**.78**				
soft	**.86**					
gentle	**.71**					
warm	**.68**					
pleasant	**.61**					
loose	.57	.39				
relaxed	.53	.45				
quiet	.35		-.42			
natural	.47					
precise		.51		.42		
articulate				**.65**		
accent-free				**.55**		
clear				**.63**		
not nasal				**.51**		
not creaky				.47		
not coarse				.47		
dark					**.79**	
low					**.90**	
short						**.64**
fast						**.73**
% Prop. Var	.14	.12	.11	.08	.06	.05
% Cumul. Var	.14	.26	.37	.45	.51	.56

Note: Values smaller than .30 are suppressed and included items are bold.

An additional factor analysis with only the bold items explains 59% of variance.

4. Acoustic Analysis

In order to further validate the factors found, typical acoustic parameters were extracted for each stimulus using Praat. Only

[1] "nicht behaucht", "unauffällig", "nicht nasal", "nicht knarrend", and "klar".

[2] "nicht behaucht".

[3] "unauffällig", "nicht nasal", "nicht knarrend", and "klar".

automatic measures were taken into account, as these facilitate the application of the factors for studying attribution processes in the future. Parameters associated with each factor are described in the list below.

HNR means harmonic to noise ratio of the voiced parts, and LTAS refers to the spectral moments of the long-term-average-spectrum. Duration is not including the leading/trailing silence.

Additionally, the following parameters were extracted: the alpha ratio, i.e. the energy difference between the bands 50 Hz..1 kHz and 1..5 kHz in dB to represent spectral tilt; Hammarberg indexes with energy of the bands in kHz, i.e. ((0..2–2..5)–(2..5–5..8)) related to the tight-breathy contrast, (2..5) related to hyper-hypo functional voice, and (0..2–2..5) related to coarseness [12]; as well as some measures related to syllable nuclei [13], namely estimates of the number of pauses and syllables. For the range of F0 and intensity, the difference between the 95% and 5% quartile was chosen to exclude outliers.

- *Darkness*: F0 (mean, min.), LTAS (spectral moments 1-4), alpha
- *Activity*: intensity (median, range, s.d.), F0 (range, s.d., slope), ((0..2 – 2..5) – (2.5 – 5.8))
- *Softness*: intensity (median, range, s.d.), jitter, shimmer, HNR, alpha, (2..5), (0..2 – 2..5)
- *Fluency*: no. of detected syllables, pauses, pause duration, estimates of articulation rate, speaking rate, syllable duration (all from [13])
- *Precision*: intensity (s.d.), F0 (s.d.), 1st spectral moment, no. of syllables & pauses, syllable duration
- *Tempo* (females only): duration, duration (only voiced parts), estimates of articulation rate and speaking rate

In order to eliminate an effect of sentence, e.g. on duration, acoustic measures were normalized to the mean, separately for both sentence groups. Separately for male and female speakers, linear models are conducted on the pre-selected measures, applying step-wise inclusion of predictors based on the Akaike information criterion (AIC penalty was increased from k=2 to k=2.705543 in order to include only predictors with p$\leq$.10).

In the following, linear models are presented, depicting parameter estimates and significance levels. They show those parameters of the list above that were actually included. For *Darkness*, the female and male model includes pitch and spectral parameters (Table 3). Please remind that the values are not standardized, but averaged between sentences, so that the impact of each parameter on the ratings can not be directly inferred. For female *Darkness*, e.g., two parameters correlate negatively, but F0 median is included positively into the model to compensate for the stronger effects of F0 min. and the 1st spectral moment.

Table 3: *Linear models for Darkness: Females (p<.0001, R^2=.703); males (p=.0199, R^2=.466).*

parameters	females	males
F0 median	0.3658*	n.a.
F0 min	-0.5888***	-0.020806**
LTAS 1st	-0.2685***	0.4268*
LTAS 2nd	n.a.	-0.3216.
LTAS 3rd	n.a.	0.5276
LTAS 4th	n.a.	-0.7238*
alpha	n.a.	-0.4920**

Also for *Activity* and *Softness* (Tables 4 & 5), there are significant models with meaningful parameter values for men and women. Interestingly, *Activity* is consistently described by an increased F0 range. For *Softness*, however, there are quite distinct models including intensity (males) or pitch perturbation (females). The positive effect of automatically estimated jitter has been reported already [14] and is not comparable to established values for sustained vowels. Common to men and women is the expected effect of a Hammarsberg index, which is negatively to coarseness (0..2)–(2..5) and, indirectly, positively related to tension (2..5) [12].

Table 4: *Linear models for Activity: Females (p=.0295, R^2=.264); males (p<.0001, R^2=.673).*

parameters	females	males
F0 s.d.	n.a.	-0.3167*
F0 range	0.2762*	0.4586***
Int s.d.	-0.1389.	n.a.

Table 5: *Linear models for Softness: Females (p=.0002, R^2=.589); males (p=.0006, R^2=.535).*

parameters	females	males
jitter	0.1706*	n.a.
shimmer	0.1533.	n.a.
(0..2)–(2..5)	0.1337	0.1301***
Int median	n.a.	-0.2042**
Int range	n.a.	-0.7316.

Table 6: *Linear models for Fluency: Females (p=.0502, R^2=.151); males (p=.0044, R^2=.442).*

parameters	females	males
speaking rate	-.2361.	n.a.
no. pauses	n.a.	.5846*
pause duration	n.a.	-.7781**
syllable duration	n.a.	-.1995*

Also as expected, *Precision* and *Fluency* are more difficult to grasp with such basic automatically obtained parameters (Tables 6 & 7). For *Fluency*, automatic syllable nucleus detection [13] works at least for males. As automatic syllable detection performs not comparable to plain duration it was not included in the model (Table 8). The estimates of speaking rate show a significant regression with duration, but not with the factor *Tempo*. Therefore, an automatic prediction of *Tempo* will be harder for heterogeneous sentences.

For *Precision*, parameters aggregating over the whole stimulus length are not appropriate. Although a significant model for females could be found, the 1st central moment, which can be related to segmental precision for consonants [15], is included, but surprisingly with a negative sign. Therefore, this model is disputable and has to be replaced by a more meaningful approach. Segmental analysis, e.g., vowel formants, should be used instead, maybe even including automatic speech recognition.

5. Discussion

The item loadings on each factor are quite similar for male and female data. *Activity* and *Darkness* are even identical. Only,

Table 7: *Linear models for Precision: Females (p=.0423, R^2=.1608); males (p=n.a., R^2=n.a.).*

parameters	females	males
LTAS 1st	-1.544*	n.a.

Table 8: *Linear model for Tempo: Only Females (p<.0001, R^2=.5222).*

parameters	females
duration	-0.44***

the additional factor *Tempo* of the female data is not found for men. For male data, two items of *Tempo* cross-load on *Fluency* and *Softness* (and *"fast"* also on *Darkness*), even when attempting six factors instead of five. This result might be caused by idiosyncrasies of some male stimuli provided and can be expected to change for other speaker sets, as Tempo represents a very typical perceptual dimension [4, 7, 2, 3].

Another slight difference is found for *Softness*, as the Factor includes the two items *"warm"* and *"pleasant"*,[4] which are excluded for the male data due to cross-loadings with *Activity*. Because of this and the different acoustic models this factor should be examined more closely. However, the small remaining inconsistencies in item loadings can be easily solved by a follow up experiment with manipulated stimuli, which is already planned for validating and completing the acoustic models. Although this questionnaire would benefit from additional, cross-gender, data, the current validation shows much consistencies in the factor analysis and acoustic modeling. We have obtained evidence to support typical perceptual dimensions, as introduced at the beginning of this paper. The dimensions found can be assigned to the categories of average pitch [4, 7, 2, 8] (the factor *Darkness* found here), intonation [4, 9] (*Activity*), rhythm [2] (*Fluency*), tempo [4, 7, 2, 3] (*Tempo*), pronunciation [4, 2, 3] (*Precision*) and timbre [4, 2, 3] (*Softness*).

Aspects of effort [7, 8], or precise articulatory descriptions like harshness, hoarseness, breathiness, or nasality [2] are not found in that distinctiveness here and should not be assumed to be manifested as general perceptual dimensions for non-experts listeners. Therefore, based on the stimuli used, this validation results in a final item set, maybe not suitable for pathological speech, but more suited for topics related to, e.g. personality attribution and speakers' likability.

6. Conclusions

The questionnaire presented in this paper is designed for non-expert German listeners to describe speakers according to their voice and speaking style. The instrument itself has been revised to its current version and shows very consistent results. Gender differences are found to be minimal and mostly refer to lacking consistency for articulatory specialties like nasality and hoarseness. In particular articulatory settings (creakiness, breathiness, harshness), as found in [2, 9], might be only visible for tailored sets of speakers, which strongly represent such settings, such as pathological [9] or affective voices. For confirmation, a final experiment is required applying manipulated and/or carefully preselected stimuli that exhibit such voice qualities.

The preliminary acoustic modeling supports the validity of the factors. Further work will have to deal with more sophisticated acoustic parameters for automatically predicting perceptual impressions of Precision, Fluency, and Tempo in order to subsequently use them for studying vocal attribution processes.

[4] "warm" and "angenehm".

7. Acknowledgements

This work was supported by the German Research Foundation (DFG, grant WE 5050/1-1).

8. Bibliographie

[1] K. Scherer, "Personality inference from voice quality: The loud voice of extraversion," *European Journal of Social Psychology*, vol. 8, pp. 467–487, 1978.

[2] ——, "Voice quality analysis of American and German speakers," *Journal of Psycholinguistie Research*, vol. 3, pp. 281–298, 1974.

[3] W. D. Voiers, "Perceptual bases of speaker identity," *Journal of the Acoustical Society of America*, vol. 36, pp. 1065–1073, 1964.

[4] W. Fagel and L. V. Herpt, "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation," *Speech Communication*, vol. 1, pp. 315–326, 1983.

[5] L. Boves, *The Phonetic Basis of Perceptual Ratings of Running Speech*. Dordrecht: Foris Publications, 1984.

[6] L. Fernández Gallardo, "Recording a high-quality german speech database for the study of speaker personality and likability," in *P&P, München*, 2016.

[7] T. Murry and S. Singh, "Multidimensional analysis of male and female voices," *Journal of the Acoustical Society of America*, vol. 68, pp. 1294–1300, 1980.

[8] S. Singh and T. Murry, "Multidimensional classification of normal voice qualities," *Journal of the Acoustical Society of America*, vol. 64, no. 1, pp. 81–87, 1978.

[9] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, pp. 265–275, 1991.

[10] Bayerisches Archiv für Sprachsignale, "PhonDat 1," München, 1995.

[11] M. Matsunaga, "How to factor-analyze your data right: Do's, don'ts, and how-to's," *International Journal of Psychological Research*, vol. 3, pp. 97–110, 2010.

[12] B. Hammerberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustical correlates of abnormal voice qualities," *Acta Otolaryngol*, vol. 90, pp. 441–451, 1980.

[13] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, pp. 385–390, 2009.

[14] B. Weiss and F. Burkhardt, "Is 'not bad' good enough? aspects of unknown voices' likability," in *Proc. INTERSPEECH*, 2012, pp. 1–4.

[15] B. Weiss, "Rate dependent spectral reduction for voiceless fricatives," in *Proc. INTERSPEECH, Brisbane*, 2008, p. 1968.

The impact of accent familiarity on the perception of difficult sound contrasts for German learners of English

Katrin Wolfswinkler[1], *Eva Reinisch*[1]

[1]Institute of Phonetics and Speech Processing, LMU Munich

k.wolfswinkler@yahoo.de, evarei@phonetik.uni-muenchen.de

Abstract

Second language (L2) learners usually retain a foreign accent even after years of training. The present study aimed at finding reasons for why accents are so persistent. We hypothesized that familiarity with an accent due to frequent exposure leads to adaptation which in turn allows listeners to understand the accent better, reducing the need for improvement. If this was the case, L2 learners should be better at understanding words spoken in a familiar than unfamiliar accent. To test this account, English minimal pairs containing two difficult sound contrasts for German learners (/ɛ/ vs. /æ/; voiceless vs. voiced word-final stops) were presented to native German listeners who had to identify the intended word. The tokens were produced by native speakers of English who served as a control, fellow Germans whose accent was highly familiar, Italians whose accent was somewhat familiar, and Finnish whose accent was unfamiliar. Results showed that words spoken by native English speakers or fellow Germans were recognized best, but words with the unfamiliar Finnish accent were recognized better than Italian accented words. A closer analysis of the acoustic differences that our speakers produced suggests that the acoustic cues available in the speech signal determined how well the words were identified.

Index Terms: Second language acquisition, Speech perception, Accent, Intelligibility, Familiarity, Foreign accent, Accent Familiarity

1. Introduction

Learners of a second language (L2) mostly retain an accent in L2 production even if they have long lasting experience with the L2. An obstacle in overcoming an accent appears to be that L2 learners perceive L2 sounds through a 'grid' of their first language (L1) sound inventory [1],[2]. That is, the sounds of the L2 tend to get assimilated to the perceptually closest L1 sound, affecting the correct perception – and consequently also the production – of L2 sounds that do not exist in the L1. Critically, if two different L2 sounds get assimilated to one single L1 category, L2 sound contrasts tend to get neutralized in perception as well as in production.

Although it has been shown that over time L2 learners get better at perceiving and producing difficult L2 sound contrasts, the ultimate attainment appears variable. One reason for this may be the quality of the input [3]. If the L2 is learned in the learners' native L1 environment, the main input is likely to come from other non-native speakers. This may include fellow learners at school and sometimes even the teachers. Therefore L2 learners usually have ample experience with listening to their own accent and hence are highly familiar with this accent. The purpose of the present study was to further investigate the role of non-native input and specifically, the role that accent familiarity plays in second language word recognition.

There are two lines of previous research that speak to this issue. First, native speakers of the target language have been shown to rapidly adapt to foreign accents. Even after brief exposure to an L2-speaker, native listeners of that L2 become better in understanding this learner's productions [4],[5].

Second, L2 learners have been shown to better understand an L2 when spoken by other learners of the same L1 background than by speakers of other L1 backgrounds. Sometimes L2 learners are even better at understanding L2 speakers with the same native language than native speakers of the target language. This has been called the '*interlanguage speech intelligibility benefit*' [6]. It is explained by the shared phonetic and phonological knowledge of the L1. Since L1 and L2 interact at segmental, phonotactic and prosodic levels this shared knowledge helps learners understand the accent.

The present study builds on these findings and asks whether, in addition to the shared L1 phonetic inventory, familiarity with an accent could affect L2 learners' performance in recognizing L2 words.

To test this issue, the present study examined the recognition of English words from minimal pairs where the contrasting sounds have been shown to be difficult to distinguish for native German learners of English [7],[8]. These were the vowels /ɛ/ - /æ/ and word-final voiced vs. voiceless stops. To investigate the influence of accent familiarity, the English minimal pairs were spoken by talkers of four different native languages whose accents differed in how frequently Germans were likely exposed to these accents. A native English accent (US) served as a control. German was the accent shared with the listeners hence frequently heard and highly familiar. An Italian accent was supposed to be somewhat familiar and Finnish was likely an unfamiliar accent.

Native German listeners were then exposed to these productions and had to identify which word of the minimal pair has been produced. We hypothesized that if familiarity with an accent affects how well this accent is understood, then listeners should be better able at understanding words spoken in a familiar than unfamiliar accent and rate this accent as less severe.

Specifically we tested four hypotheses:

H1: English words produced by native English speakers will be understood best since the words of the minimal pairs will be well differentiated.

H2: English words produced by native German speakers will be understood as well as the native English speakers' productions or second best since German is the listeners' own and hence most familiar accent. Listeners will be used to the specific cues that Germans tend to use for differentiating the words of the minimal pairs.

H3: English words produced by Italian speakers will be well understood, possibly as well as the Germans, since listeners are likely familiar with the accent.

H4: English words produced by Finnish speakers whose accent is largely unfamiliar will be understood significantly worse than the other accents (as familiarity cannot aid perception).

2. Method

2.1. Participants

Nineteen native German listeners (8 male) who were students at LMU in Munich participated. They were between 19 and 30 years old, were not enrolled in any Language and Literature class, and had not spent more than six months in an English speaking country. All had studied English for an average of nine years at school. According to self-report questionnaires administered after the experiment, all participants considered themselves as medium proficient in speaking and understanding spoken English.

2.2. Material

Twenty-two minimal word pairs were selected that contained either the vowel contrast /ɛ– æ/ (e.g., men-man, pen-pan, bed-bad,. . . ; 11 pairs) or a voicing contrast in word-final stops (e.g., bag-back, wide-white, robe-rope,. . . ; 11 pairs). Both contrasts have been shown to be difficult to distinguish for German learners of English in perception and production [7],[8].

All words were recorded by speakers of four different native languages (American English, German, Italian, and Finnish). For the recordings the words were presented at the end of semantically neutral carrier sentences (e.g., 'The next word is...') and speakers were asked to read the sentences at a comfortable pace. The list of words was randomized such that words of a minimal pair would not appear in direct succession. Native speakers of English and German had been recorded for another experiment and tokens for the present experiment were selected from this larger corpus [9]. Three native speakers of each, Italian and Finnish were recorded specifically for the present study. Two speakers per language were selected for the experiment. All speakers were female and between 18 and 30 years of age.

To assess how the speakers of the different accents produced the critical words with respect to the critical sound contrasts, acoustic measures were taken using Praat [10]. For the words with the vowel as target sounds the length as well as the first two formants of the vowel were measured. For the voicing contrast in word-final stops we measured duration of the preceding vowel and aspiration duration.

A summary of the measures is given in Table 1. The acoustic analyses showed that for the vowels, Italian and German learners produced smaller contrasts between the words of the minimal pairs than the native English or Finnish speakers, although they did not entirely neutralize the contrasts. The Finnish speakers differentiated the vowels clearly but did not produce the same vowel qualities as the native English speakers. For the stops, the German speakers produced large differences, close to the native English speakers. Italians and Finns produced smaller differences and their productions were more variable than the native English speakers' productions.

2.3. Design and Procedure

From the set of recorded words, eight minimal pairs with the vowel contrast and six pairs with word-final stops were selected for the perception experiment. The selection was based on the quality of the recordings and consistency of acoustic characteristics within the two speakers of a given accent.

For each speaker, four carrier sentences were selected and randomly combined with all words produced by this speaker. Carriers were selected such that they did not contain any hesitations and did not contain any of the critical sounds that listeners might use as an anchor for identifying the critical sound in the target word.

Listeners were seated in a sound-proof booth in front of a notebook computer wearing high-quality headphones. On each trial, listeners were presented visually the two words of a minimal pair. After 800ms preview the sentence was presented at a comfortable listening level. The listeners' task was to decide by button press (0 and 1 key on the computer keyboard with sides matched to the visual layout) which of the two words had been said. After logging their response the next trial started automatically. Each participant heard all words from all eight speakers once, resulting in a total of 224 unique trials (28 words x 8 speakers). Words and speakers were intermixed and presented in a separate random order for each participant.

In the same session after the identification test, participants were given a short accent rating task to assess the perceived strength of the speakers' accents. They were presented the carrier sentences without target words and asked to rate on a scale from 1 to 5 how strong they considered this speakers' accent. Endpoints and labels of the scale ('nearly no accent' – 'very strong accent') were counterbalanced across participants. The four sentences of the eight speakers were repeated three times in fully random order.

Table 1: *Means and Standard Deviations (in brackets) of the acoustical measures. Durations in milliseconds.*

	Vowels				**Stops**			
	/æ/		**/ɛ/**		**voiced**		**unvoiced**	
	F2-F1	**duration**	**F2-F1**	**duration**	**aspiration duration**	**vowel duration**	**aspiration duration**	**vowel duration**
English	1067 (422)	230 (60)	1266 (256)	160 (20)	30 (20)	230 (50)	80 (20)	150 (30)
German	1103 (176)	200 (60)	1198 (158)	170 (60)	200 (10)	220 (70)	50 (20)	160 (40)
Italian	975 (232)	180 (30)	1038 (276)	180 (50)	70 (20)	260 (60)	90 (40)	220 (70)
Finish	719 (119)	160 (40)	1380 (242)	140 (30)	50 (20)	230 (70)	60 (30)	200 (80)

3. Results

For the identification task, a generalized linear-mixed effects model was fitted with response (the correct/intended word coded as 1, the incorrect as 0) as the dichotomous dependent variable for which a logistic linking function was used. Fixed factors were Accent (English, German, Italian and Finnish), sound Contrast (/ɛ- æ/, voiced vs. unvoiced word-final stops) and their interaction. For the factor Accent the Level 'German' was mapped onto the intercept. The factor Contrast was added to the model since the acoustic measures of the productions had shown differences in how well speakers of the different accents had differentiated the words of the minimal pairs. The level 'Vowels' was mapped onto the intercept. The model was fitted with a full random-effects structure [11]. Figure 1 shows the results that were confirmed by the statistical analyses which are reported in Table 1.

Table 2: *Results of the mixed-effects model.*

	Fixed effect	***b***	***z***	***p***
Vowels	Intercept (German)	0.62	2.77	¡.01
	Accent (English)	0.72	2.79	¡.01
	Accent (Italian)	-0.42	-1.41	.16
	Accent (Finnish)	0.72	2.98	¡.005
Stops	Contrast	1.25	3.57	¡.001
	Contrast:Accent (English)	0.06	0.16	.88
	Contrast:Accent (Italian)	-0.41	-0.87	.38
	Contrast:Accent (Finnish)	-1.89	-5.05	¡.001

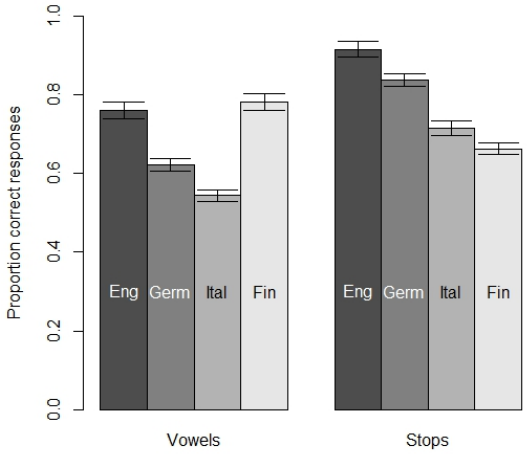


Figure 1: Proportion of correct word identifications dependent on accent for each sound contrast. Error bars indicate one standard error.

Results showed that for the vowels listeners were able to identify the correct word above chance if spoken with a German accent (see Table 2, Intercept). However, they were significantly better at identifying words when spoken by native English speakers or speakers with the Finnish accent. Words produced by speakers with the Italian accent were identified significantly worse.

For the stops listeners were overall better than for the vowels when produced by the Germans (see Table 1, Contrast). The same was the case for the productions of the English as well as the Italian speakers as evident in Figure 1. However, the stops of the native English speakers were identified slightly better compared to the German accented words while the stops produced by the Italian speakers where identified slightly worse but neither difference was significant. Just the difference to the Finnish speakers was significant as their consonant contrast was discriminated worst of all.

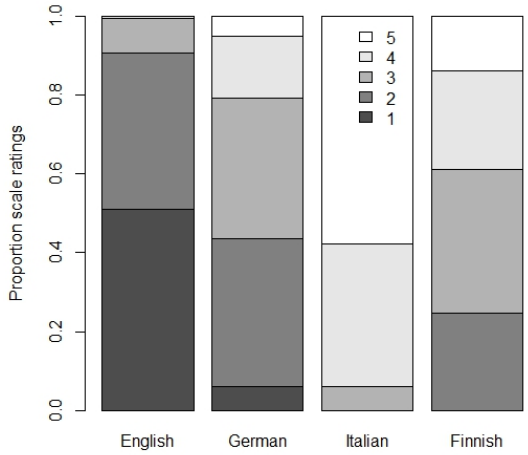


Figure 2: Results of the accent rating with the scale adjusted such that 1 = 'nearly no accent' and 5 = 'very strong accent'.

The results of the accent ratings which are presented in Figure 2 confirm the overall tendencies of the identification task (though note that for accent ratings the carrier sentences were presented without the target words). The native English speakers were rated the best suggesting that they were mostly recognized as native speakers. In the identification task their productions were understood best. German accented speakers were rated to have the next weakest accents which again matched the overall results from word identification. The accent ratings for the Finnish speakers was rated slightly stronger than the German accent, and the Italian speakers were rated as having the stronges accent.

4. Discussion

The present study was based on the observation that learners of a second language are frequently exposed not only to native speakers of the target language but also to fellow learners with either the same native language or different native languages. We addressed whether familiarity with the accent of fellow learners would help recognizing accented words in the L2 thereby assessing the boundaries of the interlanguage intelligibility benefit (ISI; [6]). The ISI suggested that shared phonetic knowledge allows learners to understand fellow learners with the same L1 (almost) as well as native speakers of the tar-

get language. We aimed at replicating this finding and extending it to other familiar vs. unfamiliar accents. Results from a word identification task with four different accents showed that German learners of English were best at identifying words from English minimal pairs when either spoken by native speakers of English or fellow German learners. This suggests that for listeners of medium proficiency native speech is well intelligible and that indeed shared L1 phonetics and frequent exposure to the accent facilitate word recognition (as was the case for the German accent). In other words, our results replicate the ISI and hypotheses 1 and 2 were confirmed with regard to the accents that were understood best. However, German accented words were identified worse than the native English tokens. Interestingly, our predictions about the other two accents were not borne out (i.e., pertaining to H3 and H4). That is, even though Germans could be considered familiar with Italian accent (e.g., Italy is a popular holiday destination and many Italians visit Munich) the Italian speakers' productions were understood less accurately than the other speakers' productions, even the Finnish speakers' productions whose accent was entirely unfamiliar.

However, our results can be explained by closer inspection of the acoustic differences that the speakers produced for the minimal pairs. For the vowels, Finnish speakers produced a large difference between the words of the minimal pairs. Therefore, even though the quality of the vowels differed from the native English speakers' productions, the intended words could be well identified – even better than the Germans' productions. The Italians in contrast, produced smaller differences between the vowels and were hence understood less accurately. For the stops, the picture looked somewhat different. Here, the Italian speakers produced larger differences than the Finnish speakers and words were hence identified better (details about the acoustic measures are reported in Table 1).

From a theoretical perspective this precedence of acoustic differences over accent familiarity may not be surprising, however, a hint of a role for familiarity can be seen when comparing the accent ratings. Although the Finnish speakers tended to produce the vowel contrast better than the German speakers, their overall accent (regardless of the any target words) was considered stronger. An additional experiment with better match of the speakers in overall accent ratings and the magnitude of differences they produce will shed further light on the issue of familiarity. With regard to the interlanguage intelligibility benefit we can conclude that L2 learners are best at understanding clearly produced words as found in native speech. In addition, the shared L1 phonetic inventory of fellow learners with the same L1 background, as well as the frequent exposure to the 'own' accent at school may make German listeners proficient at understanding the German accent. All these factors may contribute to the problem that for well understood speech there is little need for improvement which in turn could block learners from losing their accent even after years of learning.

5. Acknowledgements

We would like to thank the Lehre@LMU program for providing financial assistance for participant compensation.

6. Bibliography

[1] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, pp. 233–277, 1995.

[2] J. E. Flege, G. H. Yeni-Komshian, and S. Liu, "Age constraints on second-language acquisition," *Journal of memory and language*, vol. 41, no. 1, pp. 78–104, 1999.

[3] T. Piske, I. R. MacKay, and J. E. Flege, "Factors affecting degree of foreign accent in an l2: A review," *Journal of phonetics*, vol. 29, no. 2, pp. 191–215, 2001.

[4] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[5] M. J. Witteman, A. Weber, and J. M. McQueen, "Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation," *Attention, Perception, & Psychophysics*, vol. 75, no. 3, pp. 537–556, 2013.

[6] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *The Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1600–1610, 2003.

[7] M. Broersma and A. Cutler, "Competition dynamics of second-language listening," *The Quarterly Journal of Experimental Psychology*, vol. 64, no. 1, pp. 74–95, 2011.

[8] B. L. Smith, R. Hayes-Harb, M. Bruss, and A. Harker, "Production and perception of voicing and devoicing in similar german and english word pairs by native speakers of german," *Journal of Phonetics*, vol. 37, no. 3, pp. 257–275, 2009.

[9] N. Eger and E. Reinisch, "Is foreign-accented speech easier to understand if it is produced in one's own voice?" how Words Emerge and Dissolve: Evidence from Speech Production, Speech Perception, Acquisition and Disorders. [Online]. Available: http://www.phonetik.uni-muenchen.de/institut/veranstaltungen/how-words-emerge-and-dissolve

[10] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[11] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language*, vol. 68, no. 3, pp. 255–278, 2013.

'Chend' met <e> – 'Kind' mit <e>: using Big Data to explore phoneme-to-grapheme mapping in Lucerne Swiss German

Urban Zihlmann[1], *Adrian Leemann*[1]

[1]Phonetics Lab., Department of Theoretical and Applied Linguistics, University of Cambridge
{ubz20|al764}@cam.ac.uk

Abstract

Speakers from the canton of Lucerne are infamous for spelling Middle High German (MHG) <i> as <e> when communicating in written Swiss German, e.g. *Kind* ('child') as <Chend>. This phenomenon has been examined only impressionistically by phoneticians. This study provides a first account of this peculiarity of Lucerne Swiss German spellers: an analysis of normalised formant frequencies of two underlyingly MHG <i> vowels from 200+ speakers of the *Dialäkt Äpp* corpus revealed that the Lucerne allophone is in reality [e] for most of the localities examined, which may explain why in vernacular writing, spellers prefer <e> over <i>. Homophony due to this peculiarity can cause misunderstandings in written and oral communication, and possibly has repercussions on the reading and writing development of Lucerne students.

Index Terms: dialectology, formants, regional variation, crowdsourcing, Swiss German, iOS, Lucerne German

1. Introduction

The canton of Lucerne (LU) has a total surface area of 1,494 km^2, and approximately 394,600 inhabitants, which makes it the biggest and most populated canton of Central Switzerland [1]. Around 86.3% of its inhabitants view German (StG) as their first language [2]. Within the SwG dialect continuum, Lucerne German is a transition zone in the centre of Switzerland [3], located between the eastern and the western dialect areas. Furthermore, LU is split by the Brünig-Napf-Reuss line (applying equally to the Aargau), which is not only regarded as a cultural border between the east and the west of German-speaking Switzerland, but also a linguistic one [4].

The most seminal work on LU SwG was conducted by [3], who provided the first grammar that included general chapters on the phonetics of the dialect. To date, however, there has been no research on one of the most salient features of LU SwG: the orthographic representation of Middle High German (MHG) <i> as <e>. To illustrate this, Figure 1 depicts a text message written by a typical LU SwG speaker:

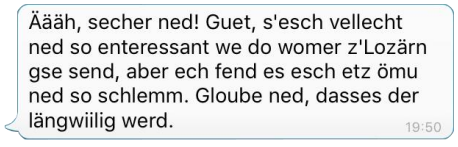


Figure1: *Text message written by a LU SwG speaker with high frequency of MHG <i> as <e>.*

The phrase reads *Ah, sicher nicht! Gut, es ist vielleicht nicht so interessant wie damals, als wir in Luzern waren, aber ich finde es ist jetzt aber nicht so schlimm. [Ich] glaube nicht, dass es dir langweilig wird*; 'Ah, definitely not! Well, it may not be as interesting as it was when we were in Lucerne together then, but I don't think it's that bad now. [I] don't think that you will be bored'. The vernacular representation features numerous MHG <i> as <e>, such as in *secher* ('definitely'), *ned* ('not') etc. A vast majority of other SwG vernacular writings would spell such words as *sicher* or *nid*, *i.e. with <i>*. This raises the question as to why most LU speakers opt for the grapheme <e> rather than <i>. What compounds the problem at hand is that some LU speakers have been shown to represent it as <i> as well (e.g. [5, 6]).

The present study contributes to fill this gap by performing an acoustic analysis of the vowels in *Ch**e**nd/Ch**i**nd* and *tr**e**nke/tr**i**nke* ('child' *Kind* and 'to drink' *trinken*, which go back to MHG *kint* and *trinken)*. It is assumed that the MHG short vowel <i> lowered its allophones to [i], [ɪ], or [e] [3, 7]. With these analyses we try to establish whether there is an acoustic basis for LU SwG writers of the vernacular to prefer <e> rather than <i> in representing MHG <i>: we predict that for most speakers, MHG <i> is indeed realized as [e] and that for this reason, LU SwG speakers tend to map MHG <i> with <e> in writing. To test this prediction, we analysed speech data from 200+ speakers stemming from the *Dialäkt Äpp (DÄ)* corpus. As the height of a vowel strongly correlates with the first formant [8], we will primarily focus on the description of *f1*.

2. Data and methods

2.1. iOS application: 'Dialäkt Äpp'

Dialäkt Äpp [9] enables users (1) to record 16 words and a short passage in their dialect and (2) to localise their dialect by choosing how they pronounce the 16 words in their SwG dialect. For the purpose of this study, we used functionality (1), introduced below. Prior to recording, the users of the app must indicate their age, sex, and dialect (see Figure 2, left panel).

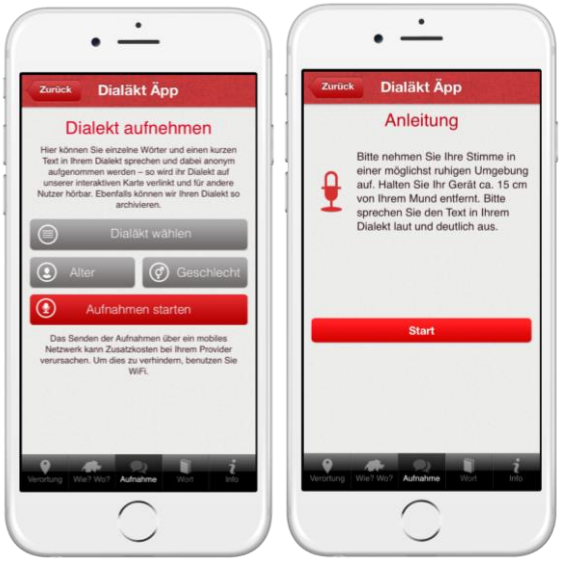


Figure 2: *User interface for dialect, age, and sex selection (left) and recording instructions (right)*

They are given instructions regarding the recording process (see Figure 2, right panel), stating: 'Please record your voice in as quiet an environment as possible. Keep an approximate distance of about 15 cm between your device and your lips. Please articulate the text loudly and clearly in your own dialectal pronunciation'. They then record the 16 words shown on individual prompts (see Figure 3, left panel). Each iOS device from the first generation onwards has sampling rates of up to 48 kHz [10]. For the purposes of this study, 48 kHz are sufficient for reliable formant measurements, as is a sampling rate of 10 kHz [11]. After the recording process the raw wav files are uploaded on a server and tagged with unique IDs. The recordings then appear on an interactive map (Figure 3, right panel, green and purple pins). After releasing *DÄ* on 22 March, 2013, it was the most downloaded free app for iPhones [12]. Presently, it has >58,000 downloads, and its database includes c. 3,000 speakers from 452 localities across German-speaking Switzerland [13, 14].

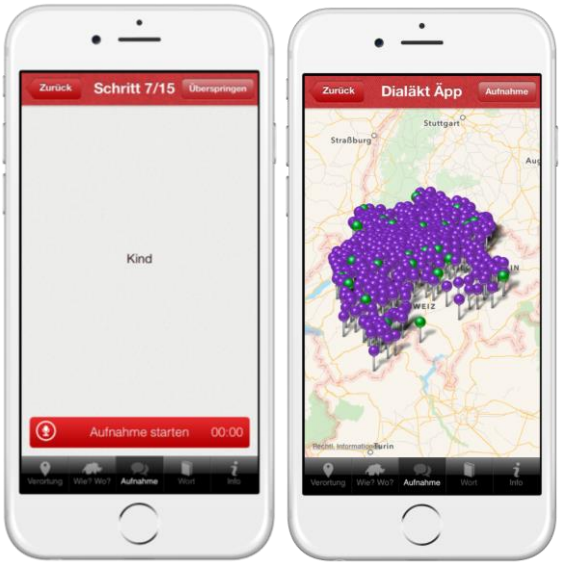


Figure 3: *User prompt for word recording (left) and interactive map of users recordings (right)*

2.2. Subjects

Users who indicated a Lucerne locality to best correspond to their dialect served as subjects. 206 speakers recorded the word *Kind* and 210 *trinken*. Speakers ranged between 10 and 77 years of age (mean=30.1; median=26.5; SD=15.0), with 47.8% males and 52.2% females. Subjects originated from virtually every corner of the canton (32 localities in total), which we divided into six regions for subsequent analyses of diatopic distributions (cf. 3.1.): Entlebuch (EB), Hinterland (HL), Lucerne-Hochdorf (L-H), Midland (ML), Mount Rigi (RG), and Schongau (SCH). The division is based on Fischer's linguistic observations on the morphological, lexical, and phonological level [3]. For instance, EB and RG speakers show differences in vowel quantity; they articulate open-syllables such as the first syllable in *jagen* ('to hunt') as [ˈjɑ.gə], while the rest of the canton produces them with long vowels, i.e. [ˈjɑː.gə], see Figure 4.

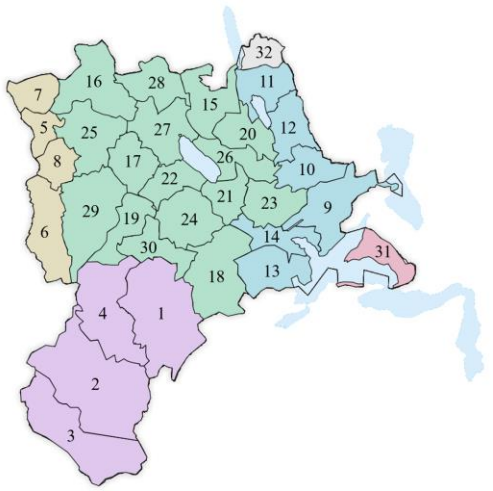


Figure 4: *Localities and broader dialect regions as used in the current sample*

2.3. Material

We chose two *DÄ* tokens with underlying MHG <i>: *Kind* 'child', and *trinken* 'to drink'. Some recordings were discarded due to background noise interference or other recording errors. The percentage of discarded tokens amounted to 17.5%.

2.4. Procedure

f1 and *f2* frequencies were measured in *Praat* [15]: if the segment was >10ms, measurements were taken 10ms after the beginning of the segment (M1), 10ms before the end of the segment (M2), and in the middle of the segment (M3; see Figure 5, top panel). If the segment was <10ms, measurements were taken at the beginning (M1) and at the end (M2) of the segments, as well as in the middle (M3; see Figure 5, bottom panel). As it is unclear which temporal value is most critical in the perception of the vowels, the mean value of M1-M3 was used for the analysis.

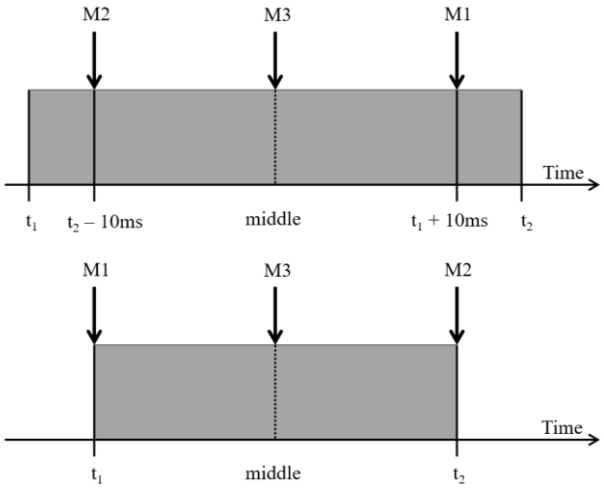


Figure 5: *Schematic of formant frequency measurements (M) >10ms (top) <10ms (bottom) (t_1 = beginning of the segment; t_2 = end of the segment)*

We normalised formant measurements using Bladon et al.'s base formula [16] which, however, only accounts for

differences in adult males and females. Thus, we adapted the formula to enable comparisons with younger speakers. To this end we considered the estimated vocal tract lengths of men and women (based on [17]) and calculated the age-appropriate amount of Barks to be subtracted from Bladon et al.'s formula. The difference between the average vocal tract length of an adult male and an adult female is 28.4 mm (m=169.3 mm; f=140.9 mm) and the difference between the respective value subtracted from Bladon et al.'s formula is 1.0 Bark (-0.53 Bark for the males; -1.53 Bark for the females). This allows us to calculate the millimetre-to-Bark ratio per millimetre difference to the mean adult vocal tract length, which is 0.035 Barks, i.e. $\frac{1}{28.4}$. We then included this as a subtraction term in Bladon et al.'s equation. This results in formula (1) for male and (2) for female speakers. The variables to be filled in are the raw formant frequencies in Hertz (f_i) and the mean vocal tract length by age (VTL_{age}).

$$(1)\quad f_i^N = 26.81\left(\frac{f_i}{1960+f_i}\right) - 0.53 - \left(\frac{1}{28.4}\left(169.3 - [VTL_{age}]\right)\right)$$

$$(2)\quad f_i^N = 26.81\left(\frac{f_i}{1960+f_i}\right) - 1.53 - \left(\frac{1}{28.4}\left(140.9 - [VTL_{age}]\right)\right)$$

As the equation results in Barks scores, we retransformed it to Hertz with *hqmisc* [18] (which uses Traunmüller's [19] formula) since the *R* package for plotting the vowels (*phonR* [20]) operates on the Hertz scale. Statistical analyses were conducted using *RStudio* [21].

3. Results

3.1. Diatopic differences

Table 1 summarises the mean formant frequencies and standard deviations (SD) by location.

	Locality	Mean *f1*	Mean *f2*	SD *f1*	SD *f2*	Area
1	Entlebuch	355	2069	42.6	192.6	EB
2	Escholzmatt	297	1934	69.4	65.4	EB
3	Marbach	355	1787	6.8	310.3	EB
4	Schüpfheim	375	2020	46.9	167.1	EB
5	Altbüron	427	2014	46.8	135.3	HL
6	Luthern	456	1823	10.1	87.8	HL
7	Pfaffnau	371	2015	9.5	38.3	HL
8	Zell	370	2112	58.1	186.6	HL
9	Ebikon	360	2018	41.8	183.3	L-H
10	Eschenbach	353	2001	89.3	200.4	L-H
11	Hitzkirch	381	2090	69.8	177.3	L-H
12	Hohenrain	432	2106	66.7	80.5	L-H
13	Horw	390	1902	43.4	203.6	L-H
14	Luzern	365	2021	50.9	188.4	L-H
15	Beromünster	367	2012	30.1	176.5	ML
16	Dagmersellen	386	1942	86.0	154.4	ML
17	Grosswangen	334	2087	27.3	185.1	ML
18	Malters	399	2132	58.4	123.8	ML
19	Menznau	442	2114	43.4	101.9	ML
20	Neudorf	358	1792	53.1	304.1	ML
21	Neuenkirch	378	2109	22.3	165.8	ML
22	Nottwil	347	1989	28.6	101.3	ML
23	Rothenburg	379	2019	58.5	186.0	ML
24	Ruswil	363	1951	44.1	161.5	ML
25	Schötz	373	2065	58.5	108.5	ML
26	Sempach	382	2140	34.0	151.7	ML
27	Sursee	365	1911	55.6	273.7	ML
28	Triengen	360	2057	51.4	139.3	ML
29	Willisau	402	2088	66.3	138.2	ML
30	Wolhusen	369	2071	44.4	155.6	ML
31	Weggis	353	1947	53.0	130.6	RG
	Total	376	2011	47.3	160.5	

Table 1: *Normalised vowel frequencies of MHG <i> and SDs by locality*

Overall, the mean *f1* frequency for the entire canton of LU is 376 Hz (SD=47.3 Hz). The lowest *f1*s (i.e. the highest articulations) are found in Escholzmatt (297 Hz), followed by Grosswangen (334 Hz), Nottwil (347 Hz), Weggis, and Eschenbach (both 353 Hz). The highest *f1*s (i.e. the lowest articulations) were found in Luthern (456 Hz), Menznau (442 Hz), Hohenrain (432 Hz), Altbüron (427 Hz), and Willisau (402 Hz). Vowel height seems to be rather stable throughout the canton (SD=47.3 Hz).

3.2. Differences by area

Table 2 summarises the mean formant frequencies and SDs by area; Figure 5 shows the values on the f1 / f2 vowel pane.

Area	Mean *f1*	Mean *f2*	SD *f1*	SD *f2*
EB	344	2004	54.8	184.7
HL	397	2043	57.3	170.9
L-H	367	2016	55.1	189.3
ML	376	2024	56.6	185.8
RG	340	1948	39.3	141.1

Table 2: *Mean normalised vowel frequencies of MHG <i> (in Hz) by area*

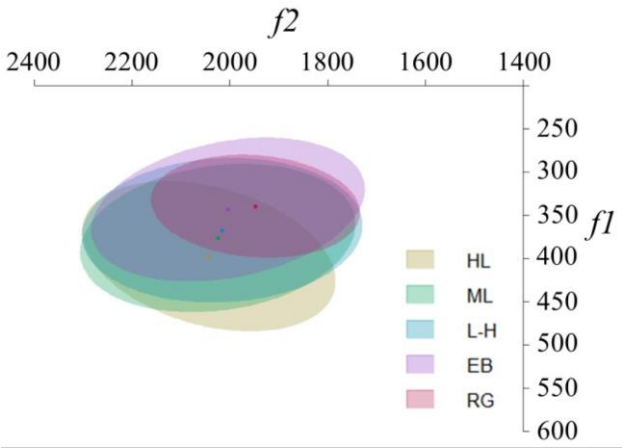


Figure 6: *Vowel ellipses of mean f1 and f2 frequencies with the corresponding SD (diameter of the oval)*

Figure 6 reveals substantial overlap between the regions. On the *f1* pane, RG reveals the lowest SD (39.3 Hz), while in HL, we observe most variation in *f1* (57.3 Hz). The highest articulation of MHG <i> is found in the RG area (340 Hz), whereas the lowest variant is found in HL (397 Hz). Both ML and L-H are in the vicinity of HL's values (ML, 376 Hz; 21 Hz lower than HL; L-H, 367 Hz; 30Hz difference to HL). EB, too approaches these values (344 Hz), although they produce a higher variant. Taken together, the northern three areas in the cantons all lie within a range of 30 Hz for *f1*, which accounts for the overlap in Figure 6. All areas exhibit values that approximate mean

frequencies of [e] of 390 Hz as suggested by Catford (as opposed to 240 Hz for [i]), but the linguistic background of the male speaker remains unspecified [22]. When data from StG are considered, such as Reubold [23], who found the formant frequencies of [e] to be 299 Hz, and 259 Hz for [i], the articulation in the entirety of LU seems to take place even lower.

4. Discussion

Our findings suggest that – on the whole – LU SwG articulations of MHG <i> are closer to [e] rather than [i]. There are regional differences, however: RG and EB demonstrate the highest variants, which has been previously documented in [3]. In Grosswangen and Nottwil, both within ML, however, we also found high articulations – yet their production is slightly lower than in RG and EB. Generally, however, the most suitable allophonic representation for MHG <i> appears to be [e]: here, mean *f1* frequencies are all in the vicinity of Catford's values for [e], and even higher (i.e. LU SwG articulates MHG <i> even lower) than the ones suggested by Reubold.

There are a number of implications to these findings. This lowering can cause confusion when LU speakers write to non-LU speakers in SwG vernacular, such as in informal texting or emails (see Figure 1). The formant frequencies reported in this study suggest that LU speakers tend to produce MHG <i> as [e], albeit with between-locality variation. If the writer chooses to represent this allophone with the grapheme <e>, misunderstandings could occur. If, for example, Zurich (ZH) SwG speakers read the message shown in Figure 1, they would likely associate <e> with the phonemes /ɛ/, /eː/, and /ə/, rather than conceiving of them as variants of MHG <i>, as intended by the LU SwG writer. Aside from potential confusion in written communication, in verbal communication, too, new homophones may emerge due to the lower articulation in LU SwG: the words *mer* ('me' *mir*), *mer* ('we' *wir*), and *Meer* ('sea' *Meer*) can all be homophonous and articulated as [meːɾ] in LU SwG. Moreover, LU SwG equivalents for the words 'seen' *gesehen* and 'been' *gewesen* are both neutralised to [g̊s̊eː], while ZH speakers maintain the [g̊s̊eː] / [g̊s̊iː] contrast. Though in isolation these words may cause misunderstandings, phrasal context typically resolves this.

The fact that the majority of LU dialect speakers use [e] for MHG <i> could also have implications for the classroom setting. German-speaking Switzerland is diglossic, yet LU children typically do not receive formal StG education until they begin school or kindergarten at age 5. By then, they will have learned to speak SwG vernacular, but will not have mastered the orthography of StG. As they grow older, they will first spell words close to what they sound like [24], followed by a simple grapheme-phoneme correspondence mechanism that will start to emerge at around age 7 [25]. However, when a given grapheme has more than one corresponding sound, or in other words, when the phoneme-grapheme correspondence is not 1:1, the spelling and reading acquisition process may be decelerated to some degree. This has been reported for English and Turkish students. When a student's native language has an irregular phoneme-grapheme correspondence as in English, they will typically master reading and spelling later than students whose native language has a more reliable sound-to-letter correspondence, such as in Turkish [26]. In the context of SwG, LU students will have to become aware that some of the [e]s they produce in SwG are orthographically represented by <e>, and some by <i> in StG – albeit vernacular writing allows for many (idiosyncratic) degrees of freedom. A speaker of ZH SwG, for example, who appears to have a more straightforward mapping of [i] to MHG <i> does not encounter this issue.

Interestingly, SwG speakers from western German-speaking Switzerland feature lowered MHG <i> as well, e.g. Bern (BE) German [27, 28]. Yet, they typically use <i> in written vernacular writing (e.g. <Chind> for *Kind*, 'child'). This suggests that LU SwG speakers conceptualise MHG <i> differently from these speakers, using an alternate strategy for phoneme to grapheme mapping. Further research is needed to explore (a) whether BE SwG speakers, in reality, have equally low articulations of MHG <i> as LU SwG speakers do and (b) whether BE and LU SwG perceive vowels equally. An exploration of both of these issues would help us better understand the peculiarity of LU SwG speakers' phoneme-to-grapheme mapping.

5. Conclusion

The findings of this study suggest that for most LU SwG dialects, the production of MHG <i> is closer to [e] rather than to [i]. Results on a more regional level revealed that speakers in the northern parts of the canton tend to articulate the phoneme closer to [e], while *f1* frequencies of RG and EB suggest the allophone to be somewhat higher for these regions (as reported in [3]). We speculate that misunderstandings may arise due to this dialect-specific phoneme-to-grapheme mapping when LU speakers are in written contact with non-LU speakers, e.g. in informal text messages. This lowering may have implications on the spelling acquisition process of StG in LU primary school students, given that students have to learn to dissociate LU-specific [e] from MHG <i>.

6. Acknowledgments

This study is one part of Zihlmann's MPhil thesis at the University of Cambridge. He wishes to thank the *Geert-und-Lore-Blanken-Schlemper Foundation* in CH-6210 Sursee, the *Hans-und-Wilma-Stutz Foundation* in CH-9102 Herisau, *Prof Beat Siebenhaar* (Leipzig), *Prof Francis Nolan (Cambridge)*, and *Marie-José Kolly (Zurich)* for their support and for co-creating the *DÄ* corpus.

7. References

[1] LUSTAT Statistics Lucerne. (2015). Bevölkerung nach Kanton. Retrieved 01 May 2016 from https://www.lustat.ch/files_ftp/daten/kt/0003/w011_001t_kt0003_zz_d_0000.html.

[2] Federal Statistics Office. (2014). *Languages and religions Data*. Retrieved 30 April 2016 from http://www.bfs.admin.ch/bfs/portal/en/index/themen/01/05/blank/key/sprachen.html.

[3] Fischer, L. (1960). Luzerndeutsche Grammatik und Wegweiser zur guten Mundart. Zurich: Schweizer Spiegel Verlag.

[4] Weiss, R. (1947). Die Brünig-Napf-Reuss-Linie als Kulturgrenze zwischen Ost- und Westschweiz auf volkskundlichen Karten. Geographica Helvetica 2(3), 153-75. doi:10.5194/gh-2-153-1947.

[5] Hotzenköcherle, R., Bigler, N., Schläpfer, R., & Börlin, R. (1984). Die Sprachlandschaften der deutschen Schweiz. Aarau: Sauerländer.

[6] Zihlmann, J. (1975). De jung Chuenz und anderi Gschichte. Luzern: Murbacher-Verlag.

[7] Haas, W. (1978). Sprachwandel und Sprachgeographie: Untersuchung zur Struktur des Dialektverschiedenheit am Beispiele der schweizerdeutschen Vokalsysteme. Wiesbaden: Steiner.

[8] Ashby, M., & Maidment, J. A. (2005). Introducing phonetic science. Cambridge: Cambridge University Press.

[9] Leemann, A., and Kolly, M.-J., Dialäkt Äpp. https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8, 2013.
[10] iPhone Specifications (2016). Retrieved 17 June 2016 from http://www.everyiphone.com/.
[11] Podesva, R. J., & Zsiga, E. (2013). Research methods in linguistics. Cambridge, UK: Cambridge University Press, 169-194.
[12] http://www.appannie.com/.
[13] Leemann, A., Kolly, M.-J., Purves, R., Britain, D., & Glaser, E. (2016). Crowdsourcing Language Change with Smartphone Applications. PLoS ONE, 11(1): e0143060. doi:10.1371/journal.pone.0143060.
[14] Leemann, A. (under review). Analyzing dialectal variation in articulation rate using crowdsourced speech data. Journal of Linguistic Geography.
[15] Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.17, retrieved 21 April 2016 from http://www.praat.org/.
[16] Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. Language and Communication, 4(1), 59-69.
[17] Goldstein, U. G. (1980). An articulatory model for the vocal tracts of growing children. PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA.
[18] Quené, H. (2014). *hqmisc* - Miscellaneous convenience functions and dataset. Version 0.1-1.
[19] Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. Journal of the Acoustical Society of America, 88(1), 97-100.
[20] McCloy, D. R. (2015). *phonR* - Tools for Phoneticians and Phonologists. Version 1.0-3.
[21] RStudio Team (2015). RStudio: Integrated Development for R [Computer program]. RStudio, Inc., Boston, MA. Version 0.99.486, retrieved 19 October 2015 from http://www.rstudio.com/.
[22] Catford, J. (2001). A practical introduction to phonetics. Oxford: Clarendon.
[23] Reubold, U. (2012). Über die Zusammenhänge zwischen Grundfrequenz und Vokalhöhe: Evidenzen aus longitudinalen Altersstimmenstudien, Perturbations- und Vokalerkennungsexperimenten. München.
[24] Ehri, L. (1987). Learning to read and spell words. Journal of Literacy Research HJLR, 19(1), 5-31. doi:10.1080/10862968709547585.
[25] Marsh, C., Friedman, M., Welch, V., & Desberg, P. (1981). A cognitive-developmental theory of reading acquisition. In G. E. MacKinnon & T. G. Waller (eds.), Reading research: Advances in theory and practice, Vol. 3. New York: Academic.
[26] Oney, B., & Goldman, S. R. (1984). Decoding and comprehension skills in Turkish and English: Effects of the regularity of grapheme-phoneme correspondences. Journal of Educational Psychology, 76(4), 557-568. doi:10.1037/0022-0663.76.4.557.
[27] Leemann, A., M.-J. Kolly, F. Nolan (2015). It's not phonetic aesthetics that drives dialect preference: the case of Swiss German. Proceedings of ICPhS 2015, 08/2015 Glasgow (UK).
[28] Marti, W. (1985). Berndeutsch-Grammatik für die heutige Mundart zwischen Thun und Jura. Bern: Francke.

AUTORENVERZEICHNIS