universität
wien

Institut für Romanistik

Akten der Konferenz "Phonetik und Phonologie im deutschsprachigen Raum (P&P14)"

Proceedings of the Conference "Phonetics and Phonology in the German Language Area (P&P14)"

ESV ERICH SCHMIDT VERLAG

facultas

F Frank & Timme
Verlag für wissenschaftliche Literatur

ISF

L
Langenscheidt

narr\f
ranck
e\atte
mpto

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

ÖH
UNI WIEN

universität wien

Botschaft
der Bundesrepublik Deutschland
Wien

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

AMBASSADE DE FRANCE
EN AUTRICHE

Ambasciata d'Italia
Vienna

Publiziert von
Institut für Romanistik
Universität Wien

Institut für Schallforschung
Österreichische Akademie der Wissenschaften

Tagungsinformationen
http://pundp2018.univie.ac.at/home/

Herausgegeben von
Pustka, Elissa/Pöchtrager, Markus A./Lenz, Alexandra N./Fanta-Jende, Johanna/Horvath, Julia/Jansen, Luise/Kamerhuber, Julia/Klingler, Nicola/Leykum, Hannah/Rennison, John (Eds.):
Akten der Konferenz „Phonetik und Phonologie im deutschsprachigen Raum (P&P14)“
Proceedings of the Conference „Phonetics and Phonology in the German Language Area (P&P14)“, Wien

Redaktion des Tagungsbandes
Barbara Tiefenbacher

DOI: 10.25365/phaidra.159

Pustka, Elissa/Pöchtrager, Markus A./Lenz, Alexandra N./Fanta-Jende, Johanna/Horvath, Julia/Jansen, Luise/ Kamerhuber, Julia/Klingler, Nicola/Leykum, Hannah/Rennison, John (Eds.):

Akten der Konferenz
"Phonetik und Phonologie im deutschsprachigen Raum (P&P14)"

Proceedings of the Conference
"Phonetics and Phonology in the German Language Area (P&P14)"

Wien
Open Access-Publikation im Repositorium *Phaidra*

VORWORT

Die Konferenz „Phonetik und Phonologie im deutschsprachigen Raum“ (P&P) fand vom 06.-07.09.2018 zum 14. Mal statt. In Österreich war es eine Premiere. Die Veranstaltung haben wir in Kooperation zwischen der Universität Wien und dem Institut für Schallforschung der Österreichischen Akademie der Wissenschaften (ÖAW) organisiert. An der Planung und Durchführung waren insgesamt 20 Personen aus unterschiedlichen Disziplinen beteiligt: Phonetiker*innen und Phonolog*innen, Germanist*innen und Romanist*innen. Bei der Begutachtung der Abstracts haben uns 22 Gutachter*innen tatkräftig unterstützt: 73 Einsendungen wurden von jeweils zwei Personen begutachtet; die 15 am besten bewerteten wurden als Vorträge angenommen, 48 als Poster und 10 abgelehnt. Für die Finanzierung der Räumlichkeiten und der Kaffeepausen waren wir auf zahlreiche externe Sponsor*innen angewiesen, um die P&P gebührenfrei zu halten: die Österreichische Akademie der Wissenschaften, die französische, italienische und deutsche Botschaft, die Verlage Facultas, Frank&Timme, Erich Schmidt, Narr und Langenscheidt sowie die Österreichische HochschülerInnenschaft (ÖH). Schließlich war es uns eine große Freude, 105 Personen zur P&P14 in Wien begrüßen zu dürfen. Allen Beteiligten sei an dieser Stelle noch einmal herzlich gedankt!

An dieser Stelle bedanken wir uns auch sehr herzlich bei John R. Rennison, der sich dem Herausgeber*innen-Team angeschlossen hat, um die englischsprachigen Artikel zu lektorieren, sowie den Sekretärinnen Barbara Tiefenbacher (Arbeitsbereich Pustka) und Theresa Ziegler (Arbeitsbereich Lenz), die uns beim Editionsprozess tatkräftig unterstützt haben.

Wien, den 15. Februar 2020
Elissa Pustka, Markus A. Pöchtrager & Alexandra N. Lenz

NACHRUF

Sylvia Mossmüller (1954 – 2018)

Mit großer Betroffenheit geben wir Nachricht vom Tod unserer langjährigen Kollegin Doz. Dr. Sylvia Moosmüller. Als Leiterin der Forschungsgruppe für Akustische Phonetik war sie maßgeblich an der erfolgreichen Forschung unseres Instituts und dem Aufbau seiner internationalen Reputation beteiligt.

Sylvia Moosmüller studierte Anglistik, Romanistik sowie Allgemeine und Angewandte Sprachwissenschaften an der Universität Wien. 1984 erlangte sie mit der Dissertation „Soziale und psychosoziale Sprachvariation: eine quantitative und qualitative Untersuchung zum gegenwärtigen Wiener Deutsch“ ihr Doktorat in Angewandter Sprachwissenschaft. Anschließend arbeitete sie in verschiedenen Projekten zum Wiener Deutsch vor allem in Zusammenarbeit mit Univ.-Prof. Dr. Wolfgang Dressler. Zeitgleich forschte sie auch immer wieder im Bereich der feministischen Linguistik und Gender Studies. Ihre Arbeit am Institut für Schallforschung (damals noch Kommission für Schallforschung) nahm sie im Jahr 1992 unter dem damaligen Leiter Doz. Dr. Werner Deutsch im Rahmen eines Projektes zum Thema forensische Phonetik auf. Mit der Approbation ihrer Habilitation „Vowels in Standard Austrian German. An Acoustic-Phonetic and Phonological Analysis“ erlangte sie 2008 die Venia legendi für Angewandte Sprachwissenschaft zuzüglich Phonetik und Phonologie. Neben ihrer Lehre an den Universitäten Wien und Graz unterrichtete sie auch an der Fachhochschule „Logopädie-Phoniatrie und Audiologie“.

Über die Jahre hinweg baute sie am Institut die Arbeitsgruppe für Akustische Phonetik auf, die unter ihrer Leitung zahlreiche national und international erfolgreiche und beachtete Projekte durchführte. In den Jahren 2008 bis 2015 übte sie zusätzlich noch die Funktion der stellvertretenden Direktorin des Instituts für Schallforschung aus.

Ihre Forschungsschwerpunkte lagen in den Bereichen der phonetischen und phonologischen Variation des Österreichischen Deutsch sowie in der akustisch-phonetischen Beschreibung der Vokale ausgewählter, bislang unzureichend beschriebener Sprachen. Daneben betrieb sie Forschung im Bereich der forensischen Phonetik, der Soziophonetik und klinischen Phonetik. Als Generalsekretärin der IAFPA (International Association for Forensic Phonetics and Acoustics) sowie als Mitglied der AG für forensische Sprach- und Audioanalyse des ENFSI war sie auch international in diesem Forschungsbereich sehr anerkannt.

Am 17. April 2018 verstarb sie nach schwerer Krankheit viel zu früh. Das Institut für Schallforschung, seine Mitarbeiterinnen und Mitarbeiter, trauern um eine hervorragende Wissenschaftlerin, die in ihrer feinen und zurückhaltenden Art eine sehr geschätzte und wertvolle Kollegin war. Unsere Gedanken sind bei ihrer Familie, der wir von Herzen unser Mitgefühl ausdrücken.

(https://www.kfs.oeaw.ac.at/index.php?option=com_content&view=article&id=1025:nachruf-sylvia-moosmueller-1954-2018&catid=171&Itemid=810&lang=de)

Abgedruckt mit freundlicher Genehmigung des Instituts für Schallforschung

INHALTSVERZEICHNIS

Storing linguistic and non-linguistic pitch contrasts

Yuki Asano[1]*, Holger Mitterer*[2]

[1]English Department, University of Tübingen, [2]Neuroscience, University of Malta

yuki.asano@uni-tuebingen.de, holger.mitterer@um.edu.mt

Abstract

This study investigates whether the memory capacity to store linguistic and non-linguistic pitch contrasts is first-language (henceforth L1)-dependent. We tested 24 participants each for the three L1s Chinese, Japanese and German, whose use of linguistic pitch contrasts differ. An online adaptive version of the Sequence Recall Task (SRT) presented sequences of between 2–9 stimuli (high-low or low-high pitch, henceforth HL and LH). The results show that the memory capacity to store both linguistic and non-linguistic pitch contrasts was L1-dependent. The use of lexical pitch contrasts in the tone system was more advantageous than those in the accentual system. A limited pattern of lexical pitch contrasts was still more advantageous than rich patterns of intonational pitch contrasts in this task.

Introduction

The ability to store prosodic information is modulated by the use of prosodic contrasts in one's L1 [1–3]. For example, French speakers, as opposed to Spanish speakers, have difficulties in storing non-words differing only in the location of stress [2]. The researchers used a SRT to investigate the effect of language on the ability to store non-native sounds [1,4]. This study investigates whether the L1-dependent ability to store prosodic contrasts can be transferred to non-speech. To this aim, we conducted an online adaptive version of the SRT by presenting linguistic and non-linguistic pitch contrasts to Chinese, Japanese and German listeners. Pitch is used to describe sound in both music and language [5] and is indispensable in all languages, although its contrast is used at different linguistic levels and in different manners across languages: First, pitch is used at different linguistic levels. In tone languages, such as in Mandarin Chinese (henceforth Chinese), each syllable is assigned one of four tones and substituting the tone of a syllable changes the meaning of the word or morpheme. A pitch-accent language, such as Tokyo Japanese (henceforth Japanese), exhibits the limited use of pitch to distinguish words in a manner similar to stress [6]. Finally, in an intonational language, such as in German, pitch is primarily used at the intonational level without bearing lexical tones. Second, the richness of the variety of pitch contrast also changes across languages. German exhibits a richer variety [7] than Japanese, because the limited pitch pattern of a pitch accent restricts its intonational patterns. One of the questions posed in the current study is whether the lexical use of pitch is more advantageous in storing linguistic and non-linguistic pitch contrasts or whether the richness of pitch contrasts in a language matters, regardless of the level at which pitch contrasts are used.

The comparison of a tone language and a pitch-accent language in this study is interesting as native Chinese listeners showed that they clearly relied on pitch information in spoken word recognition [8–10], while Japanese listeners showed controversial results [11–15]. [13] proposed the Limiting-Domain Hypothesis, claiming that the syllable is a unit of processing in lexical access. According to her hypothesis, tone can effectively be used for lexical access while pitch accent cannot, although both are perturbations of F0 and acoustically identical. In the latter case, pitch information is completed too late for spoken word recognition. This Hypothesis also explains why stress does not affect lexical access [16–18], because stress also shares the domain of the word.

The fact that Chinese listeners, but not Japanese, exploit pitch information in spoken word recognition should lead to sensitivity differences to pitch contrasts between them. [19] showed that Chinese listeners were faster and more accurate than Japanese in detecting minimally-paired words contrasting in pitch. [2], which examined the ability to learn tonal contrasts, showed that Chinese listeners had the highest ability, followed by German participants. French and Japanese participants showed the lowest ability to store them. The authors argued that the rich intonational structure of German was more advantageous in learning non-native tonal contrasts than the restricted lexical pitch contrasts of Japanese and the fewer utterance-level pitch contrasts available in Japanese and French. Based on the literature review, we formulate the following hypotheses for the current study.

Hypothesis 1: If the use of lexical pitch contrasts alone is advantageous, larger memory capacity for the pitch contrasts for Chinese and Japanese is expected than for German.

Hypothesis 2: If the use of pitch contrasts in a tone language is more advantageous than in an accentual language, larger memory capacity for Chinese is expected than for the others.

Hypothesis 3: If the richness of pitch patterns is most advantageous, German should outperform Chinese followed by Japanese.

With respect to differences in memory capacity for linguistic and non-linguistic pitch contrasts, we expect not to find differences between them within L1, but instead differences across the pitch conditions. The brain mechanisms governing language and music processing interact with each other and share an important link with respect to their underlying processing [20], and prior experience with a tone system is advantageous in differentiating pitches in music [21–23]. With respect to the memory capacity for tone, it has been theorised that a music-language connection may be found in studying the prevalence of absolute pitch (e.g. [21]). So far, no study has been conducted to investigate the memory capacity difference between linguistic and non-linguistic pitch contrasts.

Experiment

Methods

Participants: Twenty-four native listeners of Chinese, of Standard German and, of Japanese took part for a small fee. All participants were recruited in Tübingen. None of them was L1 bilingual nor had learnt any tone or pitch-accent language as an L2.

Materials: For the linguistic pitch condition, minimally-paired words that contrasted in pitch were selected in Chinese (*ai*-HL meaning *love* and *ai*-LH meaning *cancer*) and in Japanese (*ame*-HL meaning *rain* and *ame*-LH meaning *candy*). Since a pitch contrast alone in German does not distinguish words, a disyllabic word (*Imbiss* meaning *snack*) was selected and varied in its pitch (HL vs. LH). For the non-linguistic pitch condition, a complex pitch contrast was created using *Praat* [24].

For the segment condition, a minimal pair of non-words involving a segmental contrast (*muku* vs. *munu*) was constructed. None of the non-words is a real word in Chinese, Japanese or German, but they are phonotactically well-formed combinations of segments in all of these languages. The stimuli in the linguistic and the non-linguistic pitch conditions were produced using the speech synthesis program *MBOLA* by defining phonemes using the respective language packages, durations and pitch places.

Pitch values in the linguistic and the non-linguistic pitch conditions were first defined as H=280 Hz, L=150 Hz for HL stimuli and L=170 Hz, H=280 Hz for LH stimuli, as words with these values sounded most natural in the linguistic pitch condition. Then, these values were varied by multiplying them by 0.95, 0.97, 0.99, 1.00, 1.01, 1.03 and 1.05 resulting in seven tokens with slightly different pitch values. Pitch values in the segment condition were consistently 200 Hz, resulting in a flat pitch contour. The duration of each stimulus was 500 ms in all conditions.

Procedure: An online adaptive version of SRT presented sequences of 2–9 stimuli (e.g. a sequence of HL and LH). The task was to reproduce each sequence by typing the associated keys “1” and “2” in the correct order. The experiment was presented with *MatLab* on a desktop computer. The language of the experiment was English for all language groups. The experiment consisted of linguistic, non-linguistic pitch, and segment condition. Each condition started with a training session. The number of the stimuli varied between 2 and 9 according to the participant’s previous response; if the answer was correct, the next trial contained N+1 of stimuli, if incorrect, N−1. The maximal number of stimuli from the same member in a sequence was three. The number of trials presented in one condition was 30.

Results

The mean achieved length of the last three stimuli was extracted for each condition and participant to analyse participants’ language and condition-dependent memory capacity of the respective contrast. In the following analysis, statistical results from LMER models are first reported and then descriptive mean values and 95% CI error bars in plots. The full LMER-model with mean achieved length as a dependent measure, language group and condition as fixed factors and participant as a random factor showed significant interactions (overall $p < 0.001$), see Figure 1.

The between-condition analysis showed that, in the segment condition, all language groups did not differ from each other (β = -0.38, SE = 0.47, t = -0.8, p = 0.7 between Chinese and Japanese, β= 0.28, SE = 0.46, t = 0.6, p = 0.8 between Chinese and Germans, β= -0.66, SE = 0.47, t = -1.4, p = 0.3 between Japanese and Germans). In the Linguistic Pitch condition, Chinese outperformed Japanese, followed by Germans (β = 2.68, SE = 0.46, t = 5.8, p < 0.001 between Chinese and Germans, β= 1.57, SE = 0.46, t = 3.4, p < 0.01 between Chinese and Japanese, β= 1.12, SE = 0.46, t = 2.4, p < 0.05 between Japanese and Germans). Also in the non-linguistic pitch condition, Chinese outperformed Japanese, followed by Germans (β = 2.73, SE = 0.46, t = 5.9, p < 0.0001 between Chinese and Germans, β= 1.66, SE = 0.46, t = 3.6, p < 0.01 between Chinese and Japanese, β= 1.08, SE = 0.46, t = 2.3, p < 0.07 between Japanese and Germans).

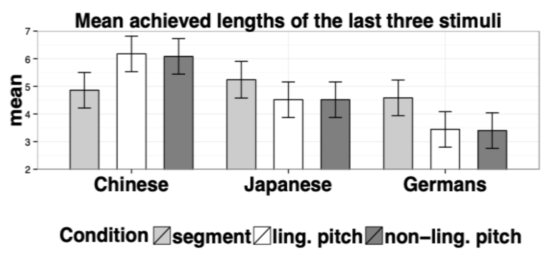


Fig. 1: Mean achieved lengths of the last three stimuli and 95% CI bars for each condition and L1.

In the between-language analysis, Chinese listeners showed larger memory capacity for the pitch contrasts in the both pitch conditions than in the segment condition (β = 1.23, SE = 0.35, t = 3.5, p < 0.001 between the non-linguistic pitch and the segment condition, β= 1.32, SE = 0.35, t = 3.7, p < 0.001 between the linguistic pitch and the segment condition and β= 9.17, SE = 0.35, t = 0.3, p = 1.0 between the both pitch conditions), while German listeners showed larger memory capacity in the segment condition than in the both pitch conditions (β = -1.18, SE = 0.35, t = -3.3, p < 0.01 between the non-linguistic pitch and the segment condition, β= -1.14, SE = 0.35, t = -3.2, p < 0.005 between the linguistic pitch and the segment condition, and β= 4.17, SE = 0.35, t = -0.1, p = 1.0 between the both pitch conditions). Japanese listeners did not show any difference between the conditions (β = 7.24, SE = 0.36, t = 2.0, p = 0.1 between the non-linguistic pitch and the segment condition, β= 7.24, SE = 0.36, t = 3.7, p = 2.0 between the linguistic pitch and the segment condition, and β= 4.44, SE = 0.35, t = 0.0, p = 1.0 between the both pitch conditions).

Discussion

The study investigated whether the memory capacity to store pitch contrasts is L1-dependent and whether this capacity can be transferred to non-linguistic pitch contrasts. The combination of Hypotheses 1 and 2 was valid. The use of lexical pitch contrasts alone, regardless of a tone or a pitch-accent language, was more advantageous than exhibiting pitch contrasts at the intonational level, so that larger memory capacity for storing the pitch contrasts by the Chinese and Japanese listeners was found than by the Germans. At the same time, the use of pitch contrasts in a tone language was more advantageous than in an accentual language, so that larger memory capacity for Chinese was found than for Japanese. Hypothesis 3 was not confirmed: The richness of pitch patterns was not advantageous in this study, so that Germans did not outperform Chinese or Japanese. The order of the participant groups did not differ between the linguistic and the non-linguistic pitch condition.

Japanese and Chinese listeners seem to encode pitch contrasts differently. We argue that this difference lies in the underlying mental representations of a Japanese pitch accent and Chinese tones. [19] proposed that pitch information itself is not stored with a word in the Japanese mental lexicon, being instead stored solely as an accent position without pitch information, whereas in the Chinese mental lexicon, pitch movement itself is represented together with words. In other words, Japanese listeners only (have to) pay attention to the position of a fall, but not to the whole contours. Moreover, the Japanese pitch accent lies beyond the syllable domain, so this information is available too late to exploit it in spoken word recognition [13]. This relates to the finding that Japanese listeners are less sensitive to pitch contrasts than Chinese [2, 19]. The use of lexical pitch contrasts therefore does not seem to necessarily create higher sensitivity to that contrast for a listener.

Whereas the German listeners were better than Japanese at learning non-words with Chinese pitch contrasts in [2], our German participants did not show larger memory capacity than our Japanese. This may lie in our experimental condition that the pitch contours used in this experiment resembled existing Japanese pitch patterns (HL vs. LH), whereas in [2], Japanese participants had to learn non-native tonal patterns. Moreover, the mental mechanism of learning new words and contours associated with pictures in [2] seems to be different to that of memorizing orders of limited patterns of sequences. To examine this assumption, future research should examine the memory capacity of pitch patterns that do not exist in Japanese lexicon (Tone 3 in Chinese).

References

[1] Dupoux, E, Peperkamp, S & N. Sebastián-Gallés, "A robust method to study stress 'deafness'," *Journal of Acoustical Society of America,* 110(3), pp. 1606–1618, 2001.

[2] Braun, B, Galts, T. & B. Kabak, "Lexical encoding of L2 tones: The role of L1 stress, pitch accent and intonation," *Second Language Research*, 30 (3), pp. 323–350, 2014.

[3] Asano, Y., *Localising foreign accents in speech perception, storage and production*, Ph.D. dissertation, University of Konstanz, Konstanz, 2016.

[4] Peperkamp, S. & E. Dupoux, "A typological study of stress 'deafness'," *Laboratory Phonology*, 7, pp. 203–240, 2002.

[5] Jänke, L., "The relationship between music and language," *Frontiers in Psychology*, 3(123), 2012.

[6] Hyman, L. M., "How (not) to do phonological typology: The case of pitch-accent," *Berkeley Phonology Lab Annual Report*, pp. 654–685, 2007.

[7] Baumann, S., Grice, M. & R. Benzmüller, "GToBI – a phonological system for the transcription of German intonation," in *Prosody 2000: Speech recognition and synthesis*, Puppel, S. & G. Demenko (Eds.) Poznan: Adam Mickiewicz University, 2001.

[8] Lee, C.-Y., "Does horse activate mother? Processing lexical tone in form priming," *Language and Speech*, 50(1), pp. 101–123, 2007.

[9] Li, X., Yang, Y. & P. Hagoort, "Pitch accent and lexical tone processing in Chinese discourse comprehension: an ERP study," *Brain Research,* 1222, pp. 192–200, 2008.

[10] Poss, N., Hung, T.-H. & U. Will, "The effects of tonal information on lexical activation in Mandarin," *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)*, Marjorie, K. C. & H. Kang (Eds.), 1, Columbus, Ohio: The Ohio State University, pp. 205–211, 2008.

[11] Otake, T. et al., "Mora or syllable? Speech segmentation in Japanese," *Journal of Memory and Language*, 32(2), pp. 258–278, 1993.

[12] Tamaoka, K. et al., "Is pitch accent necessary for comprehension by native Japanese speakers? –an ERP investigation," *Journal of Neurolinguistics*, 27(1), pp. 31–40, 2014.

[13] Walsh, D. L. "Limiting-domains in lexical access: Processing of lexical prosody," *University of Massachusetts Occasional Papers in Linguistics 19: Linguistics in the Laboratory*, Dickey, M. and S. Tunstall, (Eds.) Amherst: GLSA, pp. 133–155, 1993.

[14] Cutler, A. & T. Otake, "Pitch accent in spoken-word recognition in Japanese," *Acoustical Society of America*, 105(3), pp. 1877–1888, 1999.

[15] Minematsu, N. & K. Hirose, "Role of prosodic features in the role of prosodic features in the human process of perceiving spoken words and sentences in Japanese," *J. Acoust. Soc. Jpn.*, 16, pp. 311–320, 1995.

[16] A. Cutler, "Forbear is a homophone: Lexical prosody does not constrain lexical access," *Language and Speech*, 29(3), pp. 201–220, 1986.

[17] Cutler, A. & C. Clifton, "The use of prosodic information in word recognition," *Attention and Performance*, Bouma, X, H. & D. Bouwhuis (Eds.), Hilsdale, N.J: Erlbaum, pp. 183–196, 1984.

[18] Cutler, A. & D. Norris, "The role of strong syllables in segmentation for lexical access," *Journal of Experimental Psychology: Human Perception and Performance*, 14, pp. 113–121, 1988.

[19] Y. Asano & S. Yuen, "Asymmetric representations of lexical pitch," *Abstraction, Diversity, and Speech Dynamics*, 2017.

[20] Bidelman, G. M., S. Hutka, & S. Moreno, "Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music," *PLOS ONE* 8 (4), 2013, available from: https://doi.org/10.1371/journal.pone.0060676 [31.01.2019].

[21] Deutsch, D. et al., “Absolute pitch among American and Chinese conservatory students: Prevalence differences, and evidence for a speech-related critical period,” *Journal of Acoustical Society of America*, 119, pp. 719–722, 2006.

[22] Krishnan, A., Gandour, J. T. & G. M. Bidelman, “The effects of tone language experience on pitch processing in the brainstem,” *Journal of Neurolinguistics*, 23 (1), pp. 81–95, 2010.

[23] Giuliano, R. J. et al., “Native experience with a tone language enhances pitch discrimination and the timing of neural responses to pitch change,” *Frontiers in Psychology*, 2 (146), 2011.

[24] Boersma, P. & D. Weenink, *Praat: doing phonetics by computer* [computer program] version 5.2.20, 2011.

Vowel quality of German *äh* and *ähm* in dialogue moves

Malte Belz[1]

[1]Humboldt-Universität zu Berlin

malte.belz@hu-berlin.de

Abstract

This pilot study investigates the link between filler type (German äh *vs.* ähm*) and dialogue moves in spontaneous face-to-face dialogic interaction. Dialogic interaction is realized by a variety of dialogue moves, e.g. narratives or wh-questions. A speech corpus of eight female German speakers was annotated for filler type, segments and dialogue moves. Fillers occur mostly in narrative sequences and replies to wh-questions. Vowels in replies to wh-questions are uttered in a more open position than vowels in narratives (lower F1), for both filler types. F2 shows a curve difference between filler types, but not between move types.*

Introduction

This is an exploratory, phonetically informed corpus-based pilot study to shed light on the question whether the vowel quality of fillers in German depends on dialogue structure. Fillers, filled pauses or filler particles (FPs) take different forms, such as [ʔɛ ʔɛː ʔɛm ʔœ ʔœː ʔœːm] [1], [əː əːm] [2], [øː] [3], but also sequences of glottal stops [4] or clicks [5]. However, it is still unclear whether this variability is merely random or a function of contextual structures. In this study I will therefore focus on the vowel quality of vocalic and syllabic-nasal forms in German, orthographically often represented by *äh* and *ähm*.

Research on FPs is investigated either by using orthographic representations of *äh* and *ähm*, by borrowing from the phonological form [6, 7], or by using phonetic representations, including measures of phonetic variability, such as F0 or duration [8, 9]. This study will contribute to the latter group, investigating vowel quality of FPs. The link of FPs and discourse structure has been investigated for Dutch monologues of two speakers by Swerts [10]. He finds that “discourse structure to some extent determines the probability that an FP will occur” [10: 489]. The study of Lickley [11] indicates that discourse structure could be captured by categorizing discourse into different move types, which themselves have different distributions of FP frequency: “replies to wh-questions […] stand out as the most disfluent move type” [11: 95]. Building on this research, I hypothesize that discourse might play a role in the phonetic variation of filler form. In a first approach, I will investigate the vowel quality of *äh*/*ähm* in the context of different dialogue moves.

Method and data

Corpus

I conduct a corpus-based study on spontaneous, task-free face-to-face German dialogues using the GECO corpus [12]. Eight female students are included, with a total length of 187.42 minutes of speech and 44136 words. Annotation is added on multiple layers using Praat [13].

Annotation of fillers and dialogue moves

FPs are identified by listening to the dialogues and are marked on the FP layer. The directly adjacent micro-context to the left and to the right of the filler is annotated on the same layer, marking whether the context consists of segments, silent pauses, breath pauses, or extralinguistic entities. Vowel boundaries of FPs are annotated on a separate layer.

Dialogue moves are annotated following the dialogue coding system of Carletta et al. [14]. Dialogic interaction is realized by a variety of dialogue-initiating moves, e.g. narratives or explanations (annotation tag explain), polar questions (query-yn), wh-questions (query-wh), and response moves, e.g. positive replies (reply-yes), negative replies (reply-no), unsure replies (reply-unsure) and general replies to wh-questions (reply-wh). These moves are annotated on a separate layer as span annotations with reference to the word layer. Consider the following examples for (1) an explain move with only speaker A narrating and (2) a query-wh move of speaker A with a reply-wh move of speaker B.

(1) A: also ich äh wollte immer schon irgendwie Lehrer werden
‘well I äh always somehow wanted to become a teacher’

(2) A: Wie lang musst du denn dann noch studieren, wenn du schon im 6. Fachsemester bist?
‘How long do you have left to study then, if you’re already in the 6th semester?’
B: äh zwei Jahre so.
‘äh two years or so.’

Query and statistics

The annotated corpus was transformed into an *Emu* database [15] and queried with *emuR* version 1.0.0 [16] in *R* [17]. Formants were extracted for all vowels in *äh* and *ähm* using a Praat script [18]. Implausible F1 values outside the range of 200 – 1000 Hz and F2 values outside the range of 500 – 3000 Hz were discarded, which led to the exclusion of 0.01% of the original data.

As formant trajectories can be non-linear, generalized additive models were calculated in *R* using mgcv v.1.8-24 [19]. The dependent variable were the Lobanov normalized [20] first and second formant (F1 and F2). All formant trajectories were time normalized. To assess whether fixed factors were relevant to the model, binary factors were used, allowing for a direct evaluation whether to keep them in the model [21]. Factor smooths for speakers and micro-contexts are included as random effects. Autocorrelation is corrected for, and models are checked for fulfilling homoscedasticity and normally distributed residuals, which is achieved by using a scaled t-distribution as a link function in the model.

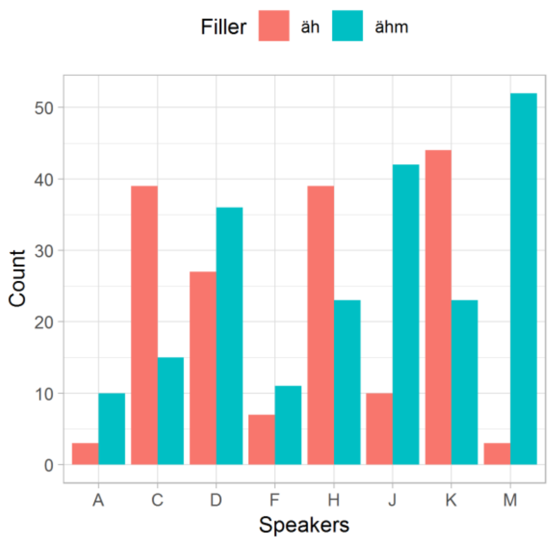


*Fig. 1: Filler count of vocalic (*äh*) and syllabic nasal (*ähm*) forms.*

Results

Frequencies and distribution

Fig. 1 shows the distribution of all FPs in the corpus. Both filler types are normally distributed over the eight speakers, as indicated by a Shapiro-Wilk test of normality for *äh* (W=0.84, p=0.075) and *ähm* (W=0.91, p=0.4). However, syllabic nasal types (N=212) are in toto used more often than vocalic types (N=172). An exact binomial test with 384 trials indicates that the true probability of success (*ähm*) is not equal to 0.5 (95% CI = [0.5, 0.6], p=0.046).

Fig. 2 shows the distribution of *äh* and *ähm* within dialogue moves. Fillers occur the most within narratives (explain), replies to wh-questions (query-wh) and unsure replies (query-unsure). Queries exhibit almost no filler use.

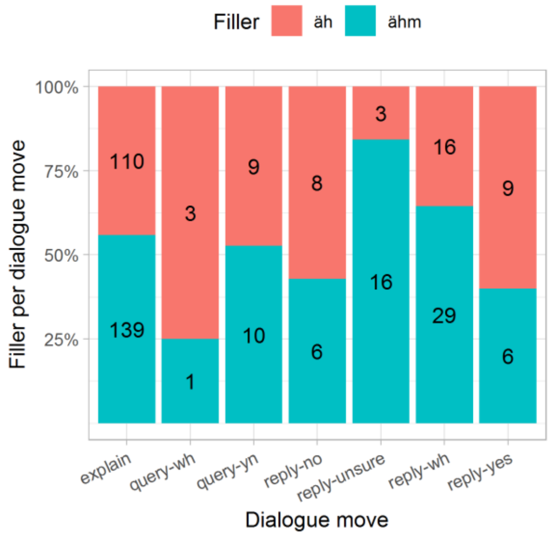


Fig. 2: Proportion of filler type per dialogue move and absolute numbers.

Vowel quality

Are fillers uttered in the same way and regardless of dialogic context, as suggested by graphematic studies? Here, I conduct an analysis for vocalic and syllabic nasal filler types (*äh* and *ähm*) used during narrative sequences or replies to wh-questions. These categories were chosen because they are the ones with the largest sample size (cf. Fig. 2).

Fig. 3 gives an impression of the great mean and standard variation of the formant trajectories of *äh* and *ähm* for the move types explain and reply-wh.

For F1, model comparison suggests that dialogue move, but not filler type should be included as a predictor ($p < 0.01$). The parametric coefficients of the model are given in Tab. 1. The intercept difference of 0.82 in Tab. 1 amounts to 124.8 Hz or 1 Bark. The approximates of smooth terms are not given here, as Fig. 4 shows that the difference lies mainly in the different intercepts of explain and reply-wh.

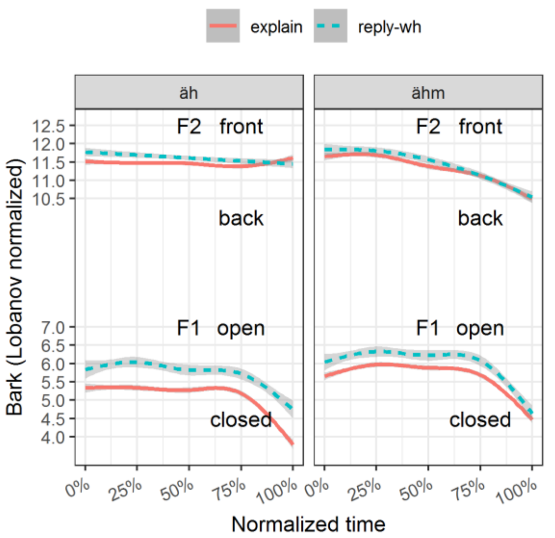


Fig. 3: F1 and F2 trajectories of all instances of äh *and* ähm *per dialogue move explain (red straight line) and reply-wh (blue dotted line) on a Lobanov normalized Bark scale.*

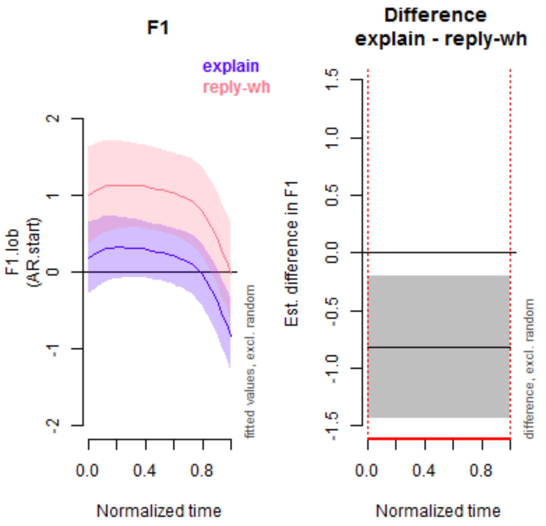


Fig. 4: The left graph shows the Lobanov normalized F1 trajectories for the dialogue moves explain (blue) and reply-wh (red). The right graph shows the difference between the two dialogue moves. The area where the difference from zero is significant is marked in red (in this case, they differ over the whole timespan).

For F2, dialogue move does not improve the model, but filler type does. Whereas the F2 intercepts for *äh* and *ähm* do not differ, they do show deviations in their curves (cf. Fig. 5), which could (at least in the end) be an effect of coarticulation towards the nasal.

	Est.	*Std. Err.*	*t*	*p*
Intercept	0.11	0.15	0.73	0.47
move *reply-wh*	0.82	0.30	2.68	0.007
R^2_{adj} = 0.404, Deviance explained = 36.8%, fREML = 7524, Scale est. = 1, n = 7410				

Tab. 1: Parametric coefficients of the model predicting F1.

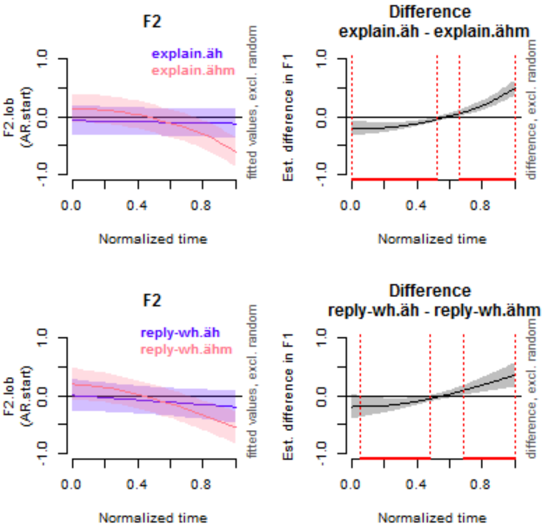


Fig. 5: The left column shows Lobanov normalized F2 trajectories for all ähs *and* ähms *within explain and reply-wh moves. The right column shows the differences between the categories on the left. The area where the difference is significant from zero is marked in red.*

Conclusion

Generally, fillers occur mostly in narrative and reply sequences, the latter corroborating the findings of [11]. Further, *ähm* forms are produced significantly more often in this sample of GECO, which is in accordance with [22] (*ähm* is more frequently produced in Germanic languages).

Some interesting relations between the height of the vowel used in fillers and the specific dialogue move was found. The data displays a difference of around +125 Hz between a vowel uttered in *äh*/*ähm* in a narrative and a vowel uttered in *äh*/*ähm* in a reply to a wh-question. Vowels in replies to wh-questions are thus uttered in a more open position, both for vocalic and syllabic nasal filler types. F2 shows no effect of move type, only of filler type, probably because of the additional nasal segment in *ähm*.

This raises several interesting questions for future research. Is the difference perceivable by the listener? Does the difference hold for new data? If FPs continue to show context-dependent structures, they could be described as form-function pairs within a constructional framework [23].

This study showed that the combination of phonetic and corpus-linguistic methods is fruitful to uncover new insights into the production of FPs and how they reflect the circumstances of their use.

References

[1] Willkop, E.-M., *Gliederungspartikeln im Dialog*, München: Iudicium, 1988.

[2] Trouvain, J., Fauth C. & B. Möbius, “Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech,” *Proceedings of Speech Prosody* (SP8), pp. 31–35, 2016.

[3] Schwitalla, J., “Kleine Wörter. Partikeln im Gespräch,” in *Über Wörter: Grundkurs Linguistik*, Dittmann, J. & C. Schmidt (Eds.), Freiburg im Breisgau: Rombach, pp. 259–283, 2002.

[4] Belz, M., “Glottal filled pauses in German,” *Proceedings of DiSS 2017: The 8th Workshop on Disfluency in Spontaneous Speech*, Stockholm, pp. 5–8, 2017.

[5] Trouvain, J., “On clicks in German,” in *Trends in phonetics and phonology: Studies from German-speaking Europe*, Leemann, A., Kolly, M.-J., Schmid, S. & V. Dellwo (Eds.), Bern: Peter Lang, pp. 21–34, 2015.

[6] Belz, M., Sauer, S., Lüdeling, A. & C. Mooshammer, “Fluently disfluent? Pauses and repairs of advanced learners and native speakers of German,” *International Journal of Learner Corpus Research*, 3(2), pp. 118–148, 2017.

[7] Clark, H. H. & J. E. Fox Tree, “Using uh and um in spontaneous speaking,” *Cognition*, 84(1), pp. 73–111, 2002.

[8] Shriberg, E. E. & R. J. Lickley, “Intonation of clause-internal filled pauses,” *Phonetica*, 50(3), pp. 172–179, 1993.

[9] Gósy, M., Gyarmathy D. & A. Beke, “Phonetic analysis of filled pauses based on a Hungarian-English learner corpus,” *International Journal of Learner Corpus Research*, 3(2), pp. 149–174, 2017.

[10] Swerts, M., “Filled pauses as markers of discourse structure,” *Journal of Pragmatics*, 30, pp. 485–496, 1998.

[11] Lickley, R. J., "Dialogue moves and disfluency rates," *ITRW on Disfluency in Spontaneous Speech* (DiSS'01), Edinburgh, pp. 93–96, 2001.

[12] Schweitzer, A. & N. Lewandowski, "Convergence of Articulation Rate in Spontaneous Speech," P*roceedings of Interspeech*, pp. 525–529, 2013.

[13] Boersma, P. & D. Weenink, *Praat: Doing phonetics by computer* [Computer program], version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/.

[14] Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. & A. H. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, 23(1), pp. 13–31, 1997.

[15] Winkelmann, R., Harrington J. & K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, 45, pp. 392–410, 2017.

[16] Winkelmann, R., Jänsch, K., Cassidy, S. & J. Harrington (2018). *emuR: Main Package of the EMU Speech Database Management System*.

[17] R Core Team (2018). *R: A language and environment for statistical computing*, Wien: R Foundation for Statistical Computing.

[18] Winkelmann, R., Pr*aatToFormants2AsspDataObj.R*, https://gist.github.com/raphywink/2512752a1efa56951f04 (Stand: 07.03.2017).

[19] Wood, S. N., *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness*, 2018.

[20] Lobanov, B. M., "Classification of Russian vowels spoken by different speakers," *The Journal of the Acoustical Society of America*, 49(2B), pp. 606–608, 1971.

[21] Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N. & R. H. Baayen, "Investigating dialectal differences using articulography," *Journal of Phonetics*, 59, pp. 122–143, 2016.

[22] Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J. & M. Liberman, "Variation and change in the use of hesitation markers in Germanic languages," *Language Dynamics and Change*, 6(2), pp. 199–234, 2016.

[23] Goldberg, A. E., *Constructions at Work: The Nature of Generalization in Language*, New York: OUP, 2006.

[ˈzɐmɐ] = „sagen wir“?
Perzeption phonetisch ambiger Reduktionsformen

Fabian Brackhane[1]

[1]Institut für Deutsche Sprache, Mannheim

brackhane@ids-mannheim.de

Abstract

Sogenannte „Pragmatikalisierte Mehrworteinheiten“ sind im Deutschen hochfrequent und unterliegen bisweilen tiefgreifenden phonetischen Reduktionsprozessen. Diese können Realisierungsvarianten hervorbringen, die in der Rückschau auf mehr als eine lexematische Ursprungsform zurückführbar sind. Die vorliegende Studie untersucht mit [ˈzɐmɐ] *einen besonders prägnanten Fall dieser Art anhand eines Perzeptionsexperimentes.*

Einleitung

In der gesprochenen Sprache begegnen häufig Wortgruppen, die in mehr oder weniger festgefügten Verbindungen stehen und mutmaßlich auch als Einheiten memoriert werden (sog. „Pragmatikalisierte Mehrworteinheiten“[1]). Wendungen wie *was weiß ich* oder *sagen wir mal* sind im spontansprachlichen Deutschen hochfrequent und erfreuen sich zunehmender Erforschung [1: 345]. Viele solcher Einheiten besitzen hierbei mehr als eine Verwendungsmöglichkeit und Bedeutung. Eine qualitative Analyse zeigt zudem, dass ihre Realisierung grundsätzlich reduktiven phonetischen Prozessen unterliegt, die sich jedoch hinsichtlich des segmentalen Substanzverlustes sehr differenziert darstellen.

Anteilig relativ selten, jedoch nicht nur vereinzelt begegnen massive Reduktionen, die eine Identifizierung des Ursprungslexems (resp. der Lexemgruppe) nur noch kontextuell zulassen, so bspw. [v̩ˈzɪç][2] für *was weiß ich* oder [ˈzɐmɐ] für *sagen wir (mal)*. In letzterem Falle besteht jedoch als weitere Komplikation das Problem der Ambiguität, das in dieser Untersuchung thematisiert werden soll.

Marker-Konstruktionen mit *sagen*[3] sind häufig und vielfältig:

(1) *sag mal, ich sag mal (so), sagen wir (mal), (ich) wollt grad sagen, sag bloß, wie gesagt, ich sage nur X, du sagst es, was du nicht sagst, das muss X gerade sagen, ehrlich gesagt, wie sagt man, ich würde sagen, das kannst du laut sagen…*

Hiervon entfallen allein auf die Konstruktionen *ich sage (jetzt) (einfach) mal (so)* und *sagen wir (doch) (jetzt) (einfach/vielleicht) (mal) (so/besser)* 431 Instanzen, wobei die jeweils knappsten Varianten *ich sage mal* und *sagen wir* knapp 70% der Treffer ausmachen. Beide Formen sind zwar funktional eng verwandt, aber nicht äquivalent; sie können neben der literalen Verwendung auch modalisierend eingesetzt werden, insbesondere die pluralische Form zusätzlich auch interpersonalisierend [2–4].

Beide Strukturen zeigen in der Marker-Verwendung eine Vielzahl phonetischer Reduktionen unterschiedlicher Intensität. Für *ich sage mal* stehen neben wenig reduzierten Varianten wie [ɪçˈzɐxmal] auch [ɪˈzɐmɐ] oder [ç̍ˈz̥ɐmɐ]. Für *sagen wir mal* sind Realisierungen wie [ˈzɐgŋ̍ʋɐmal], [ˈzɐŋəmɐ] oder [ˈzɐmɐmɐ] zu beobachten. Die Formen [ˈzɐŋʋɐ] oder [ˈzɐnʋə] lassen sich auf die knappere Struktur *sagen wir* zurückführen, bei der jedoch das fehlende *mal* womöglich bereits Ergebnis eines reduktiven Prozesses ist. In Kontexten, die die Verwendung einer der beiden Strukturen erwarten lassen, begegnet bisweilen die Lautfolge [ˈzɐmɐ]. Diese ist über Reduktionsprozesse schlüssig sowohl auf *ich sage mal* wie auch auf *sagen wir (mal)* zurückführbar und somit ambig.

Bei der Inspektion orthografischer Transkriptionen spontansprachlichen Materials im Korpus FOLK konnte festgestellt werden, dass die AnnotatorInnen zwar die Mehrzahl der [ˈzɐmɐ]-Instanzen als „sa_ma“ erfasst hatten, einige jedoch zu „ich sage mal“ bzw. „sagen wir mal“ aufgelöst hatten, obwohl diese segmentalphonetisch nicht mehr Informationen als erstere enthielten. Dabei wurde die Mehrzahl der fraglichen Instanzen pluralisch aufgelöst, nur zwei singularisch. Im Interview-Teil

[1] Definition und Terminologiegebrauch sind in der Literatur bislang nicht einheitlich (vgl. [1]).

[2] Alle Transkriptionen sind als phonetisch weit zu verstehen.

[3] Im Korpus FOLK [3] mit 10.899 Belegen das häufigste lexikalische Verb (Release 2017)

des Korpus „Deutsch heute“ [5] wiederum stellt sich der Befund exakt entgegensetzt dar: Von den dort orthografisch aufgelösten [ˈzɐmɐ]-Instanzen war die Mehrheit der Singular-Form zugewiesen worden.

Mithilfe eines Perzeptionsexperiments sollte nun untersucht werden, welche Anhaltspunkte den Transkribierenden zu ihrer – mutmaßlich intuitiven – Entscheidung verholfen hatte.

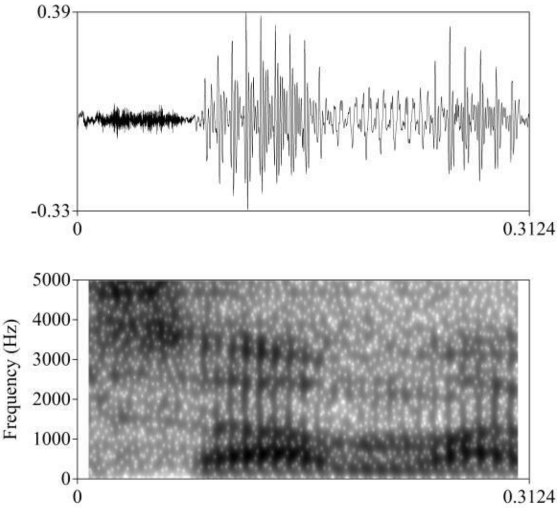


Abb. 1: Ein typischer [ˈzɐmɐ]-Beleg.

Daten

Verwendet wurden 18 [ˈzɐmɐ]-Instanzen aus spontansprachlichem Material der IDS-Korpora FOLK und „Deutsch heute“ [3, 5], die von Transkribierenden orthografisch als „ich sage mal“ oder „sagen wir (mal)“ verschriftlicht worden waren und qualitativ für ein Perzeptionsexperiment geeignet erschienen. Die Provenienz der Instanzen ist heterogen: Sie entstammen Unterrichtsgesprächen, Teambesprechungen, Interviews, Vorträgen und Map-Tasks. Ebenfalls vorhandene Belege aus informellen spontansprachlichen Interaktionen konnten aufgrund durchweg zu schlechter Aufnahmequalität nicht verwendet werden.

Tab. 1 gibt eine Aufstellung aller verwendeten Instanzen.

1 <p> Äh **sama**, wenn… wenn…

2 … die in der Nachbarschaft bei uns wohnen, weil die halt **sama** von Geburt an nichts anderes gelernt haben.

3 … sind halt so zwei… **sama** zwei Clicquen.

4 … kann man nix machen, also… **sama** im ersten Moment…

5 <p> **Sama**, wenn sie spielen und ich nichts anderes zu tun habe,

6 … ergibt sich kaum irgendwie mal, **sama**. <p>

7 Joa, es ging eigentlich. **Sama** ich habe eigentlich…

8 Jetzt wollten wir ja **sama** unabhängig von Fachkompetenz das beurteilen.

9 … ob da… **sama** in einer… Massenbildung…

10 … und in jeder Kaserne **sama** 200, 300 Mann,

11 … und dort **sama** schon zwei Autos sind,

12 … auf ähm jede Information ähm also **sama** so viel Wert legen,

13 … dass ähm die anderen Arbeiten, die äh **sama** bis zum heutigen Tag…

14 … schrittweise, so **sama** zwei, drei Schrittchen bis zum Stuhl…

15 Sie hat halt… ja **sama** einen Hauch von Nichts an.

16 Aber es geht noch über das hinaus, **sama** das ist jetzt die persönliche Schiene,

17 … dass äh durch einen Anschluss an Russland äh **sama** profitieren wird…

18 … So zwei Millimeter weiter oder **sama** drei Millimeter weiter oben.

Tab. 1: Verwendete [ˈzɐmɐ]-Instanzen.

Fragestellung / Hypothese

Untersucht werden sollte die Frage, was die Transkribierenden dazu motiviert hatte, die fraglichen [ˈzɐmɐ]-Instanzen als eine bestimmte der Grundformen zu identifizieren.

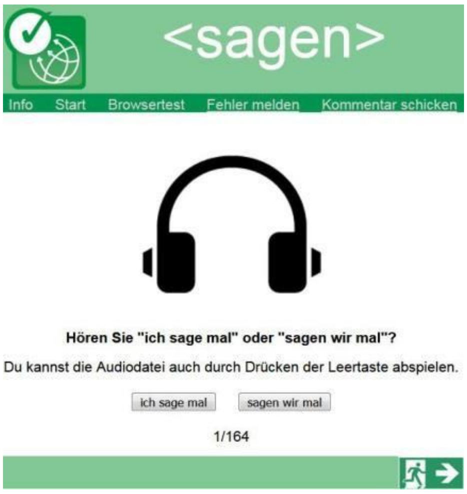
Da alle Instanzen phonetisch weitgehend identisch und ambig sind, lag es nahe, die kontextuelle Umgebung näher zu betrachten. Es wurde angenommen, dass isoliert präsentierte [ˈzɐmɐ]-Belege einigermaßen gleichverteilt den beiden angebotenen Optionen „ich sage mal“ und „sagen wir (mal)“ zugeordnet werden würden, während mit zunehmender Kontextgröße die Entscheidung deutlicher zugunsten einer Variante fallen würde.

Perzeptionsexperiment

Aus jeder Instanz wurden drei separate Stimuli generiert (=53 einzelne Stimuli[4]): 1. Der [ˈzɐmɐ]-Beleg isoliert, 2. [ˈzɐmɐ] plus ein Lexem links und 3. [ˈzɐmɐ] innerhalb der vollständigen IP.[5]

Jeder Stimulus wurde im Experiment drei Mal innerhalb dreier jeweils in sich randomisierten Durchgängen präsentiert (=159 Stimuli gesamt). Pro Präsentation konnte jeder Stimulus bis zu drei Mal angehört werden, bevor eine Bewertung abgegeben werden musste. Die Versuchspersonen wurden hierzu aufgefordert, auf eine von zwei Schaltflächen mit den Aufschriften „ich sage mal“ bzw. „sagen wir mal“ zu klicken. Position und Zuordnung der Schaltflächen wurden konstant gehalten, um eventuelle Bewertungsartefakte („Durchklicken“) besser erkennen zu können.

Das Experiment wurde mithilfe des Frameworks PERCY [6] als Online-Experiment konzipiert (Abb. 2).



Hören Sie "ich sage mal" oder "sagen wir mal"?

Du kannst die Audiodatei auch durch Drücken der Leertaste abspielen.

Abb. 2: PERCY-Interface des Perzeptionsexperiments.

Ergebnisse

Am Experiment nahmen 120 Versuchspersonen teil. Jedoch absolvierten lediglich 42 von ihnen einen vollständigen Durchgang, die übrigen brachen vorzeitig ab.

Für die Auswertung konnten jedoch alle Bewertungen verwendet werden, da ein Vergleich der Mittelwerte zeigte, dass sich gegenüber den Werten der vollständigen Datensets nur marginale Abweichungen ergaben.

Die Mehrzahl der Stimuli wurde der singularischen Ursprungsform „ich sage mal“ zugewiesen. Ein stringenter Zusammenhang zwischen Kontextgröße und Entscheidungssicherheit ist indessen nicht erkennbar (Abb. 3). Zwar gibt es Instanzen wie 10 oder 18, bei denen dieser Zusammenhang zu bestehen scheint, bei weitaus mehr ist dies jedoch offenkundig nicht der Fall. Es gibt im Gegenteil sogar Instanzen, bei denen die Stimuli mit keinem oder wenig Kontext stabiler bewertet wurden als der jeweilige Stimulus mit der gesamten IP (4, 6, 8, 13). Bei einzelnen Belegen kehrt sich die – eindeutige – Bewertung mit zunehmendem Kontext um (9, 11, 12, 14, 17, jeweils Singular → Plural).

Diskussion

Phonetisch ambige [ˈzɐmɐ]-Realisierungen waren von den Transkribenden nur vereinzelt mit der singularischen Form „ich sage mal“ assoziiert worden. Im Gegensatz hierzu steht das Ergebnis des Perzeptionsexperiments, bei dem die Mehrheit der Stimuli (28) ungeachtet der Kontextgröße dieser Ausgangsform zugeordnet wurde. Um ein Artefakt des Versuchsaufbaues auszuschließen (konstante Positionen der Bewertungsfelder im Interface), wurde für die 42 vollständigen Datensets erhoben, wie viel Prozent der Bewertungen konsistent gewesen waren, wie viele Versuchspersonen also einen Stimulus konstant drei Mal identisch bewertet hatten. Hierbei zeigt sich, dass der prozentuale Anteil absolut gesehen zwar natürlicherweise zumeist niedriger liegt als bei Betrachtung aller Bewertungen, sich die proportionalen Verhältnisse zwischen den Bewertungen nur vereinzelt ändern. Zugleich ist der Anteil der konsistenten Bewertungen von Stimulus zu Stimulus so unterschiedlich, dass ein Artefakt der Bewertungen ausgeschlossen werden kann.

[4] Für die Instanz 5 war kontextbedingt ein +Lex- Stimulus nicht möglich (Turninitiale Instanz).

[5] Für die „+Lexem“- und „IP“-Stimuli konnte dieser Grundsatz durch die für spontansprachliches Material typischen, oftmals irregulären grammatischen Konstruktionen nicht vollständig konsequent umsetzt werden. Der Stimulus 1 bspw. besitzt weder ein echtes linkes Lexem noch ist er in eine regelhafte IP eingebettet. Für derartige Fälle wurden jeweils möglichst adäquate Ausschnitte gewählt.

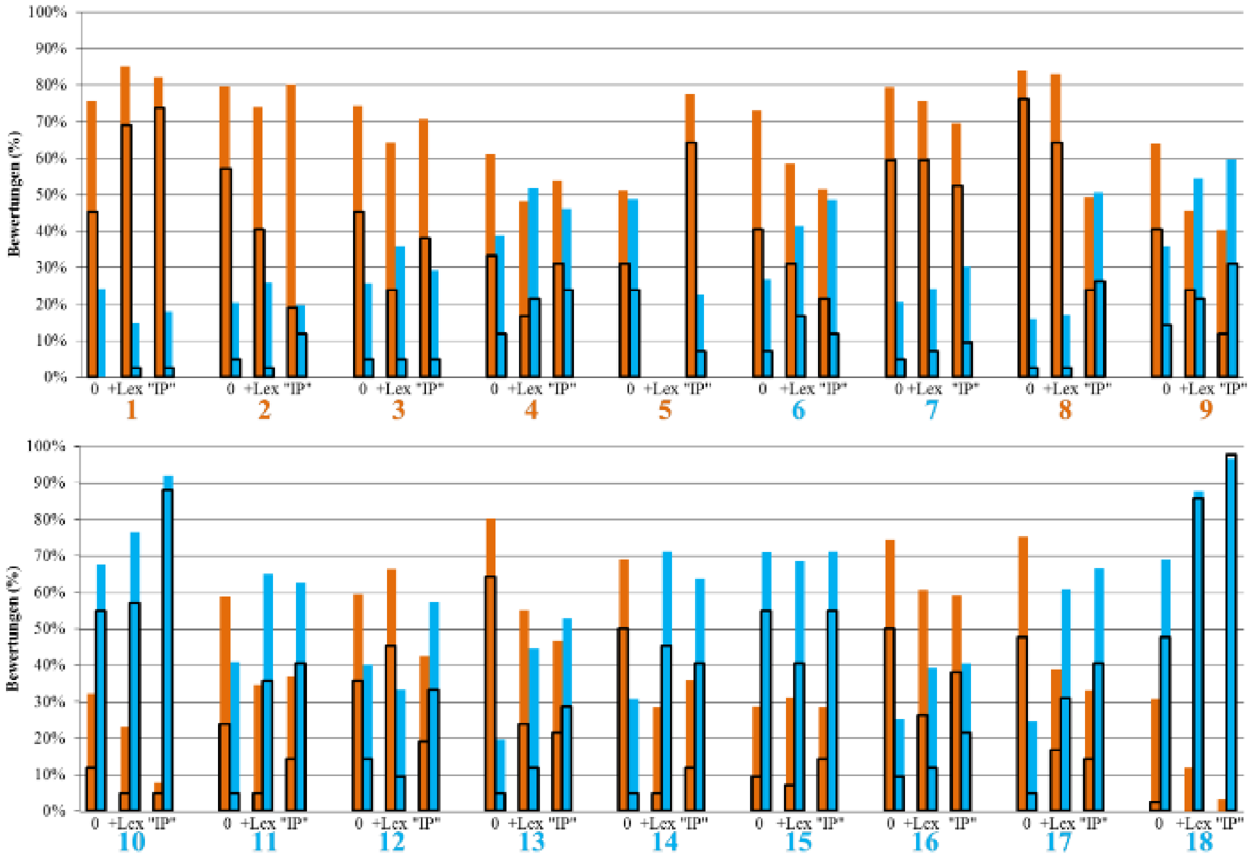


Abb. 3: Bewertungen je Stimulus. Jede Instanz (1, 2…) zeigt drei Gruppen à zwei Balken für die drei einzelnen Stimuli (0=Kontextfrei, +Lex=mit einem Lexem links, IP=mit IP-wertigem Kontext). Farb- codierung: Orange=Singular, blau=Plural. Schwarze Umrandungen: Konsistente Bewertungen (dreimalig identische Bewertung des Stimulus durch eine VPN). Die Einfärbung der Belegsiglen zeigt die Art der ursprünglichen orthografischen Auflösung

Fazit

Die vorgestellten Resultate zeigen keine eindeutige Antwort auf die Frage, welche Faktoren für Transkribierende (resp. PerzipientInnen generell) maßgeblich dafür sein könnten, die ambige Form [ˈzɐmɐ] zur einen oder anderen der beiden möglichen orthografischen Grundformen aufzulösen. Zur weiteren Untersuchung des Phänomens sind zum einen prosodische Analysen der Instanzen projektiert, zum anderen soll ein weiteres Experiment, dem die Instanzen als schriftliche Lückentexte zugrunde liegen, extraprosodische Faktoren explorieren.

Referenzen

[1] Imo, W., *Construction grammar und Gesprochene-Sprache-Forschung: Konstruktionen mit zehn matrixsatzfähigen Verben in gesprochenem Deutsch*, Tübingen: Narr, 2012.

[2] Auer, P. & S. Günthner, “Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung?,” in *Grammatikalisierung im Deutschen*, Leuschner T. & T. Mortelmans (Eds.), Berlin/New York: De Gruyter, pp. 335–362, 2005.

[3] Schmidt, Th., “Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD,” *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion,* 15, pp. 196–233.

[4] Zeschel, A., Brackhane, F. & R. Knöbl, “Reanalyse und phonetische Reduktion pragmatischer Marker mit sagen,” in *Neues vom heutigen Deutsch. Empirisch – methodisch – theoretisch.*, Eichinger, L. M. & A. Plewnia (Eds.), Berlin/Boston: De Gruyter, i. D. (=*Jahrbuch des Instituts für Deutsche Sprache* 2018).

[5] Kleiner, S., “ ‘Deutsch heute’ und der Atlas zur Aussprache des deutschen Gebrauchsstandards,” in *Regionale Variation des Deutschen. Projekte und Perspektiven*, Kehrein, R. et al. (Eds.), Berlin/Boston: De Gruyter, pp. 489–518, 2015.

[6] Draxler, Chr., “Percy – an HTML 5 framework for media rich web experiments on mobile devices,” *Proc. 12th Interspeech, Florence*, pp. 3339–3340, 2011

How to measure a pleasant voice

Jula Marie Carlsen[1], Mona Franke[1], Lena-Marie Huttner[1], André Radtke[1]

[1]Institute for Phonetics and Speech Processing, Ludwig-Maximilians-Universität Munich

j.carlsen@campus.lmu.de, m.doberass@campus.lmu.de, lenamarie.huttner@campus.lmu.de, a.radtke@campus.lmu.de

Abstract

While the evaluation of voice attractiveness has been subject of numerous studies, the mere pleasantness of a voice is a topic rarely examined. In order to investigate the pleasantness of the sound of continuous speech, we aimed to construct a listening experiment, a partial reproduction of studies already conducted. This study aims to disconnect the voice from the speaker and to evaluate solely the voice. Listeners were asked to rate 50 utterances of continuous speech for pleasantness, expressiveness, variability, intelligibility, clarity and aptitude of the voice to be used for narration in an audiobook. The subjective responses were then matched to the objective measures of the sample voices. It could be shown that voices are consistently perceived as pleasant or unpleasant by the participants. We found that deeper voices are perceived as more pleasant by both men and women. Higher jitter and lower Harmony-to-Noise values also received more favorable scores. However, the factors that contribute to a voice being perceived as pleasant are still to some degree unclear.

Introduction

Every human being prefers the sound of certain voices over others. However, the criteria by which a voice is evaluated as pleasant or unpleasant are unclear. We aim to find objective acoustic measures which will correlate to a voice being perceived as pleasant.

While pleasantness does appear as an elusive concept for scientific endeavor, the perception of voice has been of interest for a number of lines of research. Speech pathology research is engaged in finding useful parameters in voice quality to diagnose patients with dysphonia [1]. The evaluation of voice quality is predominantly done through subjective assessment. Clinical studies may therefore help inform the design of the experiment.

Further research [2] determined that f0 played a significant role in the preference for one voice over another. While this indicates little with respect to the concept of pleasantness, it does show that f0 may play a significant role in the subjective evaluation of voice [2].

One notion frequently examined is the attractiveness of a person based on their voice [3]. It is of importance to try to separate the voice and the person it belongs to for the ends of this project.

The concept of voice pleasantness is further of interest to the field of speech synthesis. One research group [4–7] aimed to find the best natural voices which could be used as models upon which to recreate a pleasant synthesized voice. Objective measurements of the model voices, such as f0, range and speaking rate, should help accomplish this objective. The results of the study are promising, yet limited. A broader approach to the evaluation of voice pleasantness may lead to more general findings.

The studies mentioned above show that pleasantness differs from the attractiveness of a voice [7] and that subjective evaluation remains the state-of-the-art method for the rating of voice.

We want to combine the different influences from the research areas to find out what contributes to a natural and healthy voice being perceived as pleasant.

Hypotheses

The aim of this study was to find objective acoustic parameters which would correlate with the perception of a voice as pleasant. This intention posed considerable difficulty for the design of the study, starting with the definition of the concept.

Pleasantness is not sufficiently defined. The Merriam Webster dictionary of English defines pleasantness as “having qualities that tend to give pleasure” [8]. While this does not indicate which qualities might be part of pleasure, it leads to a broader definition of pleasantness as a “feeling caused by agreeable stimuli” or as “what gives a sense of happy satisfaction or enjoyment”, a definition used in previous research [5]. It appears that pleasantness cannot be further divided into measurable components. Hoping to illuminate related sensations connected to *pleasantness*, auxiliary categories of perception will also be included as items in the instrument [9], described in the “Method” section.

Since different voices have different objective qualities, it is reasonable to assume that they would also vary in pleasantness [2]. We do hypothesize that the pleasantness evaluation of a voice will be homogenous among participants, which leads to the first hypothesis (H1): There are pleasant and unpleasant voices.

As studies in speech pathology and voice quality have shown, HNR, jitter, and shimmer contribute as measures to a voice being diagnosed as dysphonic [1]. We hypothesize that a voice which would rank higher on a pathological scale, would be rated as less pleasant (H2). Therefore, high HNR values will be rated more favorably (H2.1), as will voices with low jitter and shimmer values (H 2.2) [1, 4–7].

The sex of the speaker influences the values of the objective measures, most prominently f0; because of this it is likely that there will be a significant difference in the rating of male and female voices. Since a higher f0 has previously been shown to be perceived as less pleasant [2, 4] it is further to be expected that male voices will be rated as more pleasant than female voices (H3). Previous research has shown that male and female listeners often evaluate the attractiveness of a person based on their voice [3]. Since such effects cannot feasibly be excluded from occurring in this study, a fourth hypothesis (H4) states that the rating of voices will vary depending on the sex of the listener.

Method

For the purpose of measuring *pleasantness*, a subjective rating scale was devised [1, 3, 9].

In order not to significantly limit the scope of the study, the voices were also to be evaluated for *expressiveness*, *clarity*, *intelligibility*, and *variation*. All of the above items had previously been used in studies on the perception of voice and were deemed appropriate categories for the evaluation of *pleasantness* [5]. Voices were also to be evaluated for their aptitude to narrate an *audiobook* [10]. This was to give listeners an anchor for their rating, and to ensure that all voices would be evaluated on the same scale [9]. Research has shown that voice pleasantness and the completion of reading tasks are correlated, and that *audiobook* is a valid measure for such a task [4, 10]. All items were pre-tested and established measures in the field of subjective voice perception [5].

Participants were asked to evaluate voices on a seven-point scale. The questionnaire was presented to the participants in German. To analyze reliability, Cronbach’s alpha was calculated to evaluate the internal consistency of the subjective rating scale. With a value of 0.86 the reliability can be confirmed.

Stimuli

It was deemed appropriate to use utterances of continuous speech, preferably of similar content with little to no emotional component [9, 6]. The PhonDat 1 corpus of the Bavarian Archive for Speech Signals provided the stimuli for the experiment. It contains speech data of 201 German speakers recorded at four different sites (Kiel, Bonn, Bochum, Munich). All recordings have a bandwidth of 12 KHz and were scaled to an average intensity of 65db SPL [11]. The corpus contains 123 instances of readings of either *Der Nordwind und die Sonne* (the north wind and the sun) or *Die Buttergeschichte* (the butter story). Excerpts of these stories, the first sentence of *Der Nordwind und die Sonne* (M = 11.2s) and the first two sentences of *Die Buttergeschichte* (M = 17s), served as stimuli. A pretest was used to select 50 files in which the quality of the recordings and the pleasantness of the voices were evaluated. The voices used in the final experiment received more extreme ratings on the pleasantness scale in the pretest. The experiment contained voices of 25 female and 25 male speakers.

Participants

Participants were recruited through the mailing list of the Institute for Phonetics and Speech Processing at the Ludwig-Maximilians-University Munich. Listeners were required to be native German speakers and have normal hearing. Information on age and place of birth was also gathered. The age of the participants ranges from 16 to 91 years, with a median age of 28. A total of 118 female and 32 male listeners participated in the experiment.

Participants were able to leave the experiment at any time. Scores of those who chose to do so were kept in the data and included in the calculation of the results.

Acoustic measurements

The acoustic analysis of the stimuli was performed with Praat [12]. We computed values for speech rate, different pitch measures (f0, minimum f0, maximum f0, range) and measures that describe voice quality (HNR, shimmer, jitter).

Speech rate was calculated as syllables per minute. All f0-related values were measured on the psycho-acoustical scale of ERB. Minor acoustic disturbances which caused faulty f0 values in the pitch object (e.g. whistling sounds in [ʃ]) were removed manually before extracting the values. The voice quality measures jitter (local), shimmer (local) and mean HNR (dB) were calculated using the isolated vowel [a] of the word "*war*" [vaːɐ] in each utterance.

Results

The rating of the items in the questionnaire was examined using standard quantitative statistical methods in R [13], while the comments left by the participants show a valuable qualitative side to our study.

The objective measurements of one female speaker diverged considerably from the mean values of the other female speakers. The inclusion of that one voice in the statistical analysis lead to skewed values which limited the significance of the results. Ratings of this voice were therefore omitted from the evaluation. The resulting sample size was n=5000; normal distribution can be assumed. All analyses were performed using R [13].

To test the main effect, an analysis of variance was conducted which showed that the ratings for *pleasantness* differed significantly between the different speakers (F=60.43, p<0.001).

We computed correlations for all possible combinations of objective and subjective parameters with Pearson's r (see table 1). The squared values of Pearson's r then give an indication of the predictive power of the model. For each dependent variable a multivariate analysis of variance (MANOVA) was conducted to examine the significance of the correlation values with the objective parameters. The analysis shows that all values for r <-0.012 or r >0.012 are highly significant (F=18.32–294, p<0.001).

Objective Parameters

The evaluation of the results shows clear patterns: Firstly, the objective parameters were not at all able to explain the variance of our subjective parameters *clarity* and *intelligibility*. Even though some relations were significant, none of them were able to explain even 1% of the variance of the subjective parameters; for example, *intelligibility* was affected by HNR (F=10.96, p<0.001) but only 0.5% of its variance was explained by this measure (r=0.07, p<0.001). Secondly, four out of the eleven objective parameters exhibited major explanatory potential for the subjective parameters: f0, minimum f0, jitter and HNR. On the other hand, the sex of speakers and listeners, the age of the participants, speech rate, maximum f0, range and shimmer do not turn out to be predictors of *pleasantness*, *expressiveness*, *variability* and the aptitude for the narration of an *audiobook*.

Pleasantness

In regard to *pleasantness*, any effect sizes below 3% will not be taken into account in further observation. With this number in mind, *pleasantness* was only influenced by f0, minimum f0 and HNR (F=103–288, p<0.001). A regression analysis shows that these acoustic variables combined explain 8% of the cases observed (F=152.7, p<0.001) while an overall regression analysis can explain 11.3% (F=127.4, p<0.001). It is indicated that lower f0, lower minimum f0 and lower HNR contribute to a pleasant voice.

Expressiveness, variability and audiobook

HNR, f0 and minimum f0 were also the best predictors for *expressiveness*, *variability* and the aptitude for *audiobook* narration, while jitter also showed a significant impact on these subjective categories (see table 1; 18.32<F<294, p<0.001). The numbers imply that lower values for f0, minimum f0 and HNR will lead to a more favorable rating in all subjective categories (with exception of *clarity* and *intelligibility*), while higher jitter values appear to cause voices to be perceived as more pleasant. Regression analyses for all acoustical measurements combined can explain variances of 6.4% for *expressiveness* (F=67.06, p<0.001), 10.2% for *variability* (F=103.5, p<0.001) and 10.6% for *audiobook* (F=112.9, p<0.001).

Pearson's r	**Pleasantness**	**Expressiveness**	**Clarity**	**Intellegibility**	**Variability**	**Audiobook**
Syl/s	-0.01	-0.08	-0.02	-0.02	-0.13	-0.07
f0	**0.23**	**0.18**	0.01	0.04	**0.18**	**0.24**
f0 min	**0.19**	**0.19**	0.02	0.04	**0.20**	**0.21**
f0 max	0.14	0.08	-0.03	0	0.06	0.13
Range	0.05	-0.05	-0.06	-0.04	-0.09	0.01
Jitter	-0.12	**-0.20**	-0.04	-0.06	**-0.24**	**-0.16**
Shimmer	-0.08	-0.01	0	0	-0.01	-0.06
HNR	**0.16**	**0.16**	0.06	0.07	**0.16**	**0.16**

Tab.1: Correlations of subjective and objective parameters. Bold font = significant, $p < 0.001$.

Discussion

It can generally be concluded that a lower f0 is rated as more pleasant. This however does not hold true of all voices. The excluded female voice is an interesting example of this phenomenon. The voice is very deep compared to the other female voices in our dataset with an average f0 of 162Hz and a minimum f0 of 97.5. Despite the low pitch, the voice was rated negatively. Voices appear to always carry a connotation of gender, if the gender of the speaker cannot be identified, listeners appear to react to the acoustic stimulus with unease [2].

Surprisingly, the results for HNR show that huskier voices were rated better on all subjective parameters, indicating that a hoarse voice is perceived as more expressive and shows more acoustic variation, which might contribute to a more pleasant listening experience.

The result that a higher jitter value is accompanied by a higher *variability* is rather unexpected. Jitter expresses the variability in a periodic signal [1]. The subjective parameter *variability* reflects the modulation over the whole utterance whilst jitter describes the variability over a part of one vowel. According to our results, the two parameters correlate, however we have no explanation for this.

Clarity and *intelligibility* seem to be items that cannot be distinguished by listeners. This could be because of a lack of training in rating voices, or because the two items describe the same phenomenon. The two categories appear redundant when participants are rating read speech. Future studies examining the perception of freely spoken utterances are still advised to use *clarity* and *intelligibility* as items in any experiment.

It is also interesting to note that our results differ greatly from those of comparable studies [5, 6]. While we did find that f0 influenced *pleasantness*, the effect was much lower than expected. This may be explained by the size and quality of the corpus used in the experiment. Other studies on pleasantness used a much smaller data set of only female voices in their experiments and had much smaller numbers of participants than we did [4–7]. The sheer size of the corpus and the number of responses may have led to a less pronounced result.

Conclusions

The results show that there are pleasant and unpleasant voices. While pleasantness may be a subjective, individual category, different participants appeared to have similar ideas about what constituted a pleasant voice. H1 can therefore be accepted.

HNR as well as jitter values influence results in a direction opposite to the expected one, while shimmer did not have a significant influence. The observation may be attributed to the fact that none of the voices in the corpus showed any signs of pathology. The observations may therefore hold true for healthy voices in future examination. H2 and its components (H2.1. and H2.2) can be rejected. The statistical analysis further revealed that the sex of the speaker did not have any effect on the rating. H3 is therefore also rejected. Furthermore, the sex of the listener did not have any effect on the rating of male and female voices, leading to the rejection of H4. It can generally be presumed that a lower f0 is rated as more pleasant. This however does not hold true over all voices. While the results are significant, the calculated effects sizes were not overwhelmingly large. Yet, this study shows that pleasantness of voice can indeed be measured by a subjective rating scale. Factors contributing to a pleasant voice are manifold, and only a small number of these could be covered by this study alone. More exploration of the topic is desirable.

Further research may be advised to test for more categories or take an even broader approach to the subject. Researchers are encouraged to collect their own corpus data of free utterances of male and female voices, paying special attention to the secondary data gathered, such as age and accent. This study may be seen as a stepping stone towards more sophisticated explorations of voice pleasantness.

Acknowledgements

The authors would like to thank PD Dr. phil. Christoph Draxler for his continuing support of our project. We greatly appreciate all the time and effort he invested into our research. Further, we thank PD Dr. phil. Anke Werani for her input on the questionnaire and Dr. Ulrich Reubold for helping us with the statistical analysis of the results. We would further like to thank all those involved in the recording of the PhonDat 1 corpus and those who participated in the experiment.

References

[1] Schneider-Stickler, B. & W. Bigenzahn, *Stimmdiagnostik: Ein Leitfaden für die Praxis*, Wien: Springer-Verlag, 2013.

[2] Assmann, P.F. & T. M. Nearey, “Relationship between fundamental and formant frequencies in voice preference,” *Journal of the Acoustical Society of America*, 122.2, pp. EL35-EL43. 2007 Doi: https://doi.org/10.1121/1.2719045 [24.01.2018].

[3] Borkowska, B. & B. Pawlowski, “Female voice frequency in the context of dominance and attractiveness perception,” *Animal Behaviour*, 82.1, pp. 55–59, 2011.

[4] Braga D., et al., “Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality”, *Advances in Speech Technology, 14th International Workshop. June 27-29 2007, Maribor, Slovenia*, 2007.

[5] L. Coelho, et al., “Voice pleasantness: on the improvement of TTS voice quality”, *Jornadas en Tecnología del Habla,* pp. 211–214, 2008.

[6] Coelho L., et al., “Towards an Objective Voice Preference Definition for the Portuguese Language,” *Proceedings of the I Iberian SLTech 2009.* pp. 67–70, 2009.

[7] Pinto-Coelho L., et al., “On the development of an automatic voice pleasantness classification and intensity estimation system”, in *Computer Speech and Language*, 27.1, pp. 75–88, 2012.

[8] “Pleasantness” in *Merriam Webster Dictionary of English* (https://www.merriam-webster.com/dictionary/pleasantness) [05.03.2018].

[9] Barsties, B. & M. De Bodt, “Assessment of voice quality: Current state-of-the-art” in *Auris Nasus Larynx*, 42, pp. 183–188, 2014.

[10] Kreimann, J. & B.R. Gerratt, “Comparing Two Methods for reducing variability in Voice Quality Measurements,” *Journal of Speech language and Hearing Research*, 53.3, pp. 803–812, 2011.

[11] Hinterleitner F., et al., “An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks,” *Proceedings Blizzard Challenge Workshop. International Speech Communication Association (ISCA*). 2011.

[12] Boersma, P. & D. Weenink, *Praat: doing phonetics by computer* [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/, 2018.

[13] RStudio Team, *RStudio: Integrated Development for R.* RStudio, Inc., Boston, MA URL http://www.rstudio.com/, 2016.

Die Leseaussprache des Deutschen in Luxemburg

François Conrad[1], Judith Manzoni[2]

[1]Leibniz Universität Hannover, [2]Universität Trier

francois.conrad@germanistik.uni-hannover.de, manzoni@uni-trier.de

Abstract

The present study analyses phonetic interferences between German and Luxembourgish in a German text ('Nordwind und Sonne') read by 34 native speakers of Luxembourgish. The number of German or Luxembourgish variants in terms of vowels, consonants and phonological processes were counted and correlated with sociolinguistic variables. Age differences and a more direct reference to and contact with German are the significant factors explaining the interferences.

Einleitung

Die (sprach)geographische Lage Luxemburgs sowie seine Geschichte erklären die komplexe Sprachensituation im Land. An Deutschland, Belgien und Frankreich grenzend, gehört Luxemburg zur Germania, bildet aber zugleich die Grenze zur Romania. Neben den offiziellen Sprachen Französisch (Fr.) und Deutsch (Dt.) gilt die Nationalsprache Luxemburgisch (Lux.) – seit dem Sprachengesetz von 1984 zur Sprache erhoben – als identitätsstiftend. Sie gehört dialektologisch betrachtet dem Westmoselfränkischen an, wodurch sie eine große Nähe zum Dt. aufweist. Doch sie wird nicht (mehr) von ihm überdacht [1]. Die Alphabetisierung erfolgt in den Schulen jedoch weiterhin auf Deutsch, so dass jede(r) Sprecher*in neben Lux. fließend Dt. (und die weitere Schulsprache Fr.) beherrscht. Unter *Luxemburgisch* wird hier die überregionale Standardsprache verstanden, die als Verkehrssprache unter lux. Muttersprachler*innen dient.

Ziel der Studie ist die erstmalige Beschreibung der Leseaussprache des Dt. von Luxemburger*innen und damit der Interferenzen lux. Sprecher*innen im Dt.

Vergleich Deutsch vs. Luxemburgisch

Für die Aussprache des Standarddeutschen (StD) wird sich hier auf die Arbeiten von [2], [3] und [4] gestützt, während sich die des Lux. an [5], [6], [7] und [8] orientiert.

Trotz vieler Gemeinsamkeiten zwischen beiden Sprachen auf segmentaler Ebene zeigen sich Unterschiede. Dazu gehören im Konsonantismus der Zusammenfall im Lux. von /ç/ und /ʃ/ zu /ɕ/ [9] und die Aussprache von /ʀ/ als [χ] vor stimmlosen Plosiven innerhalb einer Silbe (*Spo[χ]t*) und, v.a. bei älteren Sprecher*innen, im absoluten Silbenauslaut (*mir* [mi:χ]).

Auf der Ebene der phonologischen Prozesse sind bspw. die im lux. konsequent fehlende Schwa-Elision (*spillen* [ˈʃpilən] ‚spielen') und die regressive Stimmhaftigkeitsassimilation (bei fehlendem glottalen Plosiv im vokalischen Anlaut, *ech och* [əʑ‿ɔχ] ‚ich auch') zu nennen.

Im Vokalismus stellen etwa der fehlende Gespanntheitskontrast zwischen Lang- und Kurzvokalen bei /i, o, u/, der Qualitätsunterschied bei /a/ (kurz [ɑ], lang [ɑ̟:]), die unterschiedliche Qualität des halb-offenen Vordervokals (dt. [ɛ] vs. lux. [æ]) und die größere Anzahl an Diphthongen (acht) Differenzen zum StD dar.

Folgende Transkription zeigt den untersuchten Text ('Nordwind und Sonne') mit einer maximalen Anzahl potenzieller lux. Interferenzen. Auf der Grundlage der genannten Literatur entspricht er somit einer prototypischen lux. Aussprache des Dt. (vokalische Interferenzen: orange, konsonantische Interferenzen: blau; phonologische Prozesse: grün).

[ɑɪnst ˈʃtʁitən ziɕ ˈnoχtvind‿unt ˈzonə vɛːɐ̯ fon ˈiːnən ˈbɑɪdən voːl dɛːɐ̯ ˈʃtæχkəʁə ˈvɛːʁə

ɑlz‿ɑɪn ˈvɑndɐʁɐχ dɛːʁ‿in ˈɑɪnən ˈvɑʁmən ˈmɑntəl gəˈhylt vɑ̟ːχ dæs ˈveːgəs dɑ̟ˈhɛːχkɑ̟ːm ziː ˈvuʁdən ˈɑɪniɕ dɑs ˈdɛːɐ̯ʑeːnijə fyɐ̯ deːn ˈʃtæχkəʁən ˈgæltən ˈzoltə dɛːɐ̯ deːn ˈvɑndɐʁɐχ ˈtswiŋən ˈvyʁdə ˈzɑɪnən ˈmɑntəl ˈɑptsuneːmən dɛːɐ̯ ˈnoχtvint bliːs mit ˈɑlɐ mɑχt ˈɑ̟ːbɐ ʑe mɛːʁ‿ɛːɐ̯ bliːs ˈdæsto ˈfæstɐ ˈhyltə ziɕ dɛːɐ̯ ˈvɑndɐʁɐχ in ˈzɑɪnən ˈmɑntəl ɑɪn

ˈæntliɕ gɑ̟ːp dɛːɐ̯ ˈnoχtvint deːn kɑmbv‿ɑʊf

nun æɐ̯ˈvæʁmtə diː ˈzonə diː luft mit ˈiːʁən ˈfʁoɪntliɕən ˈʃtʁɑ̟ːlən unt ʃoːnɑ̟ːχ ˈveːnigən ɑʊʀənˈblikən tsoːχ dɛːɐ̯ ˈvɑndɐʁɐχ ˈzɑɪnən ˈmɑntəl ɑʊs

dɑ̟ː ˈmustə dɛːɐ̯ ˈnoχtvint ˈtsuːgeːbən dɑs diː ˈzonə fon ˈiːnən ˈbɑɪdən dɛːɐ̯ ˈʃtæχkəʁə vɑ̟ːχ]

Methodik und Sample

Datengrundlage bilden die Aufnahmen von 34 lux. Muttersprachler*innen (20 Frauen, 14 Männer) aus Luxemburg-Stadt und direkter Umgebung. Sprecher*innen dieser Region sprechen Zentrallux., das als standardähnliche Umgangsform verstanden wird [10]. Die TeilnehmerInnen wurden in drei Altersgruppen unterteilt: eine jüngere (20–30 Jahre, Ø: 25, s=3,17), eine mittlere (40–64, Ø: 53, s=3,77) und eine ältere Generation (≥65, Ø: 80, s=6,40). Weiterhin wurden die Schulbildung (klassisches oder technisches Gymnasium) und ein besonderer Bezug zu Deutsch (etwa Deutschlehrer*in oder Aufenthalt in Deutschland) abgefragt.

Die Teilnehmer*innen lasen in ruhiger Umgebung die Fabel 'Nordwind und Sonne' vor, ohne den genauen Gegenstand der Studie zu kennen.

Die Analyse erfolgte auditiv und akustisch mithilfe von Praat (6.0.19 [11]; insbesondere Messung von F1 und F2). Für die Formantwerte wurden diejenigen von [7] als Vergleich herangezogen. Konsonantische und vokalische Zweifelsfälle wurden ausgeschlossen.

Nachstehend sind die untersuchten konsonantischen Variablen, die phonologischen Prozesse und die vokalischen Variablen (jeweils mit Beispielwörtern aus dem Text in Klammern) aufgelistet (Tab. 1).

Variable	**Varianten**
<r> __ stl. Plosiv, __$ (*Nordwind, war)*	[ɐ]/ø vs. [χ]
/ç/ (*sich*)	[ç] vs. [ɕ]
<j-> $__ (*derjenige*)	[j] vs. [ʑ]
<g> nach Vordervok., intervok. (*derjenige*)	[g] vs. [j/ʝ]
<g> nach Hintervok., intervok. (*Augen*)	[g] vs. [ʀ]
<g> __# (/g/-Spirantisierung) (*zog*)	[k] vs. [χ]
<w> nach [t͜s] (*zwingen*)	[v] vs. [w]
Schwa-Elision (*abzunehmen*)	[n̩] vs. [ən]
Regressive Stimmhaftigkeits-assimilation (*mi[d]‿aller ,mit aller'*)	ø vs. Assimilation
Qualität /i/ (*stritten*)	[ɪ] vs. [i]
Qualität /u/ (*Luft*)	[ʊ] vs. [u]
Qualität /o/ (*Sonne*)	[ɔ] vs. [o]
Qualität /ɛ/ (*fester*)	[ɛ] vs. [æ]
Qualität /a/ (*Mantel*) /a:/ (*gab*)	[a] vs. [ɑ] [a:] vs. [a̝:]
Qualität /a/ in /a͜ʊ/ (*Augenblicken*)	[a] vs. [ɑ]
Qualität /a/ in /a͜ɪ/ (*einig*)	[a] vs. [ɑ]

Tab. 1: Konsonantische (n=7) Variablen, phonologische Prozesse (n=2) und vokalische (n=7) Variablen (links), dt. vs. lux. Variante (rechts).

Für jede Variable wurden mehrere Token im Text untersucht (etwa *stritten* und *sich* für die Qualität von /i/). Im Hinblick auf die Gespanntheit wurde zum Vergleich zudem jeweils ein Bezugswort gewählt (z.B. *blies*).

Die betroffenen Segmente wurden mit *0* (lux. Aussprache) oder *1* (std. Aussprache) kodiert (abhängige Variable). Als unabhängige Variablen gelten Alter (bzw. Generation), Bildung, Geschlecht und Bezug zu Deutsch. Die statistische Auswertung erfolgte mit RBRUL (multivariate Analyse) sowie SPSS (A und *t*-Test).

Ergebnisse

Die folgende Tabelle stellt die Ergebnisse der multivariaten Analyse vor (Tab. 2). Während Bildung und Geschlecht keinen signifikanten Einfluss auf die Interferenzen zeigen, spielen das Alter und der Bezug zu Deutsch in allen lautlichen Teilbereichen eine große Rolle bei der Erklärung der Interferenzphänomene innerhalb der dt. Leseaussprache lux. Sprecher*Innen.

	Faktor-gruppe	**signifikante Faktoren**	**p**	r^2
Gesamt-sample	feste Faktoren	*Bezug zu Deutsch* *Alter*	< 0,001 < 0,001	0,16
	Zufalls-faktoren	*SprecherInnen, Wörter*		0,45
	Gesamt			0,61
Kons. Variablen (inkl. Proz.)	feste Faktoren	*Alter* *Bezug zu Deutsch*	< 0,001 < 0,001	0,19
	Zufalls-faktoren	*SprecherInnen, Wörter*		0,45
	Gesamt			0,64
Vokal. Variablen	feste Faktoren	*Bezug zu Deutsch* *Alter*	< 0,01 < 0,05	0,11
	Zufalls-faktoren	*SprecherInnen, Wörter*		0,42
	Gesamt			0,53

Tab. 2: Ergebnisse der multivariaten Analyse.

Im Folgenden werden diejenigen Variablen näher besprochen, die in Bezug auf die signifikanten Faktoren besonders relevant sind (zunächst die Konsonanten, anschließend die phonologischen Prozesse, dann die Vokale).

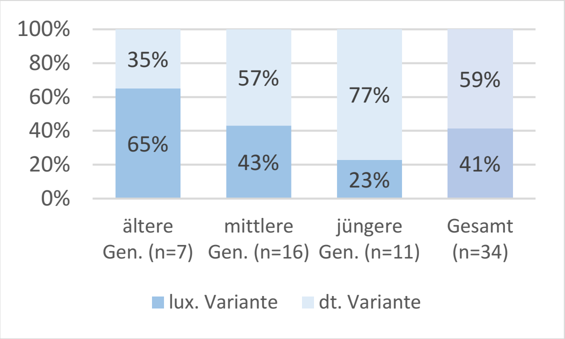


Abb.1: Einfluss des Alters auf die Aussprache konsonantischer Elemente (ANOVA, F(2, 31)=25,27, p<0,001, η =0,79).

Die drei untersuchten Generationen unterscheiden sich stark in ihren Anteilen dt. und lux. konsonantischer Varianten (Abb. 1). Mit abnehmendem Alter steigt der Anteil dt. Varianten stark an. In der jüngeren Generation werden drei von vier möglichen Segmenten in der standarddeutschen Aussprache realisiert, während es in der älteren Generation nur einer von drei ist.

Besonders stark zeigt sich dieser Unterschied bei der Realisierung von /ʀ/ im Silbenauslaut und vor stimmlosem Plosiv (als [ɐ] oder [χ]). Bei dieser Variable haben sich die Anteile nahezu umgekehrt: Die ältere Generation realisiert mit 86% fast alle möglichen /ʀ/ als [χ], die mittlere Generation noch zu 50%. Die jüngere Generation hingegen realisiert nur noch 18% konsonantische Aussprachen, währende die vokalisierte Form [ɐ] in der Aussprache des Dt. klar vorherrscht (ANOVA, F(2,31)=22,76, p<0,001, η=0,77).

Auch ein unterschiedlicher Bezug zur dt. Sprache hat einen Einfluss auf die Wahl der konsonantischen Varianten (Abb. 2):

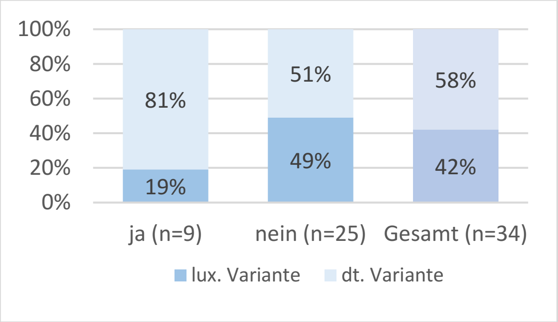


Abb. 2: Einfluss des Bezugs zu Deutsch auf die Aussprache der konsonantischen Variablen (t-Test: t(27)=5,05, p<0,001, r=0,7).

Ein sehr ähnliches Bild zeigt sich auf der Ebene der phonologischen Prozesse bei der Schwa-Elision (Abb. 3 und 4), während sich insgesamt nur sehr wenige Belege für die regressive Stimmhaftigkeitsassimilation finden lassen.

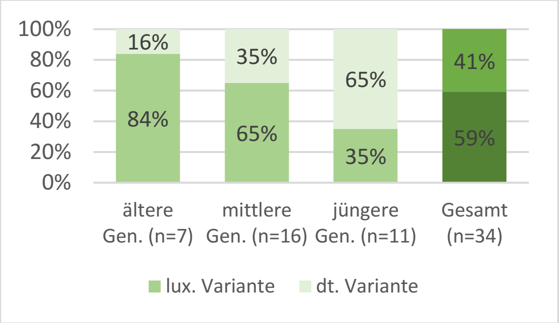


Abb. 3: Einfluss des Alters auf die Aussprache der phonologischen Prozesse (ANOVA, F(2, 31)=6,68, p<0,01, η=0,55).

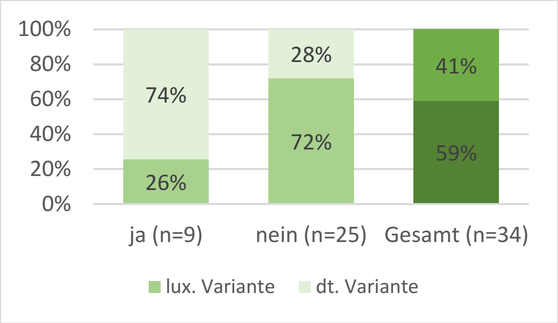


Abb. 4: Einfluss des Bezugs zu Deutsch auf die Aussprache der phonologischen Prozesse (t-Test: t(27)=5,78, p<0,001, r=0,74).

Auch im vokalischen Bereich manifestiert sich dieses Bild, wenngleich die Ergebnisse etwas weniger stark ausfallen. Erneut steigt die Anzahl standarddeutscher Realisierungen mit sinkendem Alter (Abb. 5).

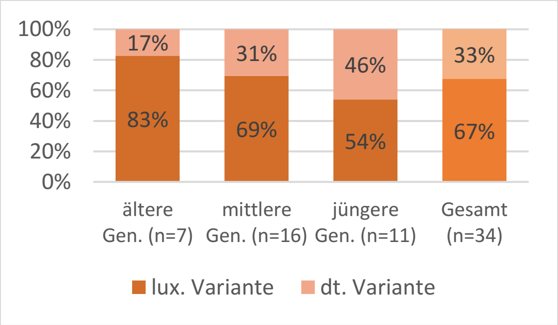


Abb. 5: Einfluss des Alters auf die Aussprache vokalischer Elemente (ANOVA, F(2, 31)=5,44, p<0,01, η=0,51).

Besonders auffällig ist hierbei der Anteil ungespannter Realisierungen von /i, o, u/. Bei der älteren Generation liegt er bei 0%, bei der mittleren Generation bei 22%, bei der jüngeren Generation jedoch bei 75% (Kruskal-Wallis, H(2)=20,77, p<0,001, η^2=0,61).

Auch ein Bezug zu Dt. erhöht erneut den Anteil standarddeutscher Realisierungen (vgl. Abb. 6).

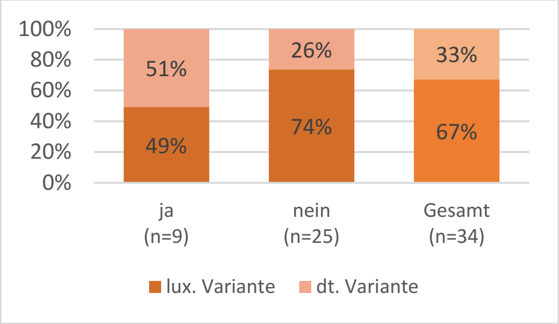


Abb. 6: Einfluss des Bezugs zu Deutsch auf die Aussprache vokalischer Elemente (t-Test, t(32)=3,41, p<0,01, r=0,54).

Diskussion und Fazit

Das Alter ist der entscheidende Faktor für die Leseaussprache des Dt. von Luxemburgern: je jünger die Sprecher*innen, desto phonetisch näher ist die Aussprache der stddt. Norm. Ebenfalls ausschlaggebend für eine interferenzärmere Aussprache ist der Kontakt mit dem Dt.: Je eher die Sprecher*innen einen direkten Bezug zum Dt. haben, desto näher ist ihre Aussprache dem StD. Diese Unterschiede gelten für alle untersuchten Lautbereiche, wenngleich sie bei den Vokalen etwas weniger ausgeprägt sind als bei Konsonanten und phonologischen Prozessen.

Diese Unterschiede lassen sich zum einen dadurch erklären, dass die ältere Generation zur Zeit der dt. Besatzung im Zweiten Weltkrieg aufgewachsen ist und sich sprachlich von den dt. Nachbar*innen abgrenzen wollte [12]. Zum anderen hatten die älteren Sprecher*innen abgesehen von den Kriegsjahren außerhalb der Schule kaum Gelegenheit Dt. zu sprechen. Die jüngere Generation hingegen kam aufgrund der Rezeption überwiegend dt. Medien [13] (insbesondere Fernsehen) schon sehr früh zumindest passiv mit Dt. in Kontakt. Überdies ist vermutlich auch eine höhere Mobilität seitens der jüngeren Sprecher*innen (Studium, Freizeitaktivitäten in Deutschland) für einen aktiven Kontakt mit dem Dt. verantwortlich. Schließlich scheint das Fr. in den letzten Jahren, zusammenhängend unter anderem mit der hohen Anzahl an französischsprachigen Grenzpendler*innen, an Prestige zu verlieren [6]. Dies fördert erneut einen stärkeren Einfluss der Kontaktsprache Dt.

Das Ergebnis, dass Sprecher*innen mit einem besonderen Bezug zu Dt. eine interferenzärmere Aussprache aufweisen, ist wenig überraschend. Es ist jedoch auffällig, dass überwiegend Sprecher*innen der jüngeren und mittleren Generation einen solchen Bezug aufweisen (8 von 9), was wiederum auf die unterschiedliche Sprach(en)situation der letzten Jahrzehnte hindeutet.

Ein überraschendes Ergebnis ist die geringe Belegzahl der regressiven Stimmhaftigkeitsassimilation. Dieses Sandhi-Phänomen ist ein stabiles Merkmal der gesprochenen lux. Sprache [5]. Sehr wahrscheinlich haben hier die Schrift (Vorleseaufgabe) und somit ggf. auch das Sprechtempo einen starken Einfluss auf dieses Phänomen.

Als stabile Variablen in dieser ersten Studie zur Leseaussprache des Dt. in Luxemburg haben sich insbesondere zwei Elemente herauskristallisiert: /ʀ/ vor stimmlosem Plosiv wird fast durchgängig als Frikativ [χ] realisiert (*No[χ]twind*) und die Qualität von /a/ wird sehr konsequent mit seiner Opposition [ɑ] vs. [ạː] in das Deutsche übertragen. Wenn von einer lux. Varietät des Dt. gesprochen werden kann – der vorliegende Beitrag möchte zu dieser Diskussion erste empirische Indizien beisteuern – können diese beiden Elemente als feste Bestandteile dieser Varietät angesehen werden.

Referenzen

[1] Gilles, P., “Mündlichkeit und Schriftlichkeit in der luxemburgischen Sprachengemeinschaft,” in *Medien des Wissens. Interdisziplinäre Aspekte von Medialität*, Mein, G. & H. Sieburg (Eds.), Bielefeld: transcript, pp. 43–64, 2011.

[2] Kohler, K., “German,” in *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, International Association of Phonetics (Ed.), Cambridge: CUP, pp. 86–89, 1999.

[3] Kleiner, S., *Atlas zur Aussprache der deutschen Gebrauchsstandards* (AADG): Unter Mitarbeit von Ralf Knöbl, http://prowiki.ids-mannheim.de/bin/view/AADG/ .

[4] Duden: *Aussprachewörterbuch*, Mannheim: Dudenverlag, ^{6}2015.

[5] Gilles, P. & J. Trouvain, “Luxembourgish: Illustration of the IPA,” *Journal of the International Phonetic Association*, 43(1), pp. 67–74, 2013.

[6] Conrad, F., *Variation durch Sprachkontakt: Lautliche Dubletten im Luxemburgischen*, Frankfurt Main: Peter Lang, 2017.

[7] Thill, T., *Une étude acoustique et comparative sur les voyelles du luxembourgeois,* Dissertation, Université du Luxemburg, Luxemburg, 2017.

[8] Manzoni, J., "Kontrastiv-phonetische Analysen: Luxemburgisch-Deutsch," Magisterarbeit, Univ. Trier, Trier, unveröff.

[9] Conrad, F., "Der Zusammenfall von /ʃ/>/ɕ/</ç/ im Luxemburgischen," in *Proceedings der Tagung Phonetik & Phonologie 2018*, pp. 29–32, 2018.

[10] Gilles, P. & C. Moulin, "Luxembourgish," in *Germanic Standardizations: Past to Present*, Vandenbusche, W. (Ed.), Amsterdam/Philadelphia: John Benjamins, pp. 303–330, 2003.

[11] Boersma, P. & D. Weenink, *Praat: Doing phonetics by computer*, 2018. [Computer program], version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/.

[12] Berg, G., *"Mir wëlle bleiwe, wat mir sin": Soziolinguistische und sprachtypologische Betrachtungen zur luxemburgischen Mundart*, Tübingen: Niemeyer, 1993.

[13] Fehlen, F., *Baleine Bis: Une enquête sur un marché linguistique multilingue en profonde mutation – Luxemburgs Sprachenmarkt im Wandel*, Luxembourg; SESOPI Centre intercomm., 2009.

Devoicing of /z/ in the Bavarian dialect: a large-scale study

Johanna Cronenberg[1], Laura-Annika Haller[1], Iveta Patáková[1], Christoph Draxler[1]

[1]Ludwig-Maximilians-Universität München

johanna.cronenberg@campus.lmu.de, l.haller@campus.lmu.de,
iveta.patakova@campus.lmu.de, draxler@phonetik.uni-muenchen.de

Abstract

This study is concerned with a phonetic feature of Bavarian dialect: the devoicing of /z/ in word and syllable-initial positions (e.g. German Sonne [sɔnə] *'sun'). In a large scale corpus analysis with 786 speakers and more than 355,000 phonetic segments, we examined differences in the pronunciation of the alveolar sibilants produced by speakers from small towns in Bavaria, from Munich, and from other German federal states. The analysis was based on measurements of the fundamental frequency. The results showed that adolescents from Bavarian small towns devoice the examined fricatives much more than their peers from Munich. In addition, the latter exhibit a degree of voicelessness similar to a control group of subjects from other German federal states. This indicates that Bavarian metropolitans are more oriented towards a Standard German pronunciation. We discuss the analysis with respect to Big Data related challenges and other possible factors that could have influenced the results.*

Introduction

Literature Review

In Germany, the use and spread of regiolects is influenced by the size of the speech community. In small towns[1], there is a relatively homogeneous population which often tends to use more dialect and demonstrates a more positive attitude towards it [1]. In contrast, people living in larger cities are surrounded by speakers of other German dialects and of other languages. While people from Bavarian villages or small towns often learn the dominant regiolect as their first (and sometimes only) language, people living in cities like Munich are more oriented towards a standardised pronunciation due to influences from other German varieties and a general diminished use of dialect [2]. These attitudes towards dialects are especially observable in the younger generations.

Aim of the Study

First, this study aims to examine the differences between adolescents from Bavaria and from other German federal states with regard to their pronunciation of the alveolar sibilants. While the Bavarian dialect is known for devoicing the voiced fricative /z/ in word and syllable-initial positions [3], the same sound is phonologically voiced in Standard German. For example, the word *Sonne* 'sun' is usually produced as [zɔnə] in Standard German, but as [sɔnə] in the Bavarian dialect. The second aim of this study is to investigate whether there are significant differences in this phonetic feature between subjects living in Bavarian small towns and subjects living in Munich.

Hypotheses

Two hypotheses can be postulated on the basis of the literature cited above [1-3]:

H1: Adolescents from Bavarian small towns devoice alveolar sibilants in word and syllable-initial positions to a stronger degree than their peers living in the city of Munich.

H2: Adolescents from the city of Munich produce voiced alveolar sibilants in word and syllable-initial positions, i.e. they are more oriented towards Standard German speakers.

As for word and syllable-final positions, we expect almost no voicing due to the final obstruent devoicing of Standard German.

[1] [1] draws the line between small towns and larger cities at a population of more than 50,000 inhabitants.

Material & Participants

Corpora

Subsets of two readily available corpora from the BAS CLARIN Repository were created and used for this study: Ph@ttSessionz[2] and RVG-J.[3] The RVG-J Corpus was recorded in 2001 at the Institute of Phonetics and Speech Processing of the University of Munich and contains both read and non-scripted German utterances by 182 adolescents (age range 13 to 20) who attended public schools in Munich. Of these students, 175 assessed themselves as speakers of Bavarian dialect.

Ph@ttSessionz is a speech database which contains recordings of 1019 adolescent speakers of German (age range 12-20). The recordings, which contain similar contents as the recordings in RVG-J, were performed in public schools in 45 locations in Germany. In this corpus, 213 speakers were from Bavaria.

Participants

From the corpora described above, three speaker groups were formed: “city”, consisting of 175 speakers (88 female) from Munich, “town”, consisting of 213 speakers (140 female) from Bavarian small towns with less than 50,000 inhabitants (e.g. Altötting), and “control”, consisting of 398 speakers (195 female) from other German federal states.

Analysis

Data Processing

The recordings from all chosen speakers were reduced to read utterances only (i.e. short sentences). They were then transformed into emuDBs and run through the following BAS Web Services [4]: G2P for tokenization, G2P for pronunciation, MAUS with SAMPA encoding, and Pho2Syl. This process resulted in segmentations into words, syllables, phonemes, and phones. The Web Services were run with default settings apart from choosing German as language.

The emuDBs were then queried using EQL [5]. The queries delivered all segments labelled as [s] or [z] in word-initial, word-final, syllable-initial, and syllable-final positions. For the initial positions, the alveolar sibilant had to be followed by a vowel, whereas for the final positions, the fricative had to be preceded by a vowel.

Voicing Measurements

Due to the massive amount of data that had to be processed (14.9 GB), voicing was measured by simply calculating the fundamental frequency. Therefore, the ksvF0 algorithm with default settings [6] was applied to all segments that resulted from the queries. The fundamental frequency was measured at the temporal midpoint of the segment and at five measuring points each before and after the midpoint. The resulting measurements were then recoded in a binary manner, such that whenever the function returned zero as a value for one measuring point, this point was taken to be voiceless; all other frequency values were considered voiced. The eleven F0 measurements per segment were used to compute the proportion of voicing.

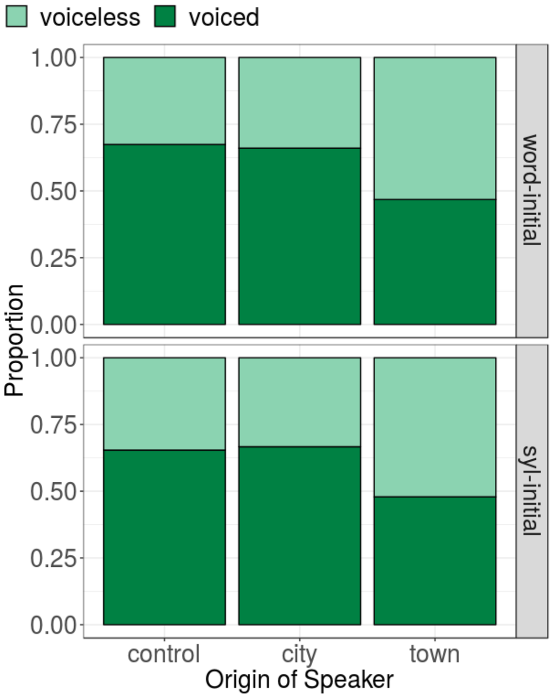


Fig. 1: Proportion of voicing in alveolar sibilants for speakers from Bavarian small towns (“town”), Munich (“city”), and other German federal states (“control”), for word and syllable-initial positions.

Results

Figure 1 shows that in word and syllable-initial positions speakers of the group “town” exhibit a far higher amount of voicelessness in

[2] http://hdl.handle.net/11022/1009-0000-0000-CC6A-4

[3] http://hdl.handle.net/11022/1009-0000-0004-AE1D-9

their alveolar sibilants than speakers from Munich. This is the expected result (see H1), considering that adolescents from Bavarian towns are likely to speak their dialect to a greater degree than their peers from Munich.

Moreover, speakers of the group "city" show almost the same degree of voicing in these segments as speakers of the group "control". This result confirms the second hypothesis, i.e. that adolescents from Munich are more oriented towards a Standard German pronunciation.

With regard to the final positions (see Figure 2), all three speaker groups show similar proportions of voicing in the alveolar sibilants. These values are unexpectedly high (around 50%), considering that the final obstruent devoicing of Standard German should usually lead to fairly voiceless fricatives at the end of both words and syllables.

For now, the significance of these results cannot be confirmed by statistical measurements, since several General Linear Mixed Models (using the *lme4* package in R) did not converge.

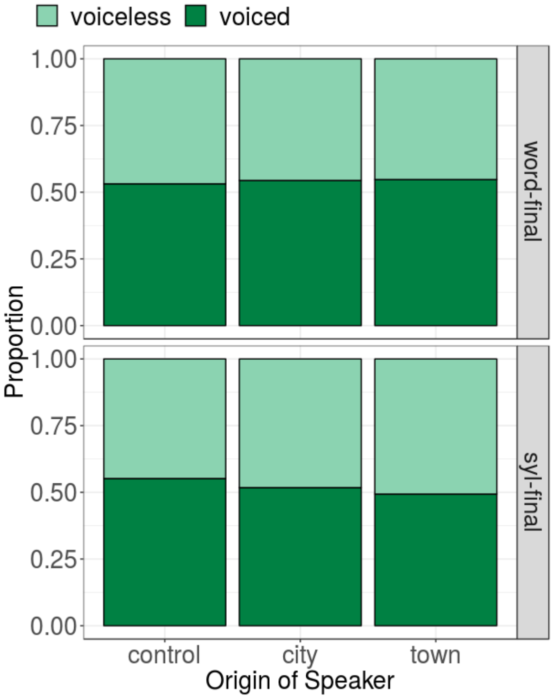


Fig. 2: Proportion of voicing in alveolar sibilants for speakers from Bavarian small towns ("town"), Munich ("city"), and other German federal states ("control"), for word and syllable-final positions.

Discussion

Managing a data volume of approx. 14.9 GB results in significant challenges related to Big Data. The higher the data volume, the greater the storage and analysis issue. Furthermore, the use of two large speech corpora leads to several different types of information that need to be analysed. Additionally, working with such a large amount of data means longer processing times.

To our knowledge, this study is the largest one in terms of amount of data in the entire field of phonetics, since it includes utterances of 786 speakers, meaning that more than 355,000 phonetic segments were analysed.

The results of the data processing were not manually corrected because of the large amount of data. On the one hand, this means that the boundaries of the segments could be partially erroneous and include acoustic overlap from other phonetic segments. On the other hand, the analysis relies on the values that were delivered by the ksvF0 function. Since this function was not personalised for each speaker, inaccurate F0 values may be included.

In future, more detailed investigations, other factors such as age and gender of participant, speaking rate and phonetic context will be taken into account. The phonetic context may well have a great influence on the results shown for word and syllable-final positions. When the observed segment in a final position is followed by a voiced segment, the voicing may also have shifted to the preceding sibilant itself.

References

[1] Steinegger, G., *Situativer Sprachgebrauch und Spracheinschätzung in Österreich und Südtirol. Auswertung einer Umfrage*, Wien: Universität Wien, 1997.

[2] Renn, M., *Die Mundart im Raum Augsburg: Untersuchungen zum Dialekt und zum Dialektwandel im Spannungsfeld großstädtisch-ländlicher und alemannisch-bairischer Gegensätze*, Heidelberg: Universitäts-Verlag Winter, 1994.

[3] Piroth, H. G. & P. M. Janker, "Speaker-dependent differences in voicing and devoicing of German obstruents," *Journal of Phonetics*, 32, pp. 81–109, 2004.

[4] Kisler, T. et al., *BAS Speech Science Web Services – An Update of Current Developments*, *LREC*, Portoroz, 2016.

[5] Winkelmann, R. et al., "EMU-SDMS: Advanced Speech Database Management and Analysis in R," *Computer Speech & Language*, 45, pp. 392–410, 2017.

[6] Schaefer-Vincent, K., "Pitch period detection and chaining: method and evaluation," *Phonetica*, 40, pp. 177–202, 1983.

Sprechtempo in Lehrvideos auf YouTube

Alexandra Ebel[1]

[1]Martin-Luther-Universität Halle-Wittenberg

alexandra.ebel@sprechwiss.uni-halle.de

Abstract

*YouTube-Videos boomen und werden auch zur Aneignung von Wissen genutzt. Seitens der Didaktiker*innen stellt sich die Frage, welche Parameter einen Einfluss auf den Erfolg von Lehrvideos auf YouTube haben könnten. Um dieser Frage nachzugehen, werden in einem Forschungsprojekt verschiedene sprachliche und sprecher*innenspezfische Merkmale in erfolgreichen und weniger erfolgreichen Videos untersucht. Eines dieser Merkmale ist das Sprechtempo. Es wurde für vierzig YouTube-Videos berechnet, um einen Zusammenhang mit dem Erfolg der Videos zu prüfen.*

Einleitung

Die Videoplattform YouTube dient längst nicht mehr nur zur Unterhaltung seiner Nutzer*innen, sondern besitzt auch wissensvermittelnde Funktionen. „Wissenskanäle boomen. […] Seit etwa fünf Jahren wird die Qualität von Wissensvideos auf YouTube besser. Wissenschaft ist dort so etwas wie Popkultur geworden.“ [1] In den *Top 200 Tools for Learning 2018* belegt YouTube wie schon in den beiden Vorjahren Platz 1 [2]. Lernen mit YouTube profitiert von der zeitlichen und örtlichen Unabhängigkeit, einer breiten Themenvielfalt sowie der multisensorischen Darstellung der Inhalte.

Verschiedene Studien bestätigen den wachsenden Einfluss des Internets und auch YouTubes [3, 4]. Kinder und Jugendliche wachsen „heute wie nie zuvor in einer mediatisierten, digitalen Gesellschaft auf.“ [5] Natürlich spielen Unterhaltungsvideos für sie eine große Rolle, aber auch Lehrvideos zu schulischen Themen sind beliebt. Ein Großteil der Kinder, die älter als zehn Jahre sind, schaut mindestens einmal wöchentlich Videos zu Schulthemen [6]. Auch für Lehrende können sich Vorteile durch die Nutzung von YouTube-Videos ergeben: In Präsenzveranstaltungen können mittels Video Vorgänge gezeigt werden, die aus verschiedenen Gründen nicht live durchgeführt werden können, z. B. weil sie zu gefährlich wären oder eine Exkursion notwendig wäre. Durch Lehrveranstaltungsaufzeichnungen können Inhalte zeit- und ortsunabhängig wiederholt werden. Oder Videos werden ergänzend zur Lehrveranstaltung produziert und als Blended Learning Inhalte auf Lernplattformen eingebunden [7]. Noch sind Wissens- und Wissenschaftsvideos auf YouTube allerdings wenig erforscht [8, 9]. Einerseits liegt das an der Personalisierung der Suchergebnisse, die eine systematische Analyse erschweren [10], und andererseits an methodischen Herausforderungen [11].

Da mittlerweile auch in den Fachdidaktiken verschiedener Disziplinen ein Interesse an YouTube-Videos aufkommt, wächst der Wunsch nach Anhaltspunkten, wie das nutzerseitige Verständnis des präsentierten Wissens sichergestellt werden kann. Es gibt bereits einige wissenschaftliche Auseinandersetzungen, die versuchen, die Merkmale erfolgreicher Videos zu ermitteln, oft bleibt es im Hinblick auf die sprecherisch-sprachliche Gestaltung aber bei Allgemeinplätzen.

Auch die Frage, welche Sprechgeschwindigkeit angestrebt werden sollte, um das Verständnis der Lerninhalte in YouTube-Videos zu unterstützten, wird in der Literatur unterschiedlich beantwortet: So empfiehlt zum Beispiel Becher lediglich generalisierend eine ‚angemessene‘ Sprechgeschwindigkeit [12], ebenso tut es Merkt [13]. Auf die Besonderheiten der Plattform YouTube, wie beliebige Reproduzierbarkeit des Schallereignisses oder die Möglichkeit, Videos mit verringerter oder erhöhter Geschwindigkeit abzuspielen, wird dabei jedoch nicht eingegangen. Welbourne & Grant konnten in einer Studie ermitteln, dass Videos mit schnell sprechenden Sprecher*innen erfolgreicher waren, als die, in denen langsamer präsentiert wurde [9], und vermuten als Grund die Wiederholbarkeit der Videos, regen jedoch weitere Studien zu dieser Thematik an.

Methode

Das Korpus besteht aus 40 YouTube-Videos (insges. 385 Minuten), in denen Wissen zu schulisch und/oder universitär relevanten Themen vermittelt wird, z. B. als Anleitung zum Verfassen einer Gedichtanalyse oder Erklärung der Bilanz in T-Konten. Die Videos wurden so ausgewählt, dass verschiedene Fachgebiete abgedeckt werden. Zudem ist das Korpus zweigeteilt in jeweils ein erfolgreiches und ein weniger erfolgreiches Video pro Thema. Wobei diese Qualitätszuschreibung zunächst an rein formalen Kriterien wie Klickzahlen, Anzahl und Art der Kommentare sowie „Daumen hoch bzw. runter“-Bewertungen ausgemacht wurde [15].

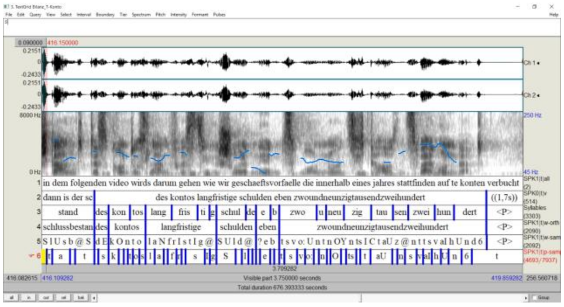

in dem folgenden video wirds darum gehen wie wir geschaeftsvorfaelle die innerhalb eines jahres stattfinden auf te konten verbucht
dann is der sc
des kontos langfristige schulden eben zwoundneunzigtausendzweihundert
((1,7s))
stand des kon tos lang fris ti g schul de e b zwo u neu zig tau sen zwei hun dert <P>
schlussbestan des kontos langfristige schulden eben zwoundneunzigtausendzweihundert <P>
S l U s b @ S d E k O n t o l a N f r I s t I g @ S U l d @ ? e b t s v o: U n t n OY n ts I C t aU z @ n t t s v a I h U n d 6 <P>
Ch 1
Ch 2
8000 Hz
0 Hz
250 Hz
45 Hz
SPK1|t|all (2)
SPK0|t|v (514)
Syllables (3303)
SPK1|t|w-orth (2090)
SPK1|t|w-sam (2092)
SPK1|t|p-sam (4693/7937)
Visible part 3.750000 seconds
Total duration 676.393333 seconds

Abb. 1: Annotation der Audiospur eines YouTube-Videos in Praat.

Die Videos wurden in Praat [16] per Skript [17] auf Silbenebene segmentiert. Für weitere Analysen wurde zudem mittels WebMAUS Basic [18] eine Segmentierung in Wörter und Laute erstellt (vgl. Abb. 1). Da die beiden automatisierten Verfahren teilweise große Ungenauigkeiten aufwiesen, z. B. durch Musikbett im Video, wurden die Segmentgrenzen manuell korrigiert. Mittels Skript [19] wurde schließlich das Sprechtempo berechnet. Folgende Werte wurden dafür bestimmt:

- Die globale Nettosprechgeschwindigkeit, d. h., die Zahl der tatsächlich gesprochenen Silben pro Sekunde inklusive Pausen [20].
- Die globale Nettoartikulationsgeschwindigkeit, d. h., die Zahl der tatsächlich gesprochenen Silben pro Sekunde exklusive Pausen.
- Das Verhältnis von Gesamtsilben- zu Gesamtpausenzahl.
- Der Anteil der Pausen an der Gesamtdauer.

Wenn im Folgenden von Sprechtempo die Rede ist, dann werden damit übergreifend die zuvor genannten vier Messungen gemeint. Die Unterscheidung zwischen Sprech- und Artikulationsgeschwindigkeit ist im Hinblick auf Videos besonders relevant, da in Videos damit zu rechnen ist, dass (längere) Pausen auftreten, in denen zwar nicht gesprochen wird, die Zuschauer*innen aber visuelle Inhalte präsentiert bekommen und die Pause anders beurteilen als in einem rein auditiven Kontext.

Ergebnisdarstellung und Diskussion

Aus den Tabellen 1 und 2 lassen sich die drei Extremwerte für jede Messung, unterteilt in erfolgreiche und weniger erfolgreiche Videos, ablesen. Die statistische Auswertung verdeutlichte, dass es für keine der vier Messungen signifikante Korrelationen zwischen Sprechtempo und Erfolg eines Videos gibt. Sowohl bei den erfolgreichen als auch bei den weniger erfolgreichen Videos werden ähnlich hohe Werte der globalen Nettoartikulationsgeschwindigkeit um 6 Silben/Sekunde und vergleichbare niedrige Werte um 3,8 Silben/Sekunde erreicht. Die globale Nettosprechgeschwindigkeit ist aufgrund der mitberechneten Pausen zwar insgesamt niedriger als die Artikulationsgeschwindigkeit, aber auch hierfür gibt es höhere sowie niedrigere Geschwindigkeiten bei beiden Videogruppen.

Extremwerte						
ERFOLG			ja		nein	
			Fallnr.	Wert	Fallnr.	Wert
SPRECH-GESCHWINDIG-KEIT	Höchster	1	10	5,19	21	5,09
		2	5	5,17	22	4,50
		3	9	5,06	35	4,23
	Niedrigster	1	2	1,93	26	1,85
		2	19	2,74	34	1,90
		3	8	2,97	31	2,27
ARTIKULATIONS-GE-SCHWINDIGKEIT	Höchster	1	10	5,95	26	6,06
		2	9	5,88	30	5,66
		3	5	5,87	21	5,63
	Niedrigster	1	2	3,90	34	3,74
		2	19	4,40	29	4,28
		3	15	4,75	31	4,28

Tab. 1: Extremwerte der Messungen zur Sprech- und Artikulationsgeschwindigkeit.

Das Verhältnis von Silben zu Pausen zeigt in den Extremwerten durchaus Unterschiede. So kommt bei den erfolgreichen Videos in Fallnr. 9 durchschnittlich auf 21,42 Silben eine Pause, das niedrigste Extrem (Fallnr. 19) kommt auf

5,56 Silben/Pause. Der höchste Wert bei den weniger erfolgreichen Videos (Fallnr. 21) liegt nur bei 16,42 Silben/Pause, im Video mit der Fallnr. 31 hingegen wird durchschnittlich alle 3,72 Silben eine Pause realisiert – die weniger erfolgreichen Videos weisen also tendenziell mehr Pausen auf. Dennoch ergibt sich für die gesamten 40 analysierten Videos keine signifikante Korrelation.

Extremwerte						
ERFOLG			ja		nein	
			Fallnr.	Wert	Fallnr.	Wert
SILBEN/PAUSEN	Höchster	1	9	21,42	21	16,42
		2	10	18,97	38	14,31
		3	3	15,06	26	12,85
	Niedrigster	1	19	5,56	31	3,72
		2	2	5,61	29	4,45
		3	18	6,95	34	5,21
PAUSEN/ GESAMT-DAUER	Höchster	1	2	50,4%	26	69,4%
		2	3	42,3%	34	49,1%
		3	12	38,2%	31	46,8%
	Niedrigster	1	5	12,0%	21	9,6%
		2	6	12,4%	35	11,5%
		3	10	12,8%	22	19,8%

Tab. 2: Extremwerte der Messungen zum Verhältnis von Silben zu Pausen sowie zum Anteil der Pausen an der Gesamtdauer des Videos.

Im Hinblick auf den Anteil der Pausen an der Gesamtdauer des Videos ließ sich vermuten, dass dieser mit dem Verhältnis von Silben zu Pausen zusammenhängen könnte: Je mehr Pausen realisiert werden, desto größer ist die Gesamtpausendauer. Allerdings können häufige, aber dafür kurze Pausen einen geringeren Anteil an der Gesamtdauer ausmachen als einige wenige lange Pausen. Beides findet sich in den Messwerten. Während Fallnr. 3 mit 15,06 Silben/Pause eines der Videos mit insgesamt wenigen realisierten Pausen ist, erreicht es mit einem Anteil von 42,3% einen verhältnismäßig hohen Pausenanteil an der Gesamtdauer. Dies spricht dafür, dass zumindest einige Pausen in diesem Video relativ lang sind. Das Video mit der Fallnr. 21 hingegen weist zwar ebenfalls wenige Pausen auf, diese haben jedoch nur einen geringen Anteil an der Gesamtdauer, sind also eher kurz. Im Vergleich der erfolgreichen mit den weniger erfolgreichen Videos zeigt sich mit Bezug auf den Anteil der Pausen an der Gesamtdauer, dass in den erfolgreichen Videos der Anteil tendenziell nicht ganz so hoch ist, wie in den weniger erfolgreichen, wo zumindest in einem Video (Fallnr. 26) mehr als 2/3 des Videos nicht artikuliert wird. Dennoch besteht auch für diese Werte kein signifikanter Zusammenhang.

Während der Arbeit an den Videos zeigten sich weitere Auffälligkeiten, die im Zusammenhang mit dem Sprechtempo stehen könnten und in der Folge untersucht werden sollten:

- Das Sprechtempo schwankt in manchen Videos stark, so dass neben der globalen auch die lokale Sprechgeschwindigkeit betrachtet werden sollte.
- Die Deutlichkeit der Aussprache variiert zwischen den YouTuber*innen teilweise enorm. So gibt es einige, die trotz eines hohen Sprechtempos wenige segmentale Auffälligkeiten aufweisen, während andere in hohem Maße assimilieren. Hier ist also das Verhältnis von Textbasis zu tatsächlicher Realisation zu ermitteln.
- Das wahrgenommene Sprechtempo muss nicht mit den gemessenen Werten übereinstimmen. Einen Hinweis darauf liefern Weirich & Simpson [21], die zeigten, dass die Größe des genutzten Vokalraums eines Sprechers bzw. einer Sprecherin einen Einfluss auf das wahrgenommene Sprechtempo hat. Es gilt also den Vokalraum der YouTuber*innen zu ermitteln, um diesem Zusammenhang nachgehen zu können.
- Jenseits der sprecherischen Parameter hat wahrscheinlich auch die Geschwindigkeit, mit der die Inhalte präsentiert werden, und welche sprachlichen Mittel dafür genutzt werden, einen Einfluss auf das wahrgenommene Tempo eines Lehrvideos. Für aussagekräftige Analysen zu diesem Punkt ist eine qualitative Inhaltanalyse notwendig.

Fazit

YouTube-Videos stellen einen interessanten und relevanten, jedoch auch sehr komplexen Analysegegenstand dar. Im Forschungsprojekt wird geprüft, welche sprachlichen und sprecher*innenspezifischen Merkmale unter Einbezug der filmischen Ebene sowie im Hinblick auf die Spezifika von Social Media einen Einfluss auf den Erfolg von Lehrvideos haben können.

Die Berechnung des Sprechtempos auf Silbenebene führte zu keinen signifikanten Ergebnissen. Jedoch zeigte sich, dass weitere Einflussgrößen wie das wahrgenommene Sprechtempo und die Deutlichkeit der Artikulation zusätzlich berücksichtigt werden sollten.

Referenzen

[1] Hollmer, K., “Wissenschaft – ein Hit auf Youtube,” *Süddeutsche.de*, 2016 (http://www.sueddeutsche.de/ digital/2.220/science-channels-warum-wissenschaft-auf-youtube-so-gut-ankommt-1.3083267) [23.01.2019].

[2] Hart, J., *Top Tools for Learning 2018,* 2018 (http://www.toptools4learning.com/youtube/) [23.01.2019].

[3] Breunig, C. & B. Engel, “Massenkommunikation 2015: Funktionen und Images der Medien im Vergleich,” *Media Perspektiven*, 7–8, pp. 323–341, 2015.

[4] Breunig, C. & B. van Eimeren, “50 Jahre ‚Massenkommunikation‘: Trends in der Nutzung und Bewertung der Medien,” *Media Perspektiven*, 11, pp. 505–525, 2015.

[5] Feierabend, S., Plankenhorn, T. & T. Rathgeb, “Jugend, Information, Multimedia. Ergebnisse der JIM-Studie 2016,” *Media Perspektiven*, 12, pp. 586–597, 2016.

[6] Feierabend, S., Plankenhorn, T. & T. Rathgeb, “Kindheit, Internet und Medien. Ergebnisse der KIM-Studie 2016,” *Media Perspektiven*, 4, pp. 206–215, 2017.

[7] Meinhard, D. B., Clames, U, & T. Koch, “Zwischen Trend und Didaktik – Videos in der Hochschullehre,” *ZFHE*, 9(3), pp. 50–64, 2014. DOI: 10.3217/zfhe-9-03/07.

[8] Körkel, T. & K. Hoppenhaus, “Über Web, Video und Wissenschaft,” in *Web Video Wissenschaft. play science. Ohne Bewegtbild läuft nichts mehr im Netz: Wie Wissenschaftsvideos das Publikum erobern,* Körkel T. & K. Hoppenhaus (Eds.), Heidelberg: Spektrum der Wissenschaft, o. S., 2016.

[9] Welbourne, D. J. & W. J. Grant “Science communication on YouTube: Factors that affect channel and video popularity,” *Public understanding of science,* 25(6), pp. 706–718, 2016. DOI: 10.1177/096366251 5572068.

[10] Allgaier, J. “Wo Wissenschaft auf Popularkultur trifft,” in *Web Video Wissenschaft. play science. Ohne Bewegtbild läuft nichts mehr im Netz: Wie Wissenschaftsvideos das Publikum erobern,* Körkel, T. & K. Hoppenhaus (Eds.), Heidelberg: Spektrum der Wissenschaft, o. S., 2016.

[11] Geipel, A., “Wissenschaft@YouTube. Plattformspezifische Formen von Wissenschaftskommunikation,” in *Knowledge in Action. Neue Formen der Kommunikation in der Wissensgesellschaft,* E. Lettkemann, Wilke R. & H. Knoblauch (Eds.), Wiesbaden: Springer Fachmedien, pp. 137–163, 2018.

[12] Becher, A., *Lernvideos auf YouTube*, Masterarbeit, Technische Universität Dresden, 2012.

[13] Merkt, M., *Didaktische Optimierung von Videos in der Hochschullehre*, 2015. Abrufbar unter: https://www.e-teaching.org/ news/eteaching_blog/didaktische-optimierung-von-videos-in-der-hochschullehre [23.01.2019].

[15] Bachl, M., “Erfolgsfaktoren politischer Y-ouTube-Videos,” in *Das Internet im Wahlkampf*, Schweitzer, E.J. & S. Albrecht (Eds.), Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 157–180, 2011.

[16] Boersma, P. & D. Weenink, *Praat: doing phonetics by computer* [Computer program]. Version 6.0.42, retrieved 15 August 2018 from http://www.praat.org/ [23.01.2019].

[17] Mayer, J., *Praat script syllable_candidates*, abrufbar unter: http://praatpfanne.ling phon.net/downloads/syllable_candidates.txt [23.01.2019].

[18] Kisler, T., Reichel U. D. & F. Schiel, “Multilingual processing of speech via web services,” *Computer Speech & Language*, 45, pp. 326–347, 2017.

[19] Mayer, J., *Praat script 5.3.2 (Sprechgeschwindigkeit),* abrufbar unter: http://praat pfanne.lingphon.net/downloads/hb-script532.txt [23.01.2019].

[20] Pfitzinger, Hr. R., *Phonetische Analyse der Sprechgeschwindigkeit,* pp. 139–140, 2001 (https://www.phone tik.uni-muenchen.de/forschung/FIPKM/ vol38/f38_hp_1.pdf) [23.01.2019].

[21] Weirich, M. & A. P. Simpson, “Differences in acoustic vowel space and the perception of speech tempo,” in *Journal of Phonetics*, 43, pp. 1–10, 2014. DOI: 10.1016/j.wocn.2014.01.001.

Sociolinguistic influences on female fundamental frequency in English

Kerstin Endes[1]

[1]Karl-Franzens Universität Graz

kerstin.endes@edu.uni-graz.at

Abstract

Previous studies on fundamental frequency (F_0) in conversation have found that a speaker's F_0 changes depending on the sex of their conversation partner. For example, Benus et al. [1] investigated the perception of social behavior in dialogue, and a link between emotions and prosodic and acoustic cues in speakers has been found in several studies [2, 3]. Benus et al. [1] reported significant differences between speakers depending on sex. Female speakers raised their F_0 when attempting to be liked by their male interlocutor, while F_0 was significantly lowered in conversation with other females, with an increase in intensity. We studied F_0 of nine female native speakers of English, measured in Hz, in two conditions: in a conversation with a male and a female interlocutor respectively, in order to observe potential changes in F_0. Counter to expectations, female F_0 range was not significantly broader when communicating with a male speaker as compared to a female speaker.

Introduction

A link between emotions and prosodic and acoustic cues in speakers has been found in several studies [2–3]. Benus et al. [1] analyzed the influence of gender in interaction in different settings: male-male, female-female and male-female. Their results showed that females attempting to be looked upon favorably by their interlocutor raised F_0 when talking to a male and lowered it when conversing with a female. Intensity and the rate of speech also increased in the second condition. Interestingly, speakers who attempted to impress their interlocutor were more likely to resemble them acoustically, regardless of their sex. Based on the findings of Benus et al. [1], this paper investigates whether female F_0 depends on the sex of their interlocutor.

However, numerous sociophonetic studies stress the impact of factors other than the speaker's sex on F_0, e.g. social status of speakers, age and language proficiency. Therefore, the theory of accommodation in communication was applied in the analysis of the results [5].

Differences in F_0 depending on the sex of the speaker might also be due to cultural norms. Van Bezooijen [6], for example, found differences in the F_0 range of Dutch and Japanese women; the former were seen to have a F_0 range similar to their male peers, given that medium to low pitch is considered vocally attractive in Dutch women (around 185 Hz), whereas Japanese women showed a much higher range than their male counterparts, rendering pitch range a marker for gender in Japanese culture. Van Bezooijen [6] also found strong differences in Japanese culture between the prototypical man and woman; a contrast which is much more pronounced than in Dutch culture.

Method

Nine female native speakers of English, three males (one native and two non-natives), and one non-native female took part in the study. The subjects were recruited in Graz, Austria. The subjects' age ranged from 19 to 48 years, with a mean of 22.8 years.

A map task was used in the study, as this task requires the subjects to interact with each other [7]. Therefore, two contrasting maps depicted the target words used for F_0 analysis. The speakers were asked to use a set of questions in their conversation, in which the target words occurred in sentence-final position. Speakers were asked to perform the task with a male and then a female speaker, and to use the set of questions ensuring that the target words were equally frequent in both conditions. The target words were chosen with the intention of keeping voiceless plosives and fricatives to a minimum, in order to facilitate the process of segmenting the audio files and to maximize the amount of voicing in the signal which could be used to measure F_0. The set of questions was embedded in semi-spontaneous speech; the speakers were also asked to inquire about their interlocutors' map, given that the two maps differed in their design. The underlying hypothesis was that female F_0 would be significantly higher when talking to a male than to a female, irrespective of the female's sexual orientation [8, 1].

All recordings took place in a sound attenuated booth using version 2.2.1 of Audacity(R) [9] recording and editing software with a sampling rate of 44.1 KHz and a Microsoft headset (LX 3000). Each participant asked a total of 14 questions in two conditions: in conversation with a male and then a female speaker, respectively. The female native speaker who was being recorded sat in the sound attenuated booth, while their interlocutor was located outside the booth but remained visible through a window in the booth. After the conversations, each participant filled out a questionnaire which included ratings of vocal attractiveness, perception of dominance and the impression which the subject intended to make on her interlocutor on a scale ranging from 1, very applicable, e.g. interlocutor was perceived as very dominant, to 5, not at all applicable.

The F_0 range was obtained from Praat [10], using the program's standard settings, i.e. with the pitch floor at 75 Hz and the pitch ceiling at 600 Hz. The target words were extracted manually with the Praat select function. The pitch contours were not corrected, even though creaky voice might pose a problem when extracting F_0 contours [11].

Results

A one-way ANOVA was conducted to test whether female F_0 differed depending on the sex of their interlocutor (male vs. female). There was no significant effect of the interlocutor's sex on F_0 of the female native subjects [F (1,232)=2.131, p=0.146]. However, a small trend was visible that F_0 was, on average, slightly higher in conversation with males than with females. For the results, Fig. 1 will serve as a reference point.

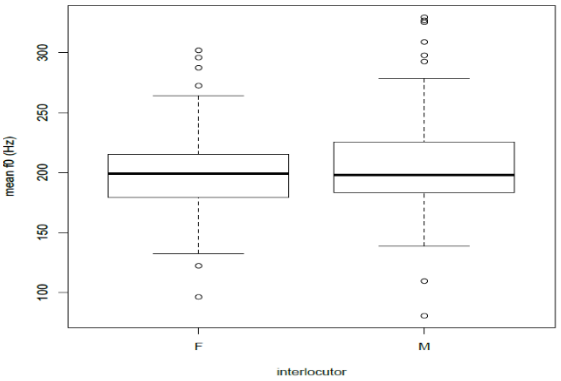


Fig. 1: Mean F_0 for nine female native speakers of English in two conditions (male vs. female).

Discussion

Contrary to our expectations, no significant differences could be found in the measures depending on the condition, i.e. female native speaker talking to a male and female respectively. These results are in accordance with accommodation theory [12], as women and men have been reported to accommodate to a greater extent to male speakers than to females. In addition, social cues may play a decisive role in accommodation and add to gender differences [12].

Only one male participant was a native speaker of English, while the female and one male interlocutor were EFL-learners with German as their L1, and one male had little exposure to the target language. Therefore, it is possible that any effects of interlocutor sex on F_0 were overruled by effects of foreigner-directed speech [2] The results of this study add to the scant number of studies on foreigner-directed speech (FDS), as well as to the numerous studies on dominance perception and pitch. It is also in line with the results of various studies which have tested the impact dialectal variation can have on F_0 range [13, 14]. Ladd et al. [14], for example, found differences in the realization of F_0 peaks in Standard Scottish and RP English.

The results of the present study show that F_0 does not depend so much on the interlocutor's sex, but rather on their linguistic needs, meaning that FDS facilitates communication [15–17].

Research has found that speech is adapted to the level of the target audience, thus infant directed speech (IDS) is characterized by higher pitch (F_0, F_0 contours), intensified emotional involvement and hyperarticulated vowels. These findings have been shown to be valid across languages, cultures and genders. Characteristics of IDS are thought to be employed subconsciously; exaggerated features of speech in IDS are thought to aim at keeping an infant's attention. It is suggested that FDS is similar to IDS in that it features a didactic element. However, the former is characterized by vowel hyperarticulation to a lesser extent than the latter [18].

Linguistic accommodation describes the options available to a speaker to render speech more (convergence) or less (divergence) analogous to that of the interlocutor. Given that native speakers are naturally characterized by a higher level of language competence, they have "linguistic status" [19], which can be compared to the higher status of more advanced EFL/ESL learners as compared to speakers whose language use is less proficient.

In this study the language competence of the female and male interlocutors differed greatly; one male was a native speaker of English, another male and one female were students of English, while one male had little exposure to the target language, thus facing difficulties (in some cases) when answering the questions of the native subjects. The subjects had to repeat several phrases to facilitate understanding for this particular speaker.

In addition to 'convergence' and 'divergence', accommodation has two directions, namely 'upward' and 'downward'. While upward convergence describes the attempt to master a foreign language by the learner, downward convergence explains the process of the native speaker adopting his/her speech to the language competence of the interlocutor, who is characteristically a non-native speaker of the language [19]. The type of accommodation strategy in this study is that of downward convergence.

Conclusion

The experiment outlined in this paper adds to the numerous studies on accommodation theory, illustrating the potential effect language competence can have on the interlocutor's fundamental frequency. In a follow-up experiment, it would be advisable to include a more comprehensive questionnaire to facilitate the interpretation of results. Information on the following aspects would have been interesting: the perception of foreignness of the interlocutor (if necessary, so as to gauge positive/negative effects in the accommodation process), the degree of simplification of language to facilitate communication, and the effort involved in communicating.

References

[1] Benus, S. et al., "Acoustic and Prosodic Correlates of Social Behavior," *Proceedings of the conference "Interspeech"*, 2011.

[2] Biersack, S. et al*., "*Fine-Tuning Speech Register: A comparison of the prosodic features of child-directed and foreigner-directed speech," *Proceedings of the "9th European Conference on Speech Communication and Technology"*, 2007.

[3] Knoll, M. & L. Scharrer, "Acoustic and affective comparison of natural and imaginary infant-, foreigner- and adult-directed speech," *Proceedings of the "9th European Conference on Speech Communication and Technology"*, 2007.

[4] Knoll, M. et al., *Emotional, linguistic or just cute? The function of pitch contours in infant- and foreigner-directed speech.* Proceedings of the conference "Speech Prosody", 2006.

[5] Shepard, C. et al., "Communication accommodation theory," in *The new handbook of language and social psychology,* Robinson, W. P. & H. Giles (Eds.), Chichester: Wiley, pp. 33–56, 2001.

[6] Van Bezooijen, R., "Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women," *Language and Speech*, 38(3), pp. 253–265, 1995.

[7] Colantoni, L. et al., *Second Language Speech. Theory and Practice,* Cambridge: CUP, 2015.

[8] Waksler, R., "Pitch range and women's sexual orientation," *WORD*, 52(1), pp. 69–77, 2001.

[9] Audacity Team, Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 2.2.1 retrieved August 14th 2017 from https://audacityteam.org/ [04.02.2019].

[10] P. Boersma & D. Weenink, *Praat: doing phonetics by computer* [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/ [04.02.2019].

[11] Mennen, I. et al., "Cross-Language differences in fundamental frequency range: A comparison of English and German," *Acoustical Society of America,* 131(3), pp. 2249–2260, 2012.

[12] Namy, L. et al., "Gender differences in vocal accommodation: The role of perception," *Journal of language and social psychology,* 21(4), pp. 422–432, 2002.

[13] Grabe, E. et al., "Pitch accent realization in four varieties of British English," *Journal of Phonetics,* 28, pp. 161–185, 2000.

[14] Ladd, R. D. et al., "Structural and dialectal effects on pitch peak alignment in two varieties of British English," *Journal of Phonetics,* 37, pp. 145–161, 2009.

[15] Dillard, J. & K. Tusing, "The Sounds of Dominance: Vocal Precursors of Perceived Dominance During Interpersonal Influence," *Human Communication Research,* 26(1), pp. 148–171, 2000.

[16] Farrell, T. et al., "Dominance signaled in an acoustic ornament," *Animal Behaviour,* 79, pp. 657–664, 2010.

[17] Charfuelan, M. et al., "Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings," *Proceedings of the conference "Interspeech"*, 2010.

[18] Burnham, D. et al. "Do you speak E-N-G-L-I-S-H? A comparison of foreigner-and infant directed speech," *Speech Communication,* 49, pp. 2–7, 2007.

[19] James, C., "Accommodation in Crosslanguage Encounters," *Papers and Studies in Contrastive Linguistics* 27, pp. 39–48, 1993.

Lieben oder leben? Diskrimination des deutschen /i:/-/e:/-Vokalkontrasts durch Sprecher*innen mit L1 Japanisch[1]

David Eppensteiner[1]

Universität Wien[1]

david.eppensteiner@outlook.com

Abstract

*In diesem Beitrag wurde die Unterscheidung des deutschen /i:/–/e:/-Vokalkontrasts durch japanische Studierende getestet und die Vorhersagen des Perceptual Assimilation Models von Best und Tyler und des Speech Learning Models von Flege überprüft. Die Hypothesen beider Modelle konnten durch das AXB-Diskriminationsexperiment bestätigt und der Einfluss des Faktors Spracherfahrung für diesen Kontrast und diese Gruppe aufgezeigt werden. Der Effekt des Speech Shaped Noise mit 0 dB und -5dB SNR auf die Hörer*innengruppe konnte nicht eindeutig bestimmt werden.*

Einleitung

Die Wahrnehmung neuer Phonemkontraste stellt mitunter ein Problem für L2-Lernende dar [1]. Ziel dieser Untersuchung ist es, herauszufinden, ob und wie gut erwachsene japanische DaF-Lernende den /i:/-/e:/-Vokalkontrast unterscheiden können, welcher im Japanischen nicht kontrastiv ist. Dazu wurden zwei Gruppen japanischer Studierender getestet, die sich hinsichtlich ihrer Deutschkenntnisse bzw. der Kontaktdauer mit der L2 Deutsch unterscheiden. Durch den Vergleich beider Gruppen soll die Relevanz des nur schwer quantifizierbaren Faktors Spracherfahrung für diesen Kontrast bestimmt werden. Die Vorhersagen zweier phonetischer Modelle – des Perceptual Assimilation Models (PAM bzw. die Weiterentwicklung PAM-L2) von Best und Tyler [2] und des Speech Learning Models (SLM) von Flege [3] – sollen dabei überprüft werden.

Methode

Die Diskrimination des deutschen /i:/-/e:/-Vokalkontrasts durch Sprecher*innen mit L1 Japanisch wurde in einem AXB-Diskriminationsexperiment untersucht. Dazu wurden zwei Gruppen japanischer Studierender zwischen 18 und 22 Jahren der Dokkyo Universität in Saitama getestet. Gruppe A besteht aus 10 Studierenden ohne Deutschkenntnisse, Gruppe BC aus 17 japanischen Deutschstudent*innen, die seit mindestens 2 Jahren Deutsch gelernt haben und etwa auf Niveau B1 des Gemeinsamen Europäischen Referenzrahmens für Sprachen sind. Die sprachliche Kompetenz der Gruppe BC in der L2 Deutsch spiegelt sich indirekt in den universitären „Niveauklassen“ wieder, worin die Lernenden eingeteilt worden sind.

Als Stimuli fungieren 24 Minimalpaare (gemischt mit anderen irrelevanten Vokalkontrasten), die bis auf den betreffenden Vokal identisch sind (z.B. *lieben – leben*). Um die Gleichheit der Silbenränder zu gewährleisten, wurde der „*i*-Vokal“ in das „*e*-Wort“ hineingeschnitten (sog. splicing). Zur Untersuchung der Fragen wurde ein MFC-Wahrnehmungsexperiment in Praat [4] konzipiert. Außerdem wurden insgesamt drei Schwierigkeitsstufen (mit und ohne Störgeräusch) eingefügt, um die Bedingungen des Hörens zu erschweren und möglicherweise größere Unterschiede herausarbeiten zu können. Als Störgeräusch wurde ein Speech Shaped Noise (SSN) in unterschiedlicher Stärke (0dB SNR und -5dB SNR) verwendet, das mithilfe eines Praatskripts aus einer vielstimmigen Aufnahme eines Stimmengewirrs in einer Bar erzeugt wurde.

Hypothesen

Aufgrund der inter- und intralingualen Ähnlichkeit des Vokalkontrasts und der Ergebnisse vorangegangener Studien [1] wurde von einem Category-Goodness Difference (CG) ausgegangen, wonach Hörer*innen mit L1 Japanisch voraussichtlich leichte Schwierigkeiten haben, die beiden Vokale zu diskriminieren. Die Erkennungsrate für die Gruppe der japanischen Studierenden ohne Spracherfahrung in der L2 Deutsch ist voraussichtlich gut, aber nicht ausgezeichnet (< 90%, gemäß CG des PAM, → H1).

[1] Dieser Beitrag basiert auf einer Masterarbeit zum selben Thema.

Dem PAM-L2 zufolge ist die Diskrimination der Gruppe der japanischen Studierenden mit Spracherfahrung voraussichtlich signifikant besser als jene ohne sprachlicher Erfahrung in der L2 Deutsch und voraussichtlich annähernd „native-like“ [2] (→ H2). Die Diskrimination ist vermutlich anfangs aufgrund der Äquivalenzklassifizierung schlechter, mit zunehmender sprachlicher Erfahrung sollte die phonetische erstsprachliche Kategorie für den unähnlicheren Laut aber soweit modifiziert werden, dass er problemlos oder zumindest besser differenziert werden kann.
Weiters wurde davon ausgegangen, dass beide Gruppen im Vergleich zur Kontrollgruppe unter erschwerten Bedingungen signifikant schlechter abschneiden, während die Unterscheidungsrate der Kontrollgruppe annähernd gleichbleibt (→ H3).

Ergebnisse und Diskussion

Die ersten beiden Hypothesen konnten für alle drei Experimentteile bestätigt werden (siehe Abb. 1–3). Die Gruppe der japanischen Studierenden ohne sprachlicher Erfahrung in der L2 Deutsch erreichte im ersten Teil des Experiments durchschnittlich 80.21%, die hohe Standardabweichung lässt allerdings auf eine heterogene Verteilung der individuellen Diskriminationsraten schließen. Die Unterschiede zwischen Gruppe A und Gruppe BC sind signifikant (p<0.05).

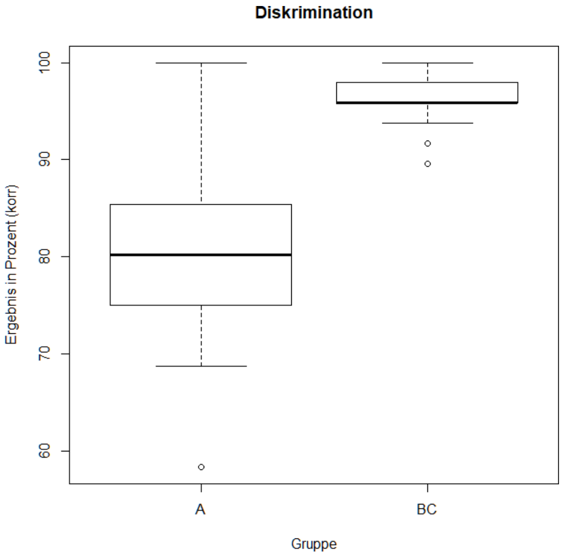


Abb. 1: Diskriminationsrate für Teil 1 in Prozent korrekt (ohne SSN).

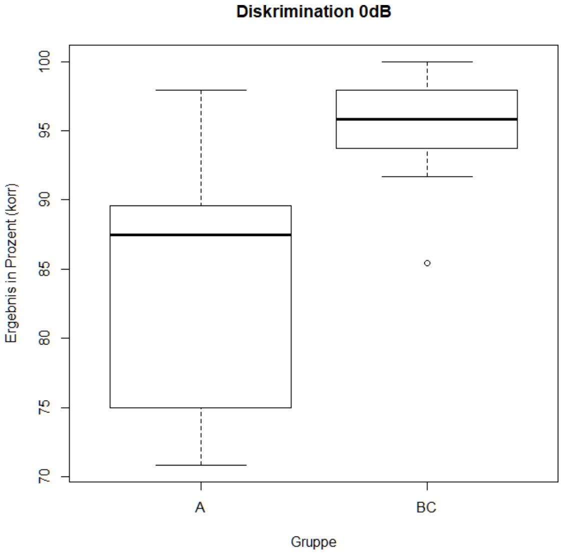


Abb. 2: Diskriminationsrate für Teil 2 in Prozent korrekt (SNR 0dB).

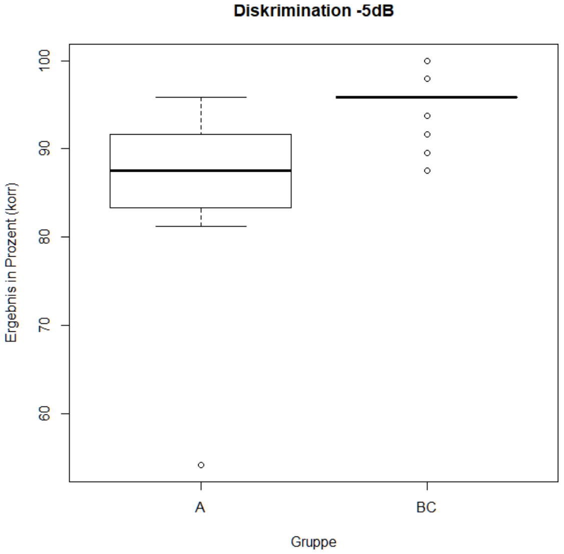


Abb. 3: Diskriminationsrate für Teil 3 in Prozent korrekt (SNR -5dB).

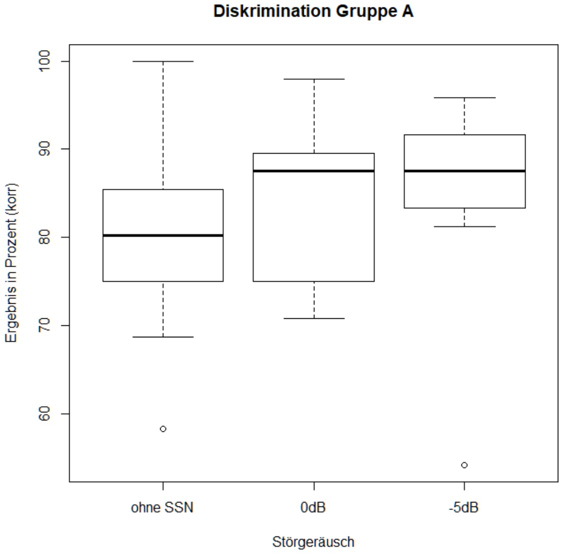


Abb. 4: Vergleich der drei Experimentteile für Gruppe A.

Die Ergebnisse der „erfahrenen“ Gruppe sind im Vergleich zur Kontrollgruppe mit Erstsprache Deutsch nur geringfügig schlechter. Die Proband*innen hatten in Übereinstimmung mit den Vorhersagen der phonetischen Modelle keinerlei Probleme die beiden Vokale zu unter- scheiden. Auch die Resultate der Gruppe A sind mit durchschnittlich >80% in allen drei Teilen relativ gut, wobei für diese Gruppe leichtere Schwierigkeiten erkennbar waren.

Das AXB-Diskriminationsexperiment hat gezeigt, dass zwischen den Ergebnissen beider Gruppen japanischer L1-Hörer*innen ein signifikanter Unterschied besteht – und zwar für alle drei Teile bzw. Störstärken. Somit kann für die Diskrimination des deutschen /iː/-/eː/-Vokalkontrasts durch Hörer*innen mit L1 Japanisch auf einen entscheidenden Einfluss sprachlicher Erfahrung geschlossen werden – in dieser Arbeit repräsentiert durch mehrjährigen Kontakt mit der L2 bzw. durch Unterricht in der Fremdsprache Deutsch im universitären Kontext in Japan. Der Faktor Spracherfahrung spielt also für diesen Kontrast definitiv eine Rolle.

Allerdings wurde das Ergebnis mit steigender Störstärke nicht wie erwartet schlechter, sondern mitunter sogar besser (siehe Abb. 4 für Gruppe A). Es hat durch die Wiederholung der Stimuli und der Einstellung auf die Hörsituation eventuell ein Lerneffekt stattgefunden, weshalb die dritte Hypothese nicht bestätigt werden kann.

Referenzen

[1] Mazuka, R., Hasegawa, M. & S. Tsuji, “Development of Non-Native Vowel Discrimination: Improvement Without Exposure,” *Developmental Psychobiology*, 56(2), pp. 192–209, 2014.

[2] Best, C. T. & M. D. Tyler, “Nonnative and second-language speech perception: Commonalities and complementarities,” in *Second language speech learning: The role of language experience in speech perception and production*, Munro, M. J. & O.-S. Bohn (Eds.), Amsterdam: John Benjamins, pp. 13–34, 2007.

[3] Flege, J. E., “Language contact in bilingualism: Phonetic system interactions,” in *Laboratory Phonology*, 9, Cole, J. & J. I. Hualde (Eds.), Berlin/New York: de Gruyter, pp. 353–381, 2007.

[4] P. Boersma & D. Weenink, *Praat: doing phonetics by computer* [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/, http://www.fon.hum.uva.nl/praat/download_win.html [20.10.2019].

Sociolinguistic implications on the variation of MHG /ei/ in Austria

Johanna Fanta-Jende[1]

[1]University of Vienna

johanna.fanta@univie.ac.at

Abstract

This study focuses on the distribution of MHG /ei/ among speakers of three villages representing Austria's major dialect regions within Bavarian: Central Bavarian, the South-Central Bavarian transition zone and South Bavarian. Different situational settings covering more controlled language data as well as 'free' conversational data give insights into inter- and intra-individual language behavior patterns on the dialect-standard-axis. Thus, this study captures not only the areal-horizontal dimension of language variation in rural Austria but also social-vertical aspects. The results demonstrate a decline of the traditional oa-diphthongs for the Eastern Bavarian language area, with a-monophthongs becoming more prominent on the speakers' individual dialect-standard-axis.

Introduction

The variation of MHG /ei/ is a well-documented phenomenon for different regions of the entire German speaking area [1: 284–298]. For the Bavarian dialects within Austria, traditionally, the realization as diphthong /ɔɐ̯/ is assumed, resulting for instance in /hɔɐ̯s/ for NHG *heiß* 'hot', /g̊lɔɐ̯n/ for NHG *klein* 'small', or /brɔɐ̯d/] for NHG *breit* 'broad' [2–4].

However, there are exceptions to this basic rule concerning the use of /aː/ instead of /ɔɐ/ (e.g. /haːs/ for NHG *heiß* 'hot'), especially for the areas of Vienna and southern Carinthia. The Wiesinger map (WEK, "heim") for NHG *heim* 'home' based on the so-called *Wenker sentences* from the late 1920s demonstrates this traditional distribution (see Fig. 1) [5]. For historical sociolinguistic and phonological reasons [6: 60–61], the *a*-monophthong replaced the former diphthong in these areas and is said to be spreading gradually to other parts of Austria since then [7: 29, 6: 63, 4].

As early as the 1910s, Pfalz [8] identified a shift in the phonological system of Deutsch-Wagram, a small town bordering Vienna. While the use of /aː/ at the beginning of the 20th century was still restricted to the local 'school youth' („Schuljugend") and 'intentionally used minor variants' („absichtlich gebrauchte nebenformen [sic!]"; translated by JFJ) [8], Unger [9] found Pfalz' observations confirmed when she looked at the same location a hundred years later. All nonstandard diphthong-variants now seemed to have been replaced by the *a*-monophthong except for low-frequent agricultural terms (e.g. /hɔɐ̯n/ for *Heiden/Buchweizen* 'buckwheat') [9: 183]. Moving (about 190 km) away from Vienna, in the 1980s for the town of Ulrichsberg in Central Bavarian Upper Austria, Scheutz found the decline of a binary system between traditional /oa/ and standard German /ai/ in favor of a three-way-paradigm including /aː/ as an intermediate vernacular form, which presumably appears 'less dialectal' to the speakers of Ulrichsberg [10]. By comparing data from interviews and informal conversations, Scheutz also identified situational effects resulting in higher numbers of *a*-monophthongs and standard-near *ai*-diphthongs in an interview setting than in an informal conversation [10].

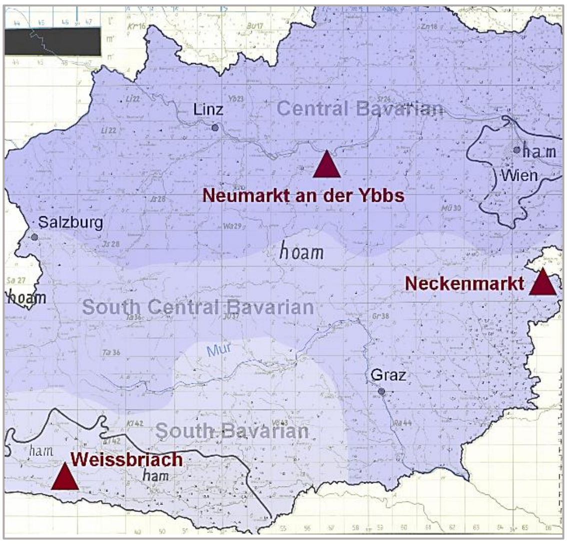


Fig. 1: Distribution of MHG /ei/ based on WEK "heim" considering different subregions of Bavarian in Austria and the analyzed locations. Generated with www.regionalsprache.de.

As these findings indicate, MHG /ei/ appears as a highly dynamic phenomenon with a broad variational range in the context of German in Austria. Unfortunately, there is quite limited empirical evidence for the degree and form of expansion, for the regional distribution of the variation, and especially for its sociolinguistic implications. The main object of this paper is to contribute, based on the studies mentioned above, to the discussion of the variation of MHG /ei/ by sharing recent results from the corpus of the Special Research Program (SFB) "German in Austria. Variation – Contact – Perception" (www.dioe.at). The emphasis will be placed not only on regional differences among Austria's major Bavarian regions, but especially on the sociolinguistic consequences of different situational settings.

Corpus and Method

To consider different subregions of the Bavarian dialect family in Austria, three rural villages were selected with a population between 500 and 2.000. The three locations are Weissbriach in Carinthia as an example of South Bavarian, Neckenmarkt in Burgenland for the (Eastern part of the) South-Central Bavarian transition zone, and Neumarkt/Ybbs in Lower Austria representing Central Bavarian (see Fig. 1).

For each place, a young woman (age between 18 and 35 years) with high formal education (high school 'Matura' graduate) and a classical NORF (non-mobile, old, rural, female; age above 60 years, see [11] for definition) were analysed (auditory analysis).

To capture the broad 'vertical' language spectrum and the speakers' individual speech repertoires, various 'natural' and standardized survey settings were used with different degrees of formality: an interview conducted by a non-local academic, a free conversation among friends, two reading-aloud tasks, and two translation tasks. For the latter, the *Wenker sentences* [12] were presented either in the speakers' local dialect or in standard German (ORF-newscaster), which then had to be translated by the participants into the respective other variety (for more information on the methods, see [13]).

Results

Beforehand, it has to be mentioned that the translations into the speakers' intended standard variety as well as the reading tasks do not show any realizations of /oa/ or /a:/ and therefore are not included in this paper (for the results on these settings and the variants concerning the standard or standard-near registers, see [14]). Rather, for each person three settings (dialect translation task, conversation among friends, and interview) were considered, resulting in 12 data sets and about 1.000 analyzed tokens in total.

a) Regional differences

Before analyzing the vertical-social dimensions of MHG /ei/, the focus is on its horizontal distribution across the three Austrian locations. Fig. 2 shows the relative distribution of (rather) standard variants regarding /ai/ (e.g. [aɛ̯], [æ(ː)e̯])[1] in comparison to the explicitly nonstandard realizations of /a:/ and /oa/ in all three settings analyzed here. These first results demonstrate that there is no strong deviation between the three villages in terms of their general use of the nonstandard variants in comparison to the standard-near form(s). Consequently, between 50% to 60% of the expressed tokens are realized in all three places with /a:/ or /oa/. However, the speakers' actual language behavior differs significantly in their use of /a:/ and /oa/ among all three villages.

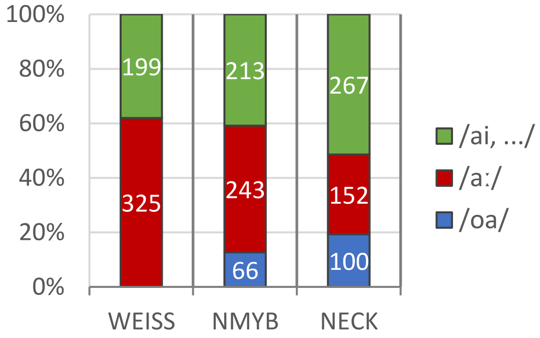


Fig. 2: Distribution of variants of MHG /ei/ for all speakers (old and young) and all situational settings divided by location (WEISS = Weissbriach, South Bavarian; NMYB = Neumarkt/Ybbs, Central Bavarian; NECK = Neckenmarkt, South-Central Bavarian).

Starting from the left, we see no single use of the *oa*-diphthong in Weissbriach, following exactly Wiesingers findings ([5], see Fig. 1) for Carinthia. For Neckenmarkt and Neumarkt/Ybbs on the other hand, a certain amount

[1] The variational linguistic classification of these diphthong-variants from a standard-orientated normative perspective is still controversial (for the discussion see [14]).

of *oa*-forms can be registered, but not to the extent as the Wiesinger map ([5], see Fig. 1) had stated for the 1920s and 1930s.

Certainly, this first analysis does not provide information about which speakers showed which specific language-behavioral patterns in which particular situation. The following graphs (Fig. 3, Fig. 4, Fig. 5) shed light on the frequency of the selected variants for each individual speaker and each situational setting.

Fig. 3 depicts all (absolute) variants for the dialect translation task for each speaker and location. The total number of tokens is quite low due to the limited general appearance of MHG /ei/ in the *Wenker sentences.* This number also varies since some of Wenker's lexemes are quite rare for the overall Bavarian nonstandard context and have therefore been replaced by the participants by more frequent lexemes (e.g. *Kleider* to *Gewand* 'clothes').

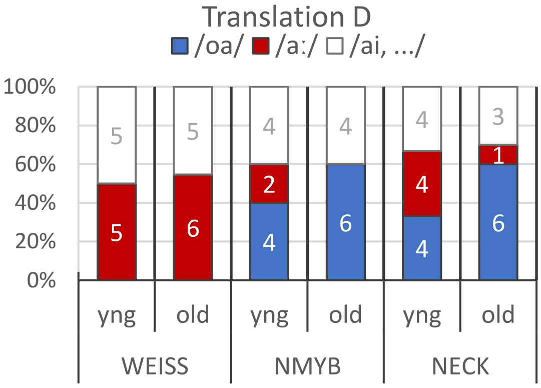


Fig. 3: Distribution of variants of MHG /ei/ in the dialect translation task for old and young speakers divided by location.

b) Generational differences

Apart from regional differences, which were already mentioned above (cf. Fig. 2), now also age differences become apparent. Especially for Neckenmarkt and Neumarkt/Ybbs, higher numbers for the *oa*-diphthong among the old participants (on the left side) can be identified in comparison to the young participants presented on the right side of each location within the graph.

This pattern recurs in the next graph (cf. Fig. 4). The effect of age on whether to use /a:/ or /oa/ becomes even more apparent. Certainly, 'free' conversational tasks yield a greater number of tokens in comparison to a more controlled translation task. Hence, when contrasting the dialect translation task and the conversation among friends, against the background of the Apparent-Time-Hypothesis, a shift toward *a*-monophthongs instead of former *oa*-diphthongs seems to emerge in the Central and South-Central Bavarian areas.

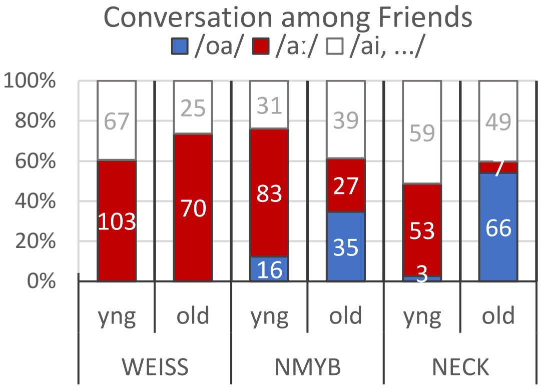


Fig. 4: Distribution of variants of MHG /ei/ in the conversation among friends for old and young speakers divided by location.

c) Situational differences

The final graph shows the results for the interview (cf. Fig. 5). By comparing all three graphs (Fig. 3, Fig. 4 and Fig. 5), also the inter-situational differences can be taken into account.

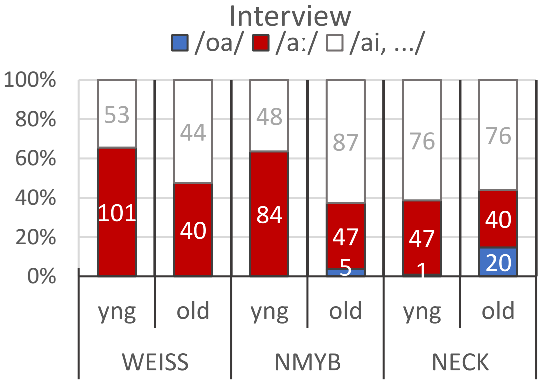


Fig. 5: Distribution of variants of MHG /ei/ in the interview for old and young speakers divided by location.

It is not only possible to identify a general decline of the nonstandard variants starting from the translation task via the conversation among friends to the interview (with the exception of the young participant from Weissbriach, whose amount of 'dialectal speech' seems to remain between 60% to 65% in all settings). Moreover, we can also identify a clear reduction of the traditional *oa*-diphthong for the benefit of the *a*-monophthong in Neckenmarkt and Neumarkt/Ybbs. While the old participants still used the diphthong in about 60% of the cases during the translation task (cf. Fig. 3) and at a rate of 33% to 48% during the conversation among friends (cf. Fig. 4), they decreased its use to less than 15% in the formal interview (cf. Fig. 5). Accordingly, the young speakers of these villages seem to have the diphthong still in mind as a traditional form but rarely seem to use it anymore in their 'free' conversations.

Discussion

MHG /ei/ appears as a highly dynamic phenomenon in the Austrian context in the intersituational and interindividual (cross speakers) contrast, as it seems to differ quite gradually between dialect and standard. This becomes apparent by the steady 'decline' of the *oa*-diphthong from a formal conversational setting, via an informal conversation among friends through to predominantly controlled translation tasks. While the participants still seem to have some kind of concept of the 'base dialect' in their village – elicited in the translation task – the traditional *oa*-variant appears to be increasingly replaced by an *a*-monophthong in 'free' conversations. Hence, compared to Wiesinger's findings for the 1920s and 1930s, the language situation in Eastern Bavarian seems to undergo a change in this regard.

Future investigations will have to consider more locations on the horizontal axis, the state of Vorarlberg as representative for the Alemannic dialect family with its own variational dynamics [14], and additional attitudinal data for a better understanding of the normative implications regarding the variation of MHG /ei/.

References

[1] Žirmunskij, V.M., *Deutsche Mundartkunde: vergleichende Laut- und Formenlehre der deutschen Mundarten*, Wien/Frankfurt am Main: Lang, 2010.

[2] Wiesinger, P., "The Central and Southern Bavarian Dialects in Bavaria and Austria," in *The Dialects of Modern German*, Russ, C.V.J. (Ed.), Stanford: Stanford University Press, pp. 438–519, 1989.

[3] Scheuringer, H., *Sprachentwicklung in Bayern und Österreich*, Hamburg: Buske, 1990.

[4] Lenz, A.N., "Bairisch und Alemannisch in Österreich," in *Language and Space – German. An International Handbook of Linguistic Variation*, Herrgen, J. & J.E. Schmidt (Eds.), Berlin/Boston: De Gruyter, 2019.

[5] Wiesinger, P., *Ergänzungskarten zum Deutschen Sprachatlas. Nacherhebungen in Süd- und Osteuropa*, Marburg, 1962.

[6] Kranzmayer, E., *Historische Lautgeographie des gesamtbairischen Dialektraumes. Verlag der Österreichischen Akademie der Wissenschaften*, Wien/Graz, 1956.

[7] Hornung, M., et al., *Die österreichischen Mundarten*, Wien: Öbv & Hpt, 2000.

[8] Pfalz A., *Lautlehre der Mundart von D. Wagram und Umgebung* Dissertation. Universität Wien, 1910.

[9] Unger J., *Der Nonstandard in Deutsch-Wagram: Unter Berücksichti-gung der Orte Aderklaa und Parbasdorf* Dissertation, Universität Wien, 2014.

[10] Scheutz H., *Strukturen der Lautveränderung: Variationslinguistische Studien zur Theorie und Empirie sprachliche Wandlungsprozesse am Beispiel des Mittelbairischen von Ulrichsberg/Oberösterreich*, Wien: Braumüller, 1985.

[11] Chambers J. K. & P. Trudgill, *Dialectology*, Cambridge: CUP, ^{2}1998.

[12] Schmidt J. E. & J. Herrgen, *Sprachdynamik: Eine Einführung in die moderne Regionalsprachenforschung*, Berlin: Erich Schmidt, 2011.

[13] Lenz A.N., "The Special Research Programme 'German in Austria. Variation – Contact – Perception'," *Sociolinguistica*, 32(1), pp. 269-277, 2018.

[14] Fanta-Jende, Johanna, "Varieties in contact. Horizontal and vertical dimensions of phonological variation in Austria", in *Variationist Linguistics meets Contact Linguistics*, Lenz, A. N. & M. Maselko (Eds.), Göttingen: Vienna University Press, in print.

Lexical transfer and realisations of /ɹ/ by bi- and monolingual learners

Kathrin Feindt[1]

[1]Universität Hamburg

kathrin.feindt@uni-hamburg.de

Abstract

To investigate which linguistic system leaves traces in bilinguals' foreign language English, the oral performance of 16 Turkish-German pupils in the 9th grade was examined for realisations of /ɹ/ phoneme with 18 monolingual German participants of the same age as control group. The study shows that bilinguals as well as monolinguals show equally high performance in target-like production of [ɹ], *with only few substitute sounds. Those can be traced back to general language acquisition mechanisms rather than a background language. Moreover, bilinguals as well as monolinguals not only show the ability to use the correct articulatory settings, in addition, they also employ superordinate phonotactics in switching to the rhotic vowel, given the appropriate phonetic surrounding. Sensitivity towards allophonic variation is most likely due to lexical transfer and the deployment of native language phonological rules.*

Introduction

Even in late stages of the acquisition process, learners are readily identified as belonging to a certain speech community by their accent. Mastering the phonological system of a foreign language seems to trouble learners more than acquiring syntactic rules or lexical items, proving that cross-linguistic influence (CLI) is very persistent in this particular linguistic area [1]. Yet, there are still many unresolved issues in the current state of research, even though phonological CLI has been recognised more lately [2]. Both the types of phonological CLI as well as their conditioning factors are to a large degree still unknown [3].

This is especially true in the case of bilingualism, with the increased number of possible source languages for transfer effects. As balanced bilinguals attain native-like competence in two languages, it is problematic to hypothesise about the speech community that would be attributed to them. This study aims to trace back non-target sounds to a specific phoneme system in order to detect the language that leaves traces in bilinguals' foreign language (FL) oral production.

Methods

The data examined here are recordings of spoken language that was gathered for the quasi-longitudinal research project *Mehrsprachigkeitsentwicklung im Zeitverlauf* (MEZ) [4]. 18 monolingual German and 16 bilingual Turkish-German pupils in grade 9 (15-16 years old) were tested for their abilities in spoken English by telling a story in response to a set of pictures. Recordings were examined for the /ɹ/ phoneme that is particularly prone to variation across languages, whereby the divergence facilitates identification of source languages. The standard realisations in the respective languages are: English ([ɹ] [5]), German ([ʁ] [6]) and Turkish ([ɾ] [7]). Moreover, the phonemes are complemented by phonotactical rules that allow for a range of allophones, of which the vocalic variant is the most prominent representative in German and RP English [5].

In total, there were 162 tokens in 6 types (76 from monolingual and 86 from bilingual speakers) that were analysed phonetically. Criteria for analysis and identification were chosen to match Standard English formant frequencies for the phoneme /ɹ/, especially focussing on F3 values, since the most distinctive feature of the alveolar approximant is a lowered F3, which is thereby only narrowly separated from F2 [8]. The corpus was assessed by opening the sound file in praat [9], listening through the individual recordings participant per participant, while firstly annotating words containing sounds of interest. In a second step, the speech stream was zoomed in on this word, a further tier was used to note down the phonemes by perception, and sounds in question were zoomed in again. After that, frequencies and movements of formants were gathered. Frequencies were calculated automatically by praat [9] via the "get first formant", "get second formant" etc. function after selecting the most stable middle part of the sound in the peak of the vowel or consonant.

After measuring formant frequencies for the language groups separately, statistical analyses were performed with the computer program SPSS [10] and the Chi^2 test was used to determine statistical significance.

Results

Overall, the number of accurately produced instances of /ɹ/ is very high; both groups show a comparable target-like usage of this particular FL sound.

In cases where the alveolar approximant was aimed at, it was generally produced as such. The outstanding performance of all participant irrespective of language background is summarised in table 1, showing very few substitutional sounds.

Phoneme	Frequency	Percent
Unclear	1	0.61
Alveolar Approximant	78	47.85
Rhotic Vowel	69	42.33
Bilabial Approximant	14	8.59
Uvular Fricative	1	0.61
Total	163	100.0

Tab. 1: Token counts and percentage of realisations of the /ɹ/ phoneme.

Most non-target sounds are bilabial approximants, which can neither be traced back to Turkish nor German. Negative transfer from German occurs only once; the Turkish tap is not used as a substitution at all. The number of background languages proves to be a negligible variable (p-value = 0,292; df = 3; χ^2 = 3,73). In addition, the variability of the /ɹ/ phoneme is accounted for by a high sensitivity for allophonic variation: Token counts for rhotic vowel nearly equal those of the consonant.

The choice of the specific realisation is determined strongly by the phonetic environment, with a highly significant p-value for this variable of 0,000 (df = 6; χ^2 = 168,83). Bilingual as well as monolingual participants prefer to use the /ɹ/ in C_V (*bread* /bɹɛd/) environments, but are nevertheless able to convert this into the vocalic variant in V_C cases (*first /fəːst/*). A graphic visualisation of learners' spoken sounds according to the phonetic environment is given in figure 1.

The figures show that [w] most often used as a substitution by both groups in a vocalic environment, while the four cases of this sound in the C_V environment were exclusively uttered by bilingual students. On the other hand, the occurrence of the German [ʁ] in this environment can be traced back to a monolingual participant. In the majority of cases, learners produce [ɐ] rather than the alveolar approximant in a V_C environment. No significant differences between the groups can be detected in this respect (p-value = 0,369; χ^2 = 0,65; df = 1).

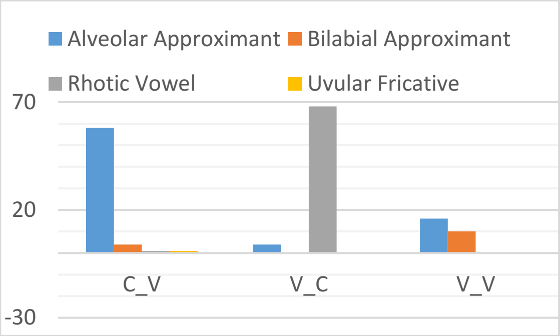


Fig. 1: Token counts in different phonetic environments.

Transfer of Lexemes

These astonishing results might be due to a positive transfer of lexemes from German. The category of V_C comprises almost exclusively the word <supermarket>, which is part of the vocabulary of all languages involved, although with varying pronunciation, displayed in (1).

(1) a. English (RP): /ˈsuːpəˌmɑːkɪt/
b. German: /ˈzuːpɐˌmɐkt/
c. Turkish: /ˈsyːpərˌmɑːɾkət/

Both English and German employ a similar phonotactical rule by which the consonant is pronounced as a vowel. As displayed in (1c), Turkish has no comparable rule, so that the /ɹ/ phoneme is realised as a consonant, regardless of the phonetic environment in which it appears.

Isolating and analysing tokens from this particular lexeme reveals no difference between the groups (p-value = 0,409; χ^2 = 0,61; df = 1), as the figures given in table 2 illustrate. Both monolinguals and bilinguals prefer to use the vocalic allophone; there are only few instances of the alveolar approximant produced, and there are no non-target-like sounds.

Language	Phoneme		
	Alveolar Approximant	Rhotic Vowel	Total
Bilingual	3.00	28.00	31.0
	9.68%	90.32%	100%
Monolingual	1.00	23.00	24.0
	4.17%	95.83%	100%
Total	4.00	51.00	55.0
	7.27%	92.73%	100%

Tab. 2: Token counts and percentage of realisations of the /ɹ/ phoneme for the word <supermarket> by mono- and bilingual speakers.

Discussion

The surprisingly high level of proficiency in target-like pronunciation of English /ɹ/ make detection of source languages difficult. Not only was the alveolar approximant successfully produced in the majority of cases, learners also followed phonotactical rules in changing from the consonantal to the vocalic variant in appropriate phonetic contexts. The high accuracy may be due to (i) accommodation to English RP or (ii) transfer from German. The advanced stage of acquisition, taken together with the overall outstanding performance in producing the alveolar approximant, suggests the former. That would imply, though, that all participants independent of language background, school type, teacher and even residence, orient towards the English variety as it is spoken in Great Britain. Extensive transfer of German phonotactics by both monolingual and bilingual participants would ascribe a prominent role to German as the preferred source language, at least in the context of positive transfer of lexemes. At this stage, results are not conclusive enough to verify one of these hypotheses. The special case of the word <supermarket> draws attention to an unresolved issue within research of phonological CLI. As [3] pointed out, the relationship between lexical and phonological transfer remains unclear. Thus, it is debateable whether the influence of one of the languages is tied to single words, or whether phonological rules are instead employed on a more general, systematic level.

The occurrence of the bilingual approximant as a substitute sound, which is neither part of the German nor the Turkish phoneme inventory, suggest that general learning mechanisms play a significant role in phonological CLI. Previous studies on the /ɹ/ phoneme in the context of language acquisition illustrate that substituting the alveolar approximant with a bilabial approximant is a strategy often observed in English children acquiring their mother tongue. For example, [11] carried out a study investigating children's phonological development. The speech production of four German children aged four to nine in their FL English was examined for realisations of the /ɹ/ phoneme, in comparison to English children of the same age group. Results show that participants acquired the foreign [ɹ] in a pattern that is similar to the phonological development of English as a native language, namely that it was pronounced as [w]. [11] concludes that the acquisition of at least this particular phoneme is not accomplished by relying on the previous L1s [11]. Other, universal constraints on human speech processing have to be inferred for older learners as well, as this investigation showed.

Conclusion

Analysing the phonological development of bilingual learners acquiring a further language points to a complex interaction of various factors involved in CLI. The synergy makes it hard to predict the role of bilingualism, as was shown in the case of the English the /ɹ/ phoneme. Tracing back non-target-like sounds to one of the background languages is not always possible, as the phone cannot easily be detached from the systematic level of phonology. Moreover, omnipresent cognitive learning processes add to the multifarious picture.

It has been shown, though, that children with a migration background are neither advantaged nor disadvantaged in learning a further language in school when compared to a monolingual peer group. At least at a more advanced stage of the acquisition process (in this case, four years of training), participants' performance is on a remarkably high level, irrespective of background languages. Although results are not conclusive concerning the status of German, it is unambiguous that the Turkish phoneme system did not leave traces in bilinguals' oral production of the FL English.

To get a clearer picture of learner's abilities, future research needs to trace phonological development over a longer period. Reliable conclusions concerning an adjustment to English phonology can only be drawn by comparing advanced learners' to beginners' realisations of /ɹ/.

References

[1] O'Brien, M. G., “Pronunciation matters,” *Die Unterrichtspraxis/Teaching German*, 37 (1), pp. 1–9, 2004.

[2] Falk, Y. & C. Bardel, “The study of the role of the background languages in third language acquisition. The state of the art,” IRA*L – International Review of Applied Linguistics in Language Teaching*, 48 (2-3), 2010.

[3] Gut, U., “Cross-linguistic influence in L3 phonological acquisition,” *International Journal of Multilingualism*, 7 (1), pp. 19–38, 2010.

[4] MEZ – *Mehrsprachigkeitsentwicklung im Zeitverlauf* (2014-2019); Projektkoordination: Prof. Dr. Dr. h. c. Ingrid Gogolin, Universität Hamburg; © MEZ 2017.

[5] Ladefoged, P., “American English,” in *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet*, International Phonetic Association (Ed.), Cambridge etc.: Univ. Press, pp. 41–44, 1999.

[6] Krech, E.M. & E. Stock, *Deutsches Aussprachewörterbuch*, Berlin: De Gruyter, 2009.

[7] Yavuz, H. & B. Ayla Balcı, *Turkish Phonology and Morphology (Türkçe Ses ve Biçim Bilgisi)*, Eskişehir: Anadolu Üniversitesi, 2011.

[8] Ladefoged, P., *Vowels and consonants. An introduction to the sounds of languages*, Malden, MA: Blackwell, 2004.

[9] Boersma, P. & D. Weenink, Praa*t: doing phonetics by computer*, 2016. Version 6.0.21. (http://www.praat.org/ retrieved: 10th of October 2016).

[10] IBM Corp (2015): *IBM SPSS Statistics for Windows. Version 23*, Armonk, NY: IBM Corp.

[11] Wode H., “Perception and Production in learning to talk,” in *Focus on phonological acquisition*, Hannahs, S.J. & M. Young-Scholten (Eds.), Amsterdam: Benjamin, pp. 17–47, 1997.

Uptalk in German: Investigation on the GECO database

Fabian Fey[1], Natalie Lewandowski[1,2]

[1]University of Stuttgart, [2]High Performance Computing Center, Stuttgart

fabian.fey@ims.uni-stuttgart.de, natalie.lewandowski@hlrs.de

Abstract

The term Uptalk refers to the occurrence of rising intonation patterns at the end of declarative phrases in spontaneous speech. To date, this phenomenon has been explored primarily in the English language. Other languages, including German, have been neglected. The present study attempts to fill this gap in the current literature by presenting findings from a corpus study on the GECO (German Conversations) database. We were able to provide conclusive evidence that high rising terminals (HRTs) are used frequently in non-questions by female native speakers of German. Moreover, the results hint at potential correlations between the speakers' personality, the modality of the conversation and the usage of Uptalk.

Introduction

The term *Uptalk* refers to the occurrence of rising intonation patterns at the end of declarative phrases in spontaneous speech. In general, a falling pitch is associated with a declarative sentence, while a rising pitch at the end of an utterance marks a question. Spontaneous speech, however, appears to operate under a different set of rules.

Up to now, this phenomenon has been explored primarily in the English language. In other languages, it has received little to no attention at all. This may be because *Uptalk* used to have a negative connotation, often being associated with insecurity on part of the speaker. Fortunately, *Uptalk* has been able to redeem itself in recent research and is now widely accepted as being a crucial part of communication because of its significant contribution to a better understanding between all interlocutors [1].

To date, no study has investigated whether or not *Uptalk* is also used in German. Therefore, the present study aims to fill this gap in the literature. It presents evidence obtained via a corpus study on the GECO (German Conversations) database [2–4] which provides conclusive evidence for the frequent existence of high rising terminals (*HRTs*) in non-questions of female native speakers of German. Moreover, it delivers first hints on potential correlations between speaker personality characteristics and *Uptalk*.

State of the Art

Formally, *Uptalk* has been defined as a rise of pitch on the final syllable(s) of the last word(s) of a declarative phrase, or a continuous rise of pitch across its last words, with a significant increase on the last one [5]. As mentioned above, the once negative connotation of *Uptalk* recently gave way to a more refined and sophisticated perspective. Nowadays, it is associated with many useful features contributing to a better understanding between all interlocutors [1]. Especially for the English language, there is a large number of existing studies focusing on these features. The following section attempts to provide an overview of the most recent literature.

Instead of communicating a general insecurity on part of the speaker, *Uptalk* may communicate his or her uncertainty, for instance concerning the accuracy of an answer. This allows the interlocutor to evaluate the answer and possibly correct it. Shepherd [6] observed this behaviour in interactions between students and their teachers, but believes it to be maintained in later life, too. Moreover, *Uptalk* is also used to assert dominance in conversations. Cheng and Warren [7] found this technique to be used by the leading figure in business meetings, most likely to exert control over the conversation. This finding agrees with what House [8] observed in her study: Her results show that *Uptalk* is used to make sure that the interlocutor is still able to follow the conversation. The discovery that *Uptalk* is often able to elicit minimal verbal reactions, like an affirmative 'Uh-hu', from the interlocutor [9] corroborates this once more.

While the features mentioned above hint at the question-like characteristics of *Uptalk*, the following studies provide a more diverse picture. Research by Wichmann and Caspers [10] shows that *Uptalk* is also used to manage and control turn-taking in conversations. A distinctive H*H% contour signals a change of speaker, while an H*% contour is an unmistakable indicator that the current speaker is not to be interrupted (also called *floor holding*). Finally, *Uptalk* can also be used to mark upcoming new information in a presentation [11, 12].

Data

The data for the present study is taken from the GECO database [2–4]. It consists of annotated recordings of 46 dialogues between 13 female native German speakers. All recordings are of spontaneous, unmoderated speech. Since each speaker was recorded separately, the database holds 92 recordings in total, each being roughly 25 minutes long. The recordings were made in two different modalities: During the monomodal recordings the interlocutors had no visual contact, whilst for the multimodal recordings the two speakers were facing each other.

Not all of the available annotation layers in GECO were used to evaluate the dialogues. The present study focused on the word, tone and syllable, as well as on information from a self-monitoring test each speaker had to take [2]. The personality features from this test included *acting, other-directedness, extraversion* and *sensitivity to expressive behavior and social cues*. The word labels were based on a transcript [13], while both the tone and syllable labels were generated automatically.

The tone labels used in the GECO database are based on the ToBI system [14], which is commonly used to annotate pitch rises, including HRTs. This system uses the character *H* to mark a high intonation and *L* to mark a low intonation. These characters can be concatenated with multiple different symbols, for instance "+" and "*", to mark more complex intonational phrases. H*, for example, marks a high accent. Additionally, the character "%" marks phrase-ends, while the character "-" marks intermediate phrase-ends. Hence, "H%" marks a high pitch at a phrase-end. To deal with the German language, Mayer [15] extended this system, which is known as the GToBI(S) system. For the present study, only the labels "H%" and "H-" are of interest: Both mark a rising pitch on an intermediate phrase-end and therefore either a question or *Uptalk*.

Methods

The data from the 92 dialogue recordings described above was analysed automatically with regard to rising intonation in phrase-end positions, using the labels provided by GTobi(S) for all syllable-related and orthographical information like punctuation marks. As mentioned above, the GTobi(S) labels H% and H- where used to identify a rising intonation in phrase-end positions, since both indicate a rising intonation at an intermediate phrase-end. Only words which were directly followed by a punctuation mark, meaning a full stop, an exclamation mark, and a question mark were considered in the analyses.

We were able to identify 9429 phrase-final words simultaneously bearing a GTobi(S) label. 4607 (48,86%) of these were labelled with either H% or H-. All H- labels where excluded from this study because they only represent a rise on an intermediate phrase-end and have a relatively low occurrence in the dataset (<1%). Furthermore, 592 questions were excluded from the analysis. All remaining H% labels (3978 words) were potential instances of *Uptalk* (see Figure 1) and were frequent enough to justify further experiments. The actual presence of *Uptalk* in this subset was subsequently verified auditorily on a random set of samples extracted from the recordings at time stamps of the target words (including preceding context) and thus could be confirmed. This initial investigation was necessary in order to verify that *Uptalk* is present in the data and provides an incentive to conduct further analyses on the dataset.

Results

Apart from providing evidence for the existence of Uptalk in the German language, another aim of the present study was to investigate whether there is a relationship between the subjects' scores in a self-monitoring test (see section "Data") and the usage of *Uptalk*. Since the dependent variable in the data set (i.e. *Uptalk*) was a binary distinction between it being either present (i.e. a H% label in phrase-final positions in non-interrogative sentences) or absent (i.e. all remaining GTobi(S) labels), a generalized linear model (glm) of the family *binomial* was used for the analysis. This glm was fitted with

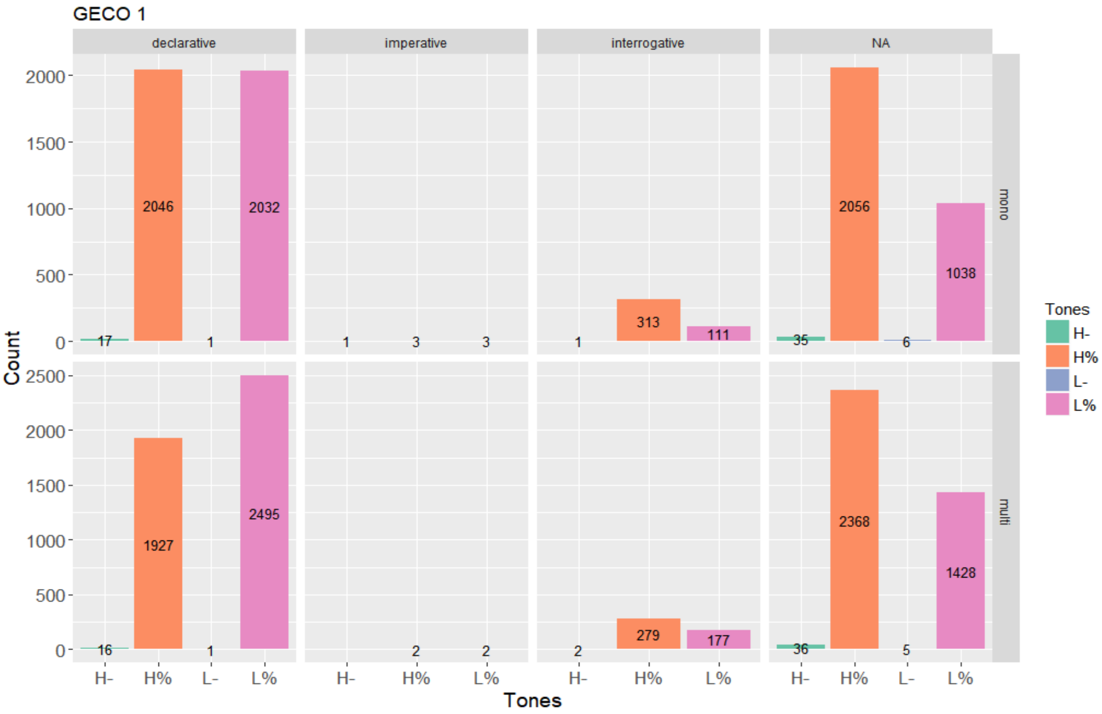


Fig. 1: Instances of phrase-final words and their GTobi(S) labels. Data are further split according to sentence type and dialog condition – mono (auditory only) vs. multi (with visual contact).

the factors *modality* (monomodal or mxulti-modal), the set of personality features (*acting, other-directedeness, extraversion* and *sensitivity to expressive behavior and social cues* from the aforementioned self-monitoring test) and the *word frequency*. This glm revealed that higher scores regarding the personality features *acting* $(\Pr(>|z|=0.0000))$ and *other-directedness* $(\Pr(>|z|=0.0099))$ significantly impacted the presence of Uptalk in the GECO database, as did *word frequency* $(\Pr(>|z|=0.0000))$. While the *modality* of a dialogue on its own proved not to have an influence on the presence of *Uptalk*, a significant interaction of *modality* combined with *acting* $(\Pr(>|z|=0.0015))$, *extraversion* $(\Pr(>|z|=0.0252))$, and *sensitivity* $(\Pr(>|z|=0.0149))$ was found, indicating that *acting* and *sensitivity* have a negative effect on the amount of *Uptalk* in the multimodal condition, while *extraversion* has a positive effect on *Uptalk* in the same condition.

A further glm with the dependent variable set to *Uptalk* was fitted with the factors *speaker* and *partner*. The glm revealed a significant interaction $(\Pr(>|Chi|=1.658e-06, df=5))$ between the subjects and the use of *Uptalk*. This result indicates strong individual differences in the amount of *Uptalk* used per speaker and within a specific speaker-partner constellation, pointing to *Uptalk* as a technique that is used very dynamically and situation-dependently.

Conclusion

By means of an analysis of the data provided by the GECO database, the present study examined whether *Uptalk* is a language-dependent phenomenon of English or can also be found in other languages, specifically German. The results were positive: Declarative sentences with a high pitch rise have been identified via orthographic information and the respective GToBI(S) label, and a subsequent random auditory test was able to verify the presence of *Uptalk* in those utterances.

Furthermore, an analysis using a generalized linear model delivered first hints on potential correlations between the speakers' personality traits and *Uptalk*. Higher scores on the features *acting* and *other-directedness* appear to increase the usage of *Uptalk* in conversations. Moreover, the same personality traits differ in their effect on the usage of *Uptalk* when the conversation's modality is taken into consideration.

Finally, a third analysis using a glm revealed a significant interaction between the amount of *Uptalk* used and the respective speaker-partner constellation. These strong individual differences in the amount of *Uptalk* used in conversations show that it is a highly dynamic and situation-dependent phenomenon.

References

[1] Douglas, N., in “Uptalk Actually Serves a Powerful Purpose”, 2017 (http://lifehacker.com/uptalk-actually-serves-a-powerful-purpose-1795688458) [12.08.2017].

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

[3] Schweitzer, A. & N. Lewandowski, “Convergence of Articulation Rate in Spontaneous Speech,” *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pp. 525–529, 2013.

[4] Schweitzer, A. & N. Lewandowski, IMS GECO Datenbank (http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.html), 2015.

[5] Britain, D., “Linguistic change in intonation: The use of high rising terminals in New Zealand English,” *Language variation and change*, 4(1), pp. 77–104, 1992.

[6] Shepherd, M., “Functional significance of rising-intonation declaratives in settings with special discursive norms,” *LSA Annual Meeting Extended Abstracts*, 2(0): 10–1–5, 2011.

[7] Cheng, W. & M. Warren, “The use of rise and rise-fall tones in the Hong Kong Corpus of Spoken English,” *International journal of corpus linguistics*, 10, pp. 85–107, 2005.

[8] House, J., “The role of prosody in constraining context selection: A procedural approach,” *Cahiers de Linguistique Francaise 28: Interfaces discours-prosodie*, 28, pp. 369–383, 2007.

[9] Allan, S., *The rise of New Zealand intonation*, Multilingual Matters, pp. 115–128, 1990.

[10] Wichman, A. & J. Caspers, *Melodic cues to turn-taking in English: evidence from perception*, 2001.

[11] McLemore, C. A., *The pragmatic interpretation of English intonation: sorority speech*, PhD thesis, University of Texas at Austin, 1991.

[12] Guy, G. R. et al., “An Intonational Change in Progress in Australian English,” *Language in Society*, 15(1), pp. 23–51, 1986.

[13] Rapp, St., “Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov models - An aligner for German,” *Proceedings of ELSNET Goes East and IMACS Workshop ¨Integration of Language and Speech in Academia and Industry” (Moscow, Russia),* 1995.

[14] Hirschberg, J. & M. E. Beckman, *The ToBI annotation conventions,* 1994.

[15] Mayer, J., *Transcription of German intonation - The Stuttgart system*, Technical report, Institute of Natural Language Processing, University of Stuttgart, 1995.

Junge oder Mädchen?
Zur Geschlechtsidentifikation präpubertärer Stimmen

Riccarda Funk[1], *Susanne Voigt-Zimmermann*[1], *Adrian P. Simpson*[2]

[1]Martin-Luther-Universität Halle-Wittenberg, [2]Friedrich-Schiller-Universität Jena

riccarda.funk@web.de, susanne.voigt-zimmermann@sprechwiss.uni-halle.de, adrian.simpson@uni-jena.de

Abstract

*Gegenstand der Studie ist die Geschlechtsidentifikation präpubertärer Stimmen durch erwachsene Hörer*innen. Dazu wurden spontansprachliche und standardisierte Stimuli von 20 Kindergartenkindern im Alter von vier bis fünf Jahren und 30 Grundschulkindern im Alter von sechs bis sieben Jahren 85 Hörer*innen in zwei Gruppen randomisiert vorgespielt. Die Hörer*innen wurden gefragt, ob die über Kopfhörer präsentierte Aufnahme von einem Jungen oder einem Mädchen gesprochen wurde und wie sicher sie sich mit ihrer Entscheidung waren. Im Anschluss wurden verschiedene akustische Korrelate des Geschlechts gemessen und mit den Ergebnissen der Identifikationsexperimente in Zusammenhang gebracht. Im Ergebnis konnten 72,5% der Stimuli dem korrekten Geschlecht der Kinder zugeordnet werden. Dabei konnten anhand der akustischen Analysen nur wenige signifikante Unterschiede nachgewiesen werden.*

Einleitung

Das auditive Erkennen des Geschlechts einer Person wird durch verschiedene Faktoren beeinflusst. Eine wichtige Rolle spielen sowohl anatomische und akustische Merkmale (wie die Größe des Kehlkopfes, die Grundfrequenz und die Formanten) als auch sozial gelernte Verhaltensmuster des Sprechenden. Die Unterschiede zwischen erwachsenen Männern und Frauen sind dabei so groß, dass Hörer*innen 'männliche' und 'weibliche' Stimmen ohne Schwierigkeiten korrekt bestimmen können. Jungen und Mädchen weisen vor der Pubertät hingegen lediglich geringe Differenzen hinsichtlich ihres Stimmapparates [1] und der Größe ihres Kehlkopfes auf [2]. Das biologische Geschlecht hat bei Kindern demnach nur einen geringen Einfluss auf die Stimmgebung. Dennoch belegen einige Studien, dass das Geschlecht von Jungen und Mädchen nur anhand der Stimme ebenfalls signifikant richtig bestimmt werden kann [3–6].

Bisher durchgeführte Studien weisen einige Nachteile auf: Meist waren die untersuchten Kinder älter als sieben Jahre (hier können beginnende anatomische und endokrinologische Veränderungen der Stimme im Zuge der Pubertät nicht vollständig ausgeschlossen werden [7]), es wurden nur wenige Kinderstimmen aufgenommen oder die Anzahl der befragten Hörer*innen im Experiment war sehr gering. Um bisher gewonnene Erkenntnisse zu überprüfen, haben wir ähnliche Experimente wie bei [3] und [6], jedoch mit einer großen Anzahl jüngerer Kinder (50; 25w, 25m) und vielen Hörer*innen (85; 48w, 37m) durchgeführt. Mithilfe von Identifikationsexperimenten und akustischen Messungen sollen folgende Hypothesen überprüft werden:

- Die Hörer*innen können das Geschlecht der Kinder zu mindestens 65% korrekt bestimmen (wie bei [3–6]). Die Trefferquote des Geschlechts ist bei älteren Kindern höher als bei jüngeren Kindern [5].
- Das Geschlecht der Kinder wird in Spontansprache häufiger korrekt bestimmt als beim Zählen (ähnlich wie bei [5]), da die Variationsmöglichkeit bei Spontansprache größer ist. Hier können mehr Verhaltensmuster durch die Kinder transportiert werden, weshalb die Geschlechtszuordnung leichter fällt.
- Die Stimuli mancher Kinder werden besonders häufig (> 80%) einem bestimmten Geschlecht zugeordnet, auch wenn dieses nicht mit dem biologischen Geschlecht des Kindes übereinstimmt (wie bei [6]).
- Die Grundfrequenzen und Halbtonumfänge der Jungen und Mädchen unterscheiden sich nicht signifikant voneinander (wie bei [5- 10]).
- Im Mittel weisen die Mädchen höhere Formantwerte, höhere Sprechgeschwindigkeiten und höhere HNR-Werte auf als die Jungen (wie bei [6, 11–13]).

Methode

Für die Überprüfung der Hypothesen wurden Aufnahmen von 20 vier- bis fünfjährigen Kindergartenkindern (12w, 8m) und 30 sechs- bis siebenjährigen Grundschulkindern (13w, 17m) angefertigt. Diese sprachen Deutsch als Muttersprache und hatten keine nachgewiesenen Sprach-, Sprech- oder Hörschwierigkeiten. Die Zustimmung der Eltern und des zuständigen Schulamts wurde im Vorfeld eingeholt.

Alle Aufnahmen fanden in einem ruhigen Raum der Schule oder des Kindergartens statt. Zunächst wurde den Kindern der Ablauf erklärt und sie wurden nach ihrem Alter gefragt. Anschließend wurde ein Zoom H4-Rekorder in ca. 20 cm Entfernung zum Kind positioniert. Die Abtastfrequenz betrug 44.1 kHz und die Amplitudenauflösung 16 Bit. Es wurden sowohl Spontansprache als auch einheitliches Sprachmaterial aus folgenden Aufnahmetypen gewonnen:

- Beschreibung von drei altersgerechten Situationsbildern (entnommen aus [14])
- Benennung von Gegenständen auf neun verschiedenen Bildern mit den Vokalen /aː/, /iː/ und /uː/ im Wortakzent
- Nachsprechen der Satzliste aus [6]
- Zählen in vorgegebenem Rhythmus von eins bis zehn

Für die Hörexperimente wurden die Beschreibung des Vogel-Bildes (siehe Abbildung 1) und das Zählen verwendet. So entstanden 40 Stimuli von den Kindergartenkindern und 60 Stimuli von den Grundschulkindern.

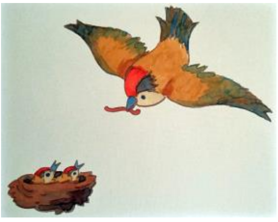


Abb. 1: Vogel-Bild

Um auszuschließen, dass das inhaltlich Gesagte in der Vogel-Aufnahme einen starken Einfluss auf die Wahrnehmung des Geschlechts im Hörexperiment besitzt, wurde ein Leseexperiment erstellt. In diesem wurden alle 50 Vogel-Stimuli orthografisch transkribiert. 30 Proband*innen (die nicht an den Hörexperimenten teilnahmen) beurteilten auf einer Fünfpunkteskala, ob der Inhalt auf sie ‘typisch weiblich’ (=1) oder ‘typisch männlich’ (=5) wirkte.

Die Hörexperimente fanden in Halle, Jena, Naumburg (Saale) und Berlin statt und wurden in einem ruhigen Raum durchgeführt. Der Ablauf, die Gerätetechnik und Wiedergabelautstärke beim Hörexperiment waren stets identisch.

Zunächst wurden demographische Angaben der Hörer*innen mittels Fragebogen erfasst. Im Anschluss wurden die Identifikationsexperimente mithilfe des Tools ExperimentMFC in Praat [15] durchgeführt.

Am Experiment nahmen insgesamt 85 Hörer*innen (48w, 37m) im Alter von 18 bis 72 Jahren (im Mittel 34 Jahre) teil. Diese wurden in zwei Gruppen aufgeteilt. Die erste Gruppe hörte randomisiert die Stimuli der Kindergartenkinder, die zweite Gruppe die Stimuli der Grundschulkinder. Nach dem Hören sollten die Proband*innen entscheiden, ob es sich bei der gehörten Stimme um einen Jungen oder um ein Mädchen handelt. Zudem wurden sie auf einer Fünfpunkteskala gefragt, wie sicher sie sich mit ihrer Aussage sind.

Die akustischen Messungen der Grundfrequenz und des Sprechhalbtonumfangs wurden in der Aufnahme des Zählens durchgeführt. Die Messung der Formanten und HNR-Werte erfolgte anhand der Vokale /aː/, /iː/ und /uː/ der Bildnennungen, die Sprechgeschwindigkeit wurde in der Aufnahme des Satzes <Die Oma mag Urlaub am Meer.> ermittelt.

Ergebnisse

Mit dem vorgeschalteten Leseexperiment konnte ein starker Einfluss des Inhalts der Vogel-Aufnahmen auf das Hörergebnis ausgeschlossen werden. Lediglich die Aussage eines Jungen zeigte mit einem Wert von über 4 eine starke Tendenz zu ‘typisch männlich’; dieser wurde im Folgenden nicht ausgewertet.

In den Hörexperimenten konnten insgesamt 72,5% der Stimuli korrekt dem Geschlecht der Kinder zugeordnet werden. Die Trefferquote der Grundschulkinder lag mit 74,6% etwas über der Trefferquote für die Kindergartenkinder mit 69,5%. Zwischen der Trefferquote der Jungen und Mädchen trat kein signifikanter Unterschied auf (*W=297,5, p=0,97*).

Die Vogel-Aufnahme führte mit 76,2% zu einer höheren Trefferquote als die Zählen-Aufnahme mit 72,9%. Dieser Unterschied ist nicht signifikant (*t=1,16, p=0,25*).

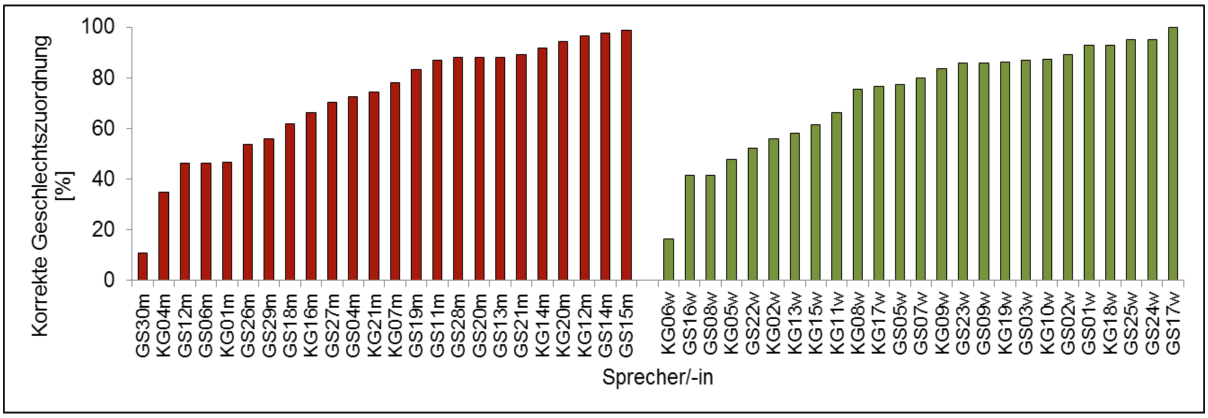


Abb. 2: Korrekte Geschlechtszuordnung der einzelnen Kinder in aufsteigender Reihenfolge (m links in rot; w rechts in grün; KG Kindergarten; GS Grundschule).

	Jungen	Mädchen
Korrekte Geschlechtszuordnung	72,1%	67,3%
	71,1%	78,8%
F0	280 Hz	273 Hz
	251 Hz	267 Hz
Halbtonumfang	8,6	8,2
	7,0	7,3
F1 /a:/	1094 Hz	1110 Hz
	878 Hz	853 Hz
F2 /a:/	1683 Hz	1798 Hz
	1600 Hz	1605 Hz
F1 /i:/	422 Hz	426 Hz
	399 Hz	398 Hz
F2 /i:/	3174 Hz*	3346 Hz*
	3047 Hz*	3248 Hz*
F1 /u:/	447 Hz	434 Hz
	371 Hz	402 Hz
F2 /u:/	1008 Hz	912 Hz
	780 Hz	822 Hz
HNR	19,3 dB	18,4 dB
	19,0 dB	17,4 dB
Silben/s (inkl. P.)	2,5	3,0
	3,7*	4,2*
Silben/s (exkl. P.)	3,2	3,5
	4,2	4,4

Tab.1: Ergebnisse der akustischen Analyse (signifikante Unterschiede=; weiß=Kindergarten, grau=Grundschule).*

Einige Kinder wurden mit einer Trefferquote von über 80% korrekt identifiziert, andere hingegen zu weniger als 20%. Bei einem Grundschulmädchen (GS17w) lag die Trefferquote bei 100%.

Die Hörerinnen beurteilten das Geschlecht signifikant häufiger richtig als die Hörer. Alle anderen abgefragten Parameter der Hörer*innen wie Alter, das Vorhandensein eigener Kinder, der Kontakt zu Kindern im Beruf oder die Expertise in der Stimmbeurteilung hatten keinen signifikanten Einfluss auf das Ergebnis.

In den akustischen Messungen ließen sich nur wenige signifikante Unterschiede finden. Der zweite Formant /i:/ liegt bei den Mädchen in beiden Gruppen über dem der Jungen: *t=-2,33, p=0,03* (Kindergarten) und *t=-2,96, p=0,006* (Grundschule). Zudem sprechen die Mädchen der Grundschule signifikant schneller als die Jungen, wenn man die Sprechpausen einberechnet: *t=-2,42, p=0,02*. Anders als erwartet liegen die HNR-Werte der Jungen über den HNR-Werten der Mädchen. Dieser Unterschied ist jedoch nicht signifikant.

Auffälliger sind die akustischen Unterschiede zwischen beiden Altersgruppen. Die Kindergartenkinder sprechen insgesamt höher und besitzen höhere Formantwerte als die Grundschulkinder, ähnlich wie bei [11, 12]. Zudem ist der Halbtonumfang der jüngeren Kinder größer. Begründet werden kann dies mit der höheren Emotionalität, mit der die jüngeren Kinder das Bild beschrieben. Weiterhin sprechen die Kindergartenkinder deutlich langsamer als die Grundschulkinder, vermutlich durch die noch schwächer ausgeprägten kognitiven und feinmotorischen Fähigkeiten [16].

Diskussion

Studien haben gezeigt, dass sich Jungen und Mädchen vor der Pubertät im Stimmapparat und Vokaltrakt anatomisch nicht voneinander unterscheiden [1, 2]. Da die Trefferquote in den Identifikationsexperimenten mit insgesamt 72,5% dennoch sehr hoch ist, scheinen die Kinder bereits früh ‘typisch männliche’ oder ‘typisch weibliche’ Sprechweisen zu lernen.

Offen bleibt jedoch die Frage, woran genau sich die Hörer*innen beim Bestimmen des Geschlechts orientieren. Zwischen Jungen und Mädchen traten zwar messbare geschlechtsspezifische Unterschiede in der Stimme und Sprechweise auf, diese waren jedoch nur in seltenen Fällen signifikant.

Welche Parameter führen stattdessen dazu, dass erwachsene Hörer*innen das Geschlecht der Kinder so gut bestimmen können? Zur Beantwortung dieser Frage sollen in kommenden Arbeiten Wahrnehmungsexperimente durchgeführt werden. Weiterhin sollen die Aufnahmen der Jungen und Mädchen mit sehr hoher Trefferquote in ihren physikalisch messbaren Werten miteinander verglichen werden, möglicherweise fallen die Differenzen hierbei stärker aus, als wenn man alle Kinder miteinander vergleicht. Spannend wäre auch zu wissen, wie sich die Kinder, die sehr ‘weiblich’ oder ‘männlich’ klingen, selbst wahrnehmen. Hier könnten Fragebögen zur Geschlechtsidentität möglicherweise Aufschluss darüber geben, wie die Kinder ihr soziales Geschlecht empfinden.

Referenzen

[1] Fitch, W. T. & J. Giedd, “Morphology and development of the human vocal tract: A study using magnetic resonance imaging,” *Journal of the Acoustical Society of America*, 106(3), pp. 1511–1522, 1999.

[2] Kahane, J., “A morphological study of the human prepubertal and pubertal larynx,” *Developmental Dynamics*, 151, pp. 11–19, 1978.

[3] Günzburger, D. et al., “Voice identication of prepubertal boys and girls by normally sighted and visually handicapped subjects,” *Language and Speech*, 30, pp. 47–58, 1987.

[4] Karlsson, I. & M. Rothenberg, “Intercultural variations in gender-based language differences in young children,” *Speech research summary report* STL-QPSR 33, pp. 1–17, 1992.

[5] Klein, C., “Acoustic and perceptual gender characteristics in the voices of pre-adolescent children,” *Phonus*, 9, pp. 221–329, 2004.

[6] Simpson A. P. et al., “Perceptual and acoustic correlates of gender in the prepubertal voice,” *Interspeech Stockholm*, pp. 914–918, 2017.

[7] Kahl, H. et al., *Sexuelle Reifung von Kindern und Jugendlichen in Deutschland. Ergebnisse des Kinder- und Jugendgesundheitssurveys*, Berlin: Springer Medizin Verlag, 2007.

[8] Robb, M. P. & J. O. Simmons, “Gender comparisons of children’s vocal fold contact behavior,” *Journal of the Acoustical Society of America*, 88(3), pp. 1318–1322, 1990.

[9] Whiteside, S. P. & C. Hodgson, “The Development of Fundamental Frequency in 6- to 10-Year Old Children: A Brief Study,” *Journal of the International Phonetic Association*, 28, pp. 55–62, 1998.

[10] Ferrand, C. T. & R. L. Bloom, “Gender Differences in Children’s Intonational Patterns,” *Journal of Voice*, 10(3), pp. 284–291, 1996.

[11] Busby, P. A. & G. L. Plant, “Formant frequency values of vowels produced by preadolescent boys and girls,” *Journal of the Acoustical Society of America*, 97(4), pp. 2603–2606, 1994.

[12] Pettinato, M. et al., “Vowel space area in later childhood and adolescence. Effects of age, sex and ease of communication,” *Journal of Phonetics*, 54, pp. 1–14, 2016.

[13] Ferrand, C. T., “Harmonics-to-Noise Ratios in Normally Speaking Prepubescent Girls and Boys,” *Journal of Voice*, 14(1), pp. 17–21, 2000.

[14] Kauschke, C. & J. Siegmüller, *Patholinguistische Diagnostik bei Sprachentwicklungsstörungen (PDSS). Diagnostikband Grammatik*, München: Elsevier, 2010.

[15] Boersma, P. & D. Weenink, *Praat. Doing phonetics by computer. Version 6.0.40.* http://www.praat.org, 2018.

[16] Haselager G. J. T. et al., “An alternative method of studying the development of speech rate,” C*linical linguistics and phonetics*, 5(1), pp. 53–63, 1991.

Vowel movement revisited – Simple or complex metrics?

Anja Geumann[1], *Denis Arnold*[1]

[1]Institut für Deutsche Sprache, Mannheim

geumann@ids-mannheim.de, arnold@ids-mannheim.de

Abstract

We use formant slope and GAMM modeling to describe English and German non-low vowel movements. The results indicate that the dynamic metrics for non-low English tense vowels are different from English lax vowels and different from German tense vowels.

Introduction

It has been suggested by a number of recent studies, e.g. [1–4], that dynamic aspects play a role in differentiating English vowel qualities that are traditionally considered monophthongal in nature. Other studies, e.g. [5], have argued that, while there are dynamic differences, they are not so important for differentiating vowel qualities.

Metrics for describing the characteristic movement of a vowel in F1/F2 space are movement length, F1-slope, F2-slope. Movement length is usually associated with a distinction between a monophthong and a diphthong. F1-slope was suggested by Slifka [3] for distinguishing between English tense and lax non-low vowels. Chládková et al. [1] present F2-slope as a cue for front-back distinctions in English vowels. In [6], F2-slope in English vowels was shown to be similar for non-low front tense vowels and fronting diphthongs, while non-low back tense vowels have a similar F2-slope to retracting diphthongs. Here we re-examine systematically the data regarding formant slope.

In addition, we model the data using generalized additive mixed models (GAMMs) [7, 8].

In this paper we investigate non-low English tense and lax, back and front vowels and compare them to corresponding German vowels to inquire whether the dynamic behavior can be observed across languages for a certain distinctive feature opposition.

Data and Methods

We have used two well established corpora of read speech, namely the *TIMIT corpus* for American English [9] and the *Kiel Corpus of read speech* for Standard German [10], to investigate the dynamics of English and German vowels. Both corpora are provided with manually verified segmental annotations. We have examined all the data but report here only on the female speakers.

We consider place of articulation of the context of a vowel to be a major factor affecting formant transitions and potentially even the formants throughout the vowel. For that reason, we are looking here at movements in a symmetric coronal context. These coronal contexts include (aspirated) plosives, fricatives, nasals and approximants. Lateral and rhotic contexts have been excluded. The symmetric coronal context was chosen to best cover all vowels in both languages (see the occurrence frequencies in Table 1 for English (f, TIMIT) and in Table 2 for German (f, Kiel read speech). The labial_coronal context was fairly frequent in both languages as well, but in the German data, tense and lax vowels were less balanced.

V	*c_c*	*c_d*	*c_l*	*d_c*	*d_d*	*d_l*	*l_c*	*l_d*	*l_l*
i:	215	108	256	15	24	18	311	65	65
ɪ	429	169	107	71	12	29	200	107	20
e:	136	44	61	25	0	21	114	34	39
ɛ	160	51	93	45	5	7	143	44	16
u:	12	3	11	7	0	6	11	0	17
ʊ	27	9	6	42	4	1	62	4	9
o:	226	64	39	21	10	2	80	7	5
ɔ	35	19	14	24	3	7	229	13	10

Tab. 1: English 192 female speakers' vowel context frequencies, c=coronal, d=dorsal, l=labial.

V	*c_c*	*c_d*	*c_l*	*d_c*	*d_d*	*d_l*	*l_c*	*l_d*	*l_l*
i:	271	55	276	2	0	43	53	4	15
ɪ	106	545	23	30	7	9	481	35	1
e:	186	17	103	132	39	36	93	36	8
ɛ	108	34	30	69	2	0	169	2	0
u:	99	199	39	69	7	0	32	1	0
ʊ	65	100	49	16	27	9	195	11	0
o:	50	20	30	15	7	4	44	1	3
ɔ	91	130	17	39	19	56	163	31	2

Tab. 2: German 26 female speakers' vowel context frequencies, c=coronal, d=dorsal, l=labial.

Formant values were automatically extracted with Praat [11] for all the non-low phonemes /i:/, /ɪ/, /e:/, /ɛ/, /u:/, /ʊ/, /o:/, /ɔ/, Burg algorithm, 5 formants between 0-5500 Hz for female speakers and 0-5000 Hz for male speakers, time step 5 ms, window length in Praat: 25 ms, viz. a 50 ms Gaussian window, pre-emphasis from 50 Hz. Formant values were converted within Praat to Bark values.

Analysis I

The metrics we computed were:

- movement length: Euclidian distance between 25% and 75% of the vowel in the F1/F2 space.
- movement upward/downward: F1-slope between 25% and 75% in Bark.
- movement to the front/back: F2-slope between 25% and 75% in Bark.

In English and German, F1-slope did not reveal consistent significant differences between all pairs of tense and lax vowels.

For the English data only, F2-slope and movement length were significant for all vowel pairs; for the German data only F2-slope showed significant differences for all vowel pairs.

If we look at the F2-slope values for English and German directly, as in Figures 1 and 2, we find rather dissimilar patterns in the two languages.

As can be seen in Figure 1, in English we find higher F2-slopes for front tense vowels; for the back tense vowels we have lower F2-slopes. This can be interpreted as English tense non-low vowels moving outward in vowel space. The F2-slopes of German were significantly different for tense vs. lax vowel pairs. However, they do not appear to show a pattern similar to the English data, see Figure 2.

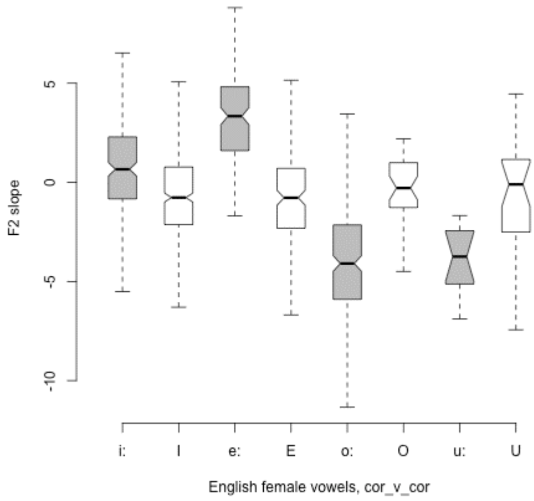


Fig. 1: English F2-slope for vowels in symmetric coronal context.

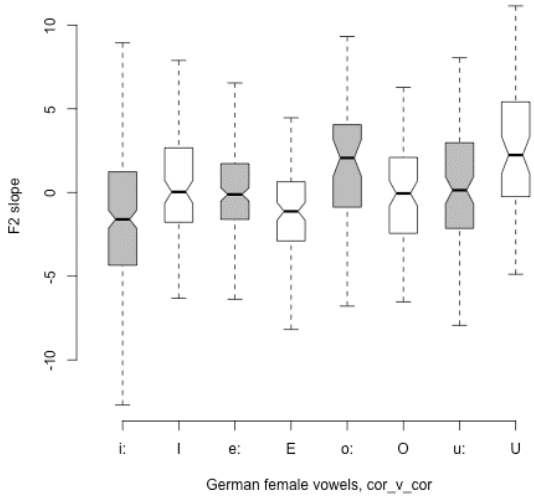


Fig. 2: German F2-slope for vowels in symmetric coronal context.

Analysis II

For modeling dynamic behavior of the dynamic formant data, we have used generalized additive mixed models (GAMMs) [7], [8] using the R-package mgcv.

Predictors
formant: formant f1, f2
relt: normalized duration from 0-1
dur: vowel duration
v: phonemes /i:/, /ɪ/, /e:/, /ɛ/, /u:/, /ʊ/, /o:/, /ɔ/
pre: prevowel context: lab, cor, dor
post: postvowel context: lab, cor, dor

Model formula

formantbark ~ te(relt, dur, by = formant)

+ s(v,relt, by = formant, bs = "re")

+ s(v, by = formant,bs = "re")

+ s(pre, relt, by = formant, bs = "re")

+ s(post, relt, by = formant, bs = "re")

+ s(v,pre, by = formant, bs = "re")

+ s(v, post, by = formant, bs = "re")

+ s(pre, by = formant, bs = "re")

+ s(post, by = formant, bs = "re")

+ formant

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.0873	0.2771	14.752	< 2e-16 ***
f2	5.8984	0.9170	6.433	1.27e-10 ***

Tab. 3: German females' parametric coefficients.

	edf	Ref.df	F	p-value
te(relt,dur):f1	11.4769	13.44	1.224e+01	< 2e-16 ***
te(relt,dur):f2	10.3690	12.21	8.908e+00	< 2e-16 ***
s(v,relt):f1	0.9342	7.00	3.769e+02	0.34219
s(v,relt):f2	6.7343	7.00	2.086e+06	8.28e-14 ***
s(v):f1	6.8930	7.00	9.363e+04	< 2e-16 ***
s(v):f2	6.8978	7.00	2.367e+07	< 2e-16 ***
s(pre,relt):f1	1.6705	2.00	4.802e+03	0.03737 *
s(pre,relt):f2	1.9675	2.00	3.117e+06	< 2e-16 ***
s(post,relt):f1	1.8333	2.00	1.783e+04	4.60e-05 ***
s(post,relt):f2	1.9681	2.00	2.052e+06	< 2e-16 ***
s(v,pre):f1	12.2372	23.00	1.586e+03	0.00302 **
s(v,pre):f2	13.9603	23.00	1.344e+05	1.08e-11 ***
s(v,post):f1	13.3927	23.00	4.076e+03	1.29e-05 ***
s(v,post):f2	14.1776	23.00	1.684e+05	4.61e-09 ***
s(pre):f1	1.3465	2.00	4.126e+03	0.13538
s(pre):f2	1.8296	2.00	3.630e+06	8.52e-06 ***
s(post):f1	0.6914	2.00	1.864e+03	0.18706
s(post):f2	1.6309	2.00	1.477e+06	0.00344 **

R-sq.(adj) = 0.963 Deviance explained = 96.3%

fREML = 46988 Scale est. = 0.51812 n = 42882

Tab. 4: German females' approximate significance of smooth terms.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.2834	0.2604	20.288	< 2e-16 ***
f2	5.4478	0.7349	7.413	1.24e-13 ***

Tab. 5: English females' parametric coefficients.

	edf	Ref.df	F	p-value
te(relt,dur):f1	1.506e+01	17.07	4.322e+01	< 2e-16 ***
te(relt,dur):f2	1.268e+01	14.52	1.941e+01	< 2e-16 ***
s(v,relt):f1	6.926e+00	7.00	7.858e+05	< 2e-16 ***
s(v,relt):f2	6.974e+00	7.00	3.320e+07	< 2e-16 ***
s(v):f1	6.802e+00	7.00	3.574e+06	5.51e-13 ***
s(v):f2	6.891e+00	7.00	5.803e+07	< 2e-16 ***
s(pre,relt):f1	1.878e+00	2.00	9.042e+04	2.62e-15 ***
s(pre,relt):f2	1.994e+00	2.00	6.339e+07	< 2e-16 ***
s(post,relt):f1	1.655e+00	2.00	1.155e+04	0.00146 **
s(post,relt):f2	1.977e+00	2.00	3.612e+05	< 2e-16 ***
s(v,pre):f1	1.481e+01	23.00	4.685e+03	0.13387
s(v,pre):f2	1.386e+01	23.00	4.496e+05	< 2e-16 ***
s(v,post):f1	1.203e+01	23.00	1.019e+03	0.09604 .
s(v,post):f2	1.535e+01	23.00	1.292e+05	7.28e-09 ***
s(pre):f1	7.750e-01	2.00	2.011e+04	0.08349 .
s(pre):f2	1.972e+00	2.00	6.462e+07	< 2e-16 ***
s(post):f1	1.756e+00	2.00	1.653e+04	0.00313 **
s(post):f2	1.733e-04	2.00	0.000e+00	0.96610

R-sq.(adj) = 0.959 Deviance explained = 95.9%

fREML = 92793 Scale est. = 0.40763 n = 95326

Tab. 6: English females' approximate significance of smooth terms.

In Figure 3 and Figure 4 we present GAMM model predictions for 9 time steps between 30% and 70% of normalized duration, and vowel duration of 90 ms, which represents tense and lax vowels fairly equally. The consonantal context of the vowel is again, as in analysis I, coronal on both sides, since this is frequent for all vowel qualities in both languages.

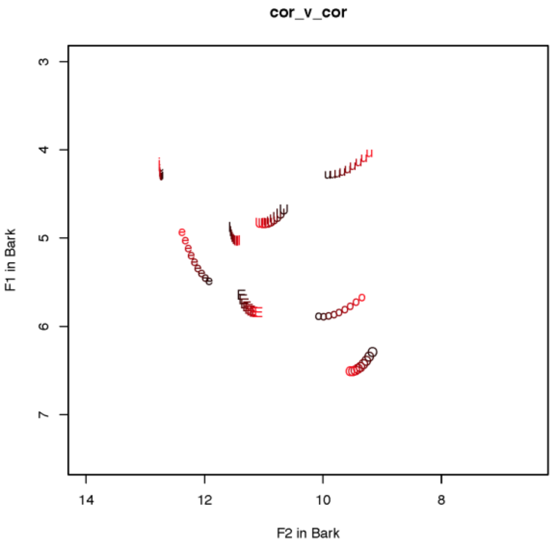


Fig. 3: GAMM model prediction for English female speakers. Vowels /i:/, /ɪ/, /e:/, /ɛ/, /o:/, /ɔ/, /u:/, /ʊ/ in coronal context before and after. Begin of the vowel in black, end of the vowel in red.

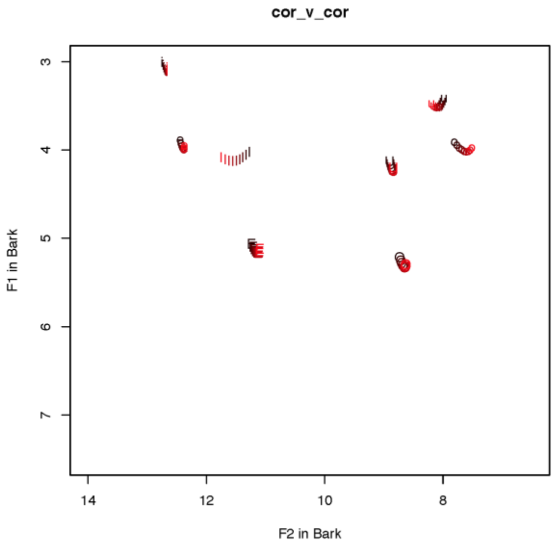


Fig. 4: GAMM model prediction for German female speakers. Vowels /i:/, /ɪ/, /e:/, /ɛ/, /o:/, /ɔ/, /u:/, /ʊ/ in coronal context before and after. Begin of the vowel in black, end of the vowel in red.

The model predictions in Figure 3 for English tense vowels show outward movements, while /ɪ/, /ɛ/, /ʊ/, /ɔ/ move in the opposite direction, towards the center of the vowel space. Model predictions for German in Figure 4 show only small movements going mostly in the same direction. The movement here is most likely just induced by coarticulatory effects of the surrounding context of each vowel.

Discussion

We have used two approaches to describe and model the dynamic behavior of English and German non-low vowels. The English F2-slopes show a pattern that discriminates between tense and lax vowels and is distinct for back and front vowels.

As already pointed out, this can be interpreted as English tense non-low vowels moving outward. This pattern is not to be found in the German data. Modeling of formant movements in our second analysis with GAMMs supports the pattern found previously, and reinforces our view that the pattern found with the English but not the German data is robust and trustworthy. We still see shortcomings in both approaches, as we have to rely on a certain amount of comparable data that is hard to come by, even in comparatively large corpora as TIMIT. We find the unequal distribution in the contexts of the vowels an interesting aspect worthy of further investigation in its own right.

References

[1] Chládková, K. et al., “F2 slope as a perceptual cue for the front-back contrast in Standard Southern British English,” *Language and Speech,* First published date: May-30-2016, available from: https://doi.org/10.1177/0023830916650991 [23.01.2019].

[2] Hillenbrand, J. M, “Static and dynamic approaches to vowel perception,” in *Vowel Inherent Spectral Change*, Morrison, G. S. & P. F. Assmann (Eds.), Berlin/Heidelberg: Springer, pp. 9–30, 2013.

[3] Slifka, J. “Tense/lax vowel classification using dynamic spectral cues,” *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, pp. 921–924, 2003.

[4] Strange, W. & O.S. Bohn, “Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies,” *JASA* 1004(1), pp. 488–504, 1998.

[5] Watson, C. I. & J. Harrington, “Acoustic evidence for dynamic formant trajectories in Australian English vowels,” *JASA*, 106(1), pp. 458–468, 1999.

[6] Raffelsiefen, R. & A. Geumann, “Diphthongs versus monophthongs in English,” *Proceedings of the conference on phonetics & phonology in German-speaking countries (P&P 13),* M. Belz et al. (Eds.), Berlin: ZAS / Humboldt-Universität zu Berlin, pp. 157–160, 2018, (https://doi.org/10.18452/18805) [23.01.2019]

[7] Wood, S., *Generalized additive models: an introduction with R*. CRC press, 2006.

[8] Wood, S.N., Goude Y. & S. Shaw, “Generalized additive models for large data sets,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1), pp. 139–155, 2015.

[9] Garofolo, J. et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[10] Kohler, K. J. (ed.), P*honetisch-Akustische Datenbasis des Hochdeutschen. Kieler Arbeiten zu den PHONDAT-Projekten 1989–1992*, 1992 (Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 26).

[11] Boersma, P. & D. Weenink, Pr*aat: doing phonetics by computer* [Computer program]. Version 6.0.17, retrieved 21 April 2016 from http://www.praat.org/ [04.02.2019].

Normalisierung von Vokalen im DaF-Kontext: Vergleich von Normalisierungsmethoden

Christoph Gube[1]

[1]Humboldt-Universität zu Berlin

christoph.gube@gmail.com

Abstract

*Für den DaF-Unterricht wäre es wünschenswert, Artikulationsstellen von Vokalen zwischen deutschen L1-Sprecher*innen und DaF-Lerner*innen direkt miteinander vergleichen zu können. In dieser Arbeit vergleiche ich deshalb anhand von Audio-Daten von Sprecher*innen mit L1 Deutsch und L1 Chinesisch 16 Methoden der Normalisierung von Vokal-Formanten miteinander, wobei eine möglichst hohe Übereinstimmung (Schnittmenge) der individuellen Vokalräume der Sprecherinnen erzielt werden soll. Ich komme zu dem Ergebnis, dass die erzielbare Schnittmenge der individuellen Vokalräume nicht nur von der verwendeten Normalisierungsmethode, sondern auch von der betrachteten Sprache abhängt.*

Einleitung

DaF-Lerner*innen mit L1 Chinesisch haben oft Schwierigkeiten mit dem deutschen Vokalsystem. Besonders das Konzept der Lang- und Kurzvokale und die damit einhergehende Änderung der Vokalqualität stellt eine häufige Fehlerquelle dar [1: 87]. Für den DaF-Unterricht wäre es daher wünschenswert, Mundöffnung und Zungenposition deutscher L1-Sprecher*innen und chinesischer Lerner*innen direkt vergleichen zu können, um die Produktion der Lerner*innen gezielter korrigieren zu können.

'Rohe' Formantdaten in Hz sind jedoch für einen direkten Vergleich zwischen verschiedenen Sprecher*innen nicht geeignet, weil diese Werte mit der individuellen Anatomie des Vokaltrakts stark variieren [2: 159; 3: 1f.]. Es ist deshalb für Vergleiche zwischen Sprecher*innen nötig, die Formantdaten zu normalisieren.

Am Beispiel von chinesischen und deutschen Audio-Daten aus Produktionsexperimenten vergleiche ich 16 verschiedene Normalisierungsmethoden mit dem Ziel, ihre Eignung für den direkten Vergleich von Artikulationsstellen zwischen Sprecher*innen im DaF-Kontext zu beurteilen.

Daten

Für diese Untersuchung wurden Audio-Daten von 12 Sprecher*innen mit L1 Deutsch und 10 Sprecher*innen mit L1 Chinesisch verwendet. Die chinesischen Sprecherinnen studierten zum Zeitpunkt des Produktionsexperiments den Studiengang Germanistik im dritten Studienjahr am *Science and Technology College der Universität Nanchang* (Jiangxi, VR China) und waren zwischen 20 und 23 Jahren alt. Die deutschen Sprecher*innen waren zum Zeitpunkt der Datenerhebung zwischen 20 und 34 Jahren alt.

Von beiden Gruppen von Sprecher*innen wurden alle Monophthonge des Deutschen elizitiert, von den chinesischen Sprecher*innen zusätzlich dazu insgesamt 27 Vokale aller 9 Vokalkategorien des Chinesischen nach [1] in verschiedenen Tönen. Die Untersuchung beschränkte sich später auf die chinesischen Monophthonge /i, y, u, ə, ɤ, a/ [4: 110].

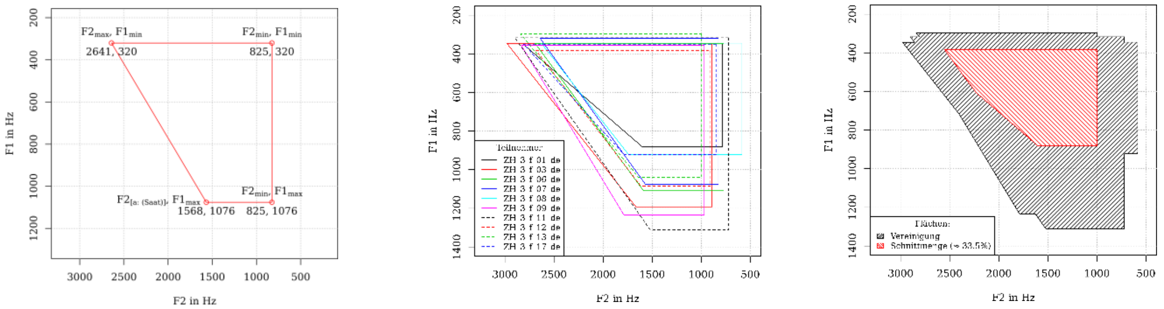
Normalisierung

Ausgehend von den Daten in Hz wurde für jede Sprecherin der theoretische maximale Vokalraum nach [3: 14] konstruiert (Abb. 1a). Die Vokalräume der Sprecher*innen wurden für jede Gruppe und Sprache separat übereinandergelegt (Abb. 1b) und der Prozentsatz der Schnittmenge von der Vereinigung der Flächen errechnet (Abb. 1c). Tab.1 zeigt alle Schnittmengen:

Gruppe	S*
Deutsche Sprecher*innen (Deutsch)	47,0%
Chinesische Sprecher*innen (Chinesisch)	32,6%
Chinesische Sprecher*innen (Deutsch)	33,5%

* S = Schnittmenge in % der Vereinigung

Tab. 1: Schnittmengen nach Sprache.



(a) Vokalraum nach [3]. **(b)** Vokalräume (ZH-de). **(c)** Vereinigung & Schnittmenge.

*Abb.1: Vokalräume (Hz) von ZH-3-f-07 und allen chinesischen Sprecher*innen (Deutsch).*

Grob lassen sich in dieser Arbeit drei Arten von Normalisierungsmethoden unterscheiden:

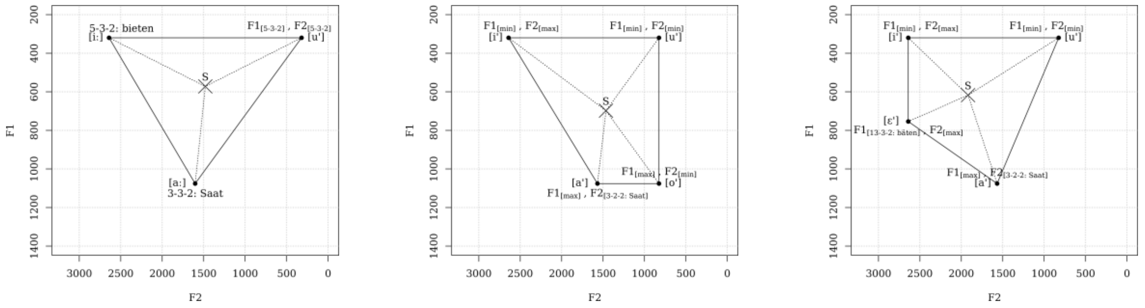
[1] Transformationen überführen die Formantdaten von Hz in andere Skalen und sind formant- und vokal-intrinsisch (*Bark*, *Mel*, *ERB*, *Log10*, *Logn*), [2] Skalierungen berücksichtigen alle Vokale einer Sprecherin (vokal-extrinsisch) und skalieren die Formantdaten an ihren Extrem- und/oder Mittelwerten (*LCE*, *Gerstman*, *Lobanov*, *Nearey* sind formant-intrinsisch, *NeareyGM* ist formant-extrinsisch) und [3] geometrische Methoden skalieren anhand eines geometrisch konstruierten Vokalraums (Beispiele in Abb. 2), in dieser Arbeit sind dies: *WF1-3* [2], *Bigham* und *LettER* (formant-intrinsisch, vokal-extrinsisch).

Geometrische Methoden sind in ihrer Implementierung leicht an die charakteristischen Vokalräume unterschiedlicher Sprachen anpassbar: *WF2* legt einen dreieckigen Vokalraum zugrunde und normalisiert relativ zu dessen Mittelpunkt *S* (siehe Abb. 2a; [5: 420 f.]). [6: 134-136] hingegen konstruiert für seine Untersuchung einen viereckigen Vokalraum mit Mittelpunkt *S* (Abb. 2b), um den Vokalraum des von ihm untersuchten amerikanischen Englisch besser zu repräsentieren. In dieser Studie erzielten Experimente mit der *Bigham*-Modifikation *Bigham-ε* (Abb. 2c) höhere Übereinstimmungen der chinesischen Formantdaten als *Bigham*.

Ergebnisse

Die Resultate aller 16 untersuchten Normalisierungsmethoden sind in Tabelle 1 dargestellt, sortiert in absteigender Rangfolge nach erzielter Übereinstimmung der individuellen Vokalräume der Sprecher*innen. Alle Methoden unter der Zeile für Hz ergaben keine Verbesserung der Schnittmengen gegenüber den ‘rohen’ Formantdaten in Hz und führten im Gegenteil zu Fehlausrichtungen der Vokalräume, entweder durch Parallelversatz der Flächen oder durch Vergrößerung der sich nicht überschneidenden Bereiche (z.B. *Bark*, Abb. 3a).

Gerstman steht in allen drei Gruppen an der Spitze der Tabelle, weil es durch seine Implementierung zwangsläufig immer fast kongruente Vokalräume liefert, die nur in der horizontalen (F_2) Position von /a/ variieren (Abb. 3b) und dadurch durchgängig sehr hohe Übereinstimmungen der individuellen Vokalräume der Sprecher*innen erzielt. Nach *Gerstman* folgen in allen drei Gruppen geometrische Methoden.



(a) *WF2*. **(b)** *Bigham*. **(c)** *Bigham-ε*.

Abb. 2: Vokalräume (Deutsch) in geometrischen Methoden, Probandin `ZH-3-f-07`.

Deutsche Sprecherinnen (Deutsch)				Chinesische Sprecherinnen (Chinesisch)				Chinesische Sprecherinnen (Deutsch)			
	Methode	V*	S**		Methode	V*	S**		Methode	V*	S**
1	*Gerstman*	E	86,8	1	*Gerstman*	E	87.1	1	*Gerstman*	E	86,2
2	*WF2*	E	63,0	2	*Bigham-ɛ*	E	61.3	2	*Bigham*	E	49,0
3	*Bigham*	E	60,9	3	*WF2*	E	60.4	3	*WF2*	E	48,7
4	*WF3*	E	59,7	4	*LCE*	E	58.2	4	*WF3*	E	48,0
5	*LCE*	E	58,4	5	*Bigham*	E	57.9	5	*LCE*	E	46,3
6	*Bigham-ɛ*	E	52,1	6	*WF3*	E	57.1	6	*Bigham-ɛ*	E	45,0
7	*WF1*	E	51,3	7	*WF1*	E	56.3	7	*Lobanov*	E	41,4
8	*Hz*	—	47,0	8	*Lobanov*	E	49.9	8	*WF1*	E	38,5
9	*Bark*	I	46,5	9	*Nearey*	E	46.7	9	*Hz*	—	33,5
	Mel	I	46,5	10	*NeareyGM*	E	43.3	10	*Mel*	I	32,8
10	*ERB*	I	45,3	11	*Bark*	I	33.8	11	*Bark*	I	32,6
11	*NeareyGM*	E	43,7	12	*Mel*	I	33.5	12	*NeareyGM*	E	32,2
12	*Log10*	I	43,5	13	*LettER*	E	33.3	13	*ERB*	I	32,0
	Logn	I	43,5	14	*ERB*	I	32.9	14	*Nearey*	E	31,3
13	*Lobanov*	E	43,0	15	*Hz*	—	32.6	15	*Log10*	I	30,9
14	*Nearey*	E	42,2	16	*Log10*	I	31.7		*Logn*	I	30,9
15	*LettER*	E	32,7		*Logn*	I	31.7	16	*LettER*	E	28,2

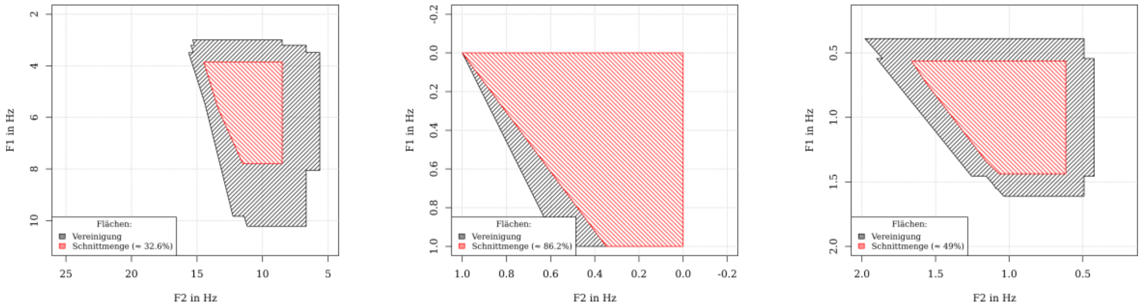
* V = Vokal-intrinsisch (I) bzw. -extrinsisch (E), ** S = Schnittmenge in % der Vereinigung

Tab. 1: Rangfolge nach Schnittmenge S.

Es sollte jedoch beachtet werden, dass einzelne Methoden je nach Sprache unterschiedlich performant sein können. So steht z.B. *Bigham* für die deutschen Audio-Daten der chinesischen Sprecher*innen an zweiter Stelle in der Tabelle, für die chinesischen Daten jedoch an fünfter Stelle. Statt dessen erzielte die Modifikation *Bigham-ɛ*, welche den Eckpunkt [o'] gegen [ɛ'] tauscht (Abb. 2c), hier eine höhere Schnittmenge der Vokalräume; möglicherweise wird der Vokalraum des Chinesischen durch die Form des Vierecks in *Bigham-ɛ* besser repräsentiert. Außerdem ist auffällig, dass trotz annähernd gleicher Überschneidungswerte in Hz für die chinesischen Sprecher*innen in beiden Sprachen die chinesischen Daten durch jede Normalisierungsmethode außer den logarithmischen Skalierungen *Log10* und *Logn* eine Verbesserung erfahren, während für die deutschen Daten der deutschen und chinesischen Sprecher*innen die 'rohen' Hz-Daten in der Rangliste viel weiter oben angesiedelt sind und wesentlich mehr Methoden zu einer Verschlechterung der Übereinstimmungswerte führen.

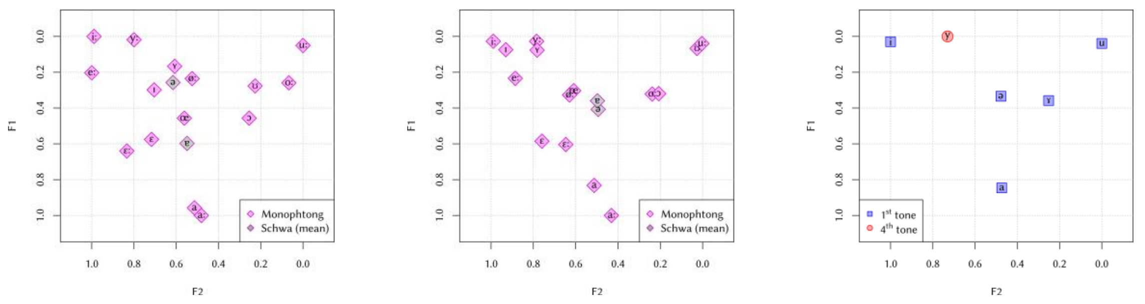
Diskussion

Basierend auf den Ergebnissen aus Tab. 1 sehe ich *Gerstman* als am besten geeignet, um den eingangs erwähnten Vergleich von Artikulationsstellen individueller Vokale zwischen Sprecher*innen im DaF-Kontext durchzuführen. *Gerstman* ist einfach zu implementieren und



(a) *Bark.* **(b)** *Gerstman.* **(c)** *Bigham.*

Abb. 3: Vereinigungen und Schnittmengen der Vokalräume für je eine transformierende (a), skalierende (b) und geometrische (c) Methode, chinesische Sprecherinnen (Deutsch).



(a) DE-2-f-11. **(b)** ZH-3-f-07 (Deutsch). **(c)** ZH-3-f-07 (Chinesisch).

*Abb. 4: Vergleich zwischen Proband*innen und Sprachen (Gerstman).*

liefert fast kongruente Vokalräume, welche sich gut für einen direkten Vergleich eignen, wobei auch *Bigham* und *WF2* für die vorliegenden Daten gute Ergebnisse erzielen.

Abb. 4 zeigt exemplarisch den Vergleich von Vokalpositionen einer Sprecherin mit L1 Deutsch und einer Sprecherin mit L1 Chinesisch. Gut zu erkennen ist bei ZH-3-f-07 (Abb. 4b) im Vergleich zu DE-2-f-11 (Abb. 4a) das Clustering der Lang- und Kurzvokale /iː - ɪ, yː - ʏ, uː - ʊ, oː - ɔ, øː - œ/, die leicht abweichende Position von /eː/ sowie die fehlende Differenzierung zwischen /ə/ und [ɐ], welche sich beide an der Position des chinesischen /ə/ (Abb. 4c) der Sprecherin ZH-3-f-07 befinden.

Abschließend sollte erwähnt werden, dass je nach Ziel der Untersuchung individuelle Normalisierungsmethoden ungeeignet sein können, auch wenn sie hohe Übereinstimmungen der Vokalräume erzielen. So verwendet z.B. *Gerstman* für die meisten Sprachen die Vokale /i, u, a/ (höchste/niedrigste F_1/F_2-Werte) als Eckpunkte für die Normalisierung [3: 5]. Dies führt dazu, dass diese Vokale feste Koordinaten (im Falle von /a/ eine feste F_1-Position) in der Formantkarte erhalten und Variationen zwischen einzelnen Sprecher*innen bei diesen Vokalen u.U. nicht mehr erkennbar sein können. Sind die relativen Positionen dieser Vokale von besonderem Interesse, sollte auf andere Methoden ausgewichen und ggf. die Vergleichsprozedur angepasst werden; für die Daten in dieser Untersuchung wären z.B. *Bigham* oder *WF2* ebenfalls gut geeignet.

Referenzen

[1] Hunold, C., *Untersuchungen zu segmentalen und suprasegmentalen Ausspracheabweichungen chinesischer Deutschlernender*, Frankfurt am Main et al: Peter Lang, 2009.

[2] Watt, D. & A. H. Fabricius, “Evaluation of a technique for improving the mapping of multiple speakers’ vowel spaces in the F1~F2 plane,” *Leeds Working Papers in Linguistics and Phonetics*, 9 (9), pp. 159–173, 2002.

[3] Flynn, N., “Comparing vowel formant normalisation procedures,” *York papers in linguistics*, 2 (11), pp. 1–28, 2011.

[4] Lee, W.-S. & E. Zee, “Standard Chinese (Beijing),” *Journal of the International Phonetic Association*, 33 (1), pp. 109–112, 2003.

[5] Fabricius, A. H. et al., “A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics,” *Language Variation and Change*, 21 (3), pp. 413–435, 2009.

[6] Bigham, D. S., *Dialect Contact and Accommodation Among Emerging Adults in a University Setting*. Dissertation, University of Texas at Austin, 2008 (https://repositories.lib.utexas.edu/handle/2152/17759) [06.02.2020].

Kategoriale Wahrnehmung von *Voice Onset Time* durch österreichische Hörer*innen

Petra Hödl[1]

[1]Institut für Sprachwissenschaft (Universität Graz)

petra.hoedl@uni-graz.at

Abstract

*Die hier präsentierte Studie beschäftigt sich mit der Frage, ob die Tendenz zur reduzierten Aspiration durch österreichische Sprecher*innen Auswirkungen auf deren Perzeption der Stimmhaftigkeitsopposition zwischen /b, d, g/ und /p, t, k/ hat. Mittels Identifikationstest wurden für österreichische und deutsche Hörer*innen die Kategoriengrenzen zwischen 'stimmhaft' und 'stimmlos' innerhalb von VOT-Kontinua erhoben sowie die Form ihrer psychometrischen Kurven modelliert. Die Ergebnisse deuten darauf hin, dass VOT nicht das einzige akustische Merkmal ist, das für die Perzeption der Stimmhaftigkeit im Deutschen relevant ist. Zudem wurden individuelle Unterschiede hinsichtlich der Kategorialität der Perzeption festgestellt.*

Einleitung

Das Standarddeutsche weist eine phonologische Opposition zwischen stimmhaften (/b, d, g/) und stimmlosen (/p, t, k/) Plosiven auf. Diese phonologische Opposition wird allerdings insbesondere im Anlaut meist als Aspirationskontrast und nicht als 'echter' Stimmtonkontrast realisiert [1], [2]. Somit haben wir es phonetisch gesehen in beiden Fällen mit physiologisch stimmlosen Lauten zu tun, die durch unterschiedlich lange positive *Voice Onset Times* (VOT) [3] gekennzeichnet sind. Diese phonetische Unterscheidung ist für die Aussprache norddeutscher Sprecher*innen empirisch bestätigt worden [1]. Bei Sprecher*innen einiger (südlicher) Varietäten des Deutschen – wie auch des österreichischen Deutsch – kommt es laut Literatur jedoch häufig zu einer Reduzierung der Aspiration [2–5]. Dies wird insbesondere für Bilabiale und Alveolare angenommen.

Es stellt sich nun die Frage, ob diese Tendenz zur reduzierten Aspiration Auswirkungen auf die Wahrnehmung des phonetischen Parameters VOT durch österreichische Hörer*innen hat. Wenn man aufgrund der Tendenz zur Deaspirierung die VOT als unzuverlässiges Unterscheidungsmerkmal bei produzierten österreichischen Plosiven ansieht, dann könnte sich dies auch in der Perzeption dieses akustischen Parameters widerspiegeln.

Im Speziellen soll die Position und der Schärfegrad der Kategoriengrenzen zwischen 'stimmhaft' und 'stimmlos' bei österreichischen Hörer*innen sowie einer Kontrollgruppe deutscher Hörer*innen erhoben werden.

Methode

Es wurde ein 2AFC-Identifikationstest durchgeführt, bei dem 33 österreichische Hörer*innen als Testgruppe und 47 deutsche Hörer*innen als Kontrollgruppe semi-manipulierte Wortäußerungen einer österreichischen Sprecherin identifizieren mussten.

Sowohl die österreichischen Proband*innen als auch die Sprecherin des Testmaterials stammten aus der Steiermark. Die deutschen Proband*innen waren in Graz wohnhafte Personen, die in unterschiedlichsten Teilen Deutschlands aufgewachsen sind. Sie werden in weiterer Folge sehr grob und terminologisch nicht exakt in 'Süddeutsche' und 'Norddeutsche' untergliedert (die Gliederung orientiert sich an der Speyerer Linie und der Grenze zwischen oberdeutschem und mitteldeutschem Dialektgebiet).

Als Stimuli wurden die Wörter *backen*, *packen*, *danken*, *tanken*, *Gasse* und *Kasse* verwendet, wobei die VOT der Äußerungen in Stufen von 5 Millisekunden akustisch manipuliert wurde. Die sechs so entstandenen Kontinua bestanden aus jeweils 13 Stimuli, die VOT-Werte von 0 bis 60 Millisekunden aufwiesen. Sonstige akustische Manipulationen der Stimuli (beispielsweise eine Angleichung der Aspirationsintensität zwischen den einzelnen Artikulationsorten) wurden vermieden, um ein möglichst natürlich klingendes Testmaterial zu gewährleisten. Das einzige akustische Merkmal, das neben der VOT noch manipuliert wurde, war die Dauer des Lösungsgeräusches, die auf 5 Millisekunden nivelliert wurde.

Insgesamt mussten die Teilnehmer*innen während des Experiments 468 manipulierte Stimuli (13 VOT-Schritte * 6 Kontinua * 6 Wiederholungen) identifizieren. Die Durchführung des Experiments sowie alle akustischen Manipulationen erfolgten in Praat [6].

Auswertung

Für jede Versuchsperson wurden auf Basis ihrer Antworten im Identifikationstest psychometrische Kurven erstellt. Dabei wurde die Wahrscheinlichkeit einer ‘stimmlos’-Antwort als Funktion der VOT-Dauer des Stimulus modelliert. Diese Modellierung und die Auswertung der Ergebnisse erfolgte mittels binomialer Regressionsanalyse separat für jedes Kontinuum und jede Versuchsperson. Als Kategoriengrenze wurde der 50%-Schwellenwert gewählt, d.h. jener VOT-Wert, bei dem die Wahrscheinlichkeit einer ‘stimmlos’-Antwort bei 50% liegt (Abb.1). Wenn der 50%-Schwellenwert innerhalb der präsentieren VOT-Werte von einer Versuchsperson nicht erreicht wurde, wurden ihre Daten für das betroffene Kontinuum nicht berücksichtigt.

Die Auswertung der Identifikationsergebnisse erfolgte in R mithilfe des Packages quickpsy [7].

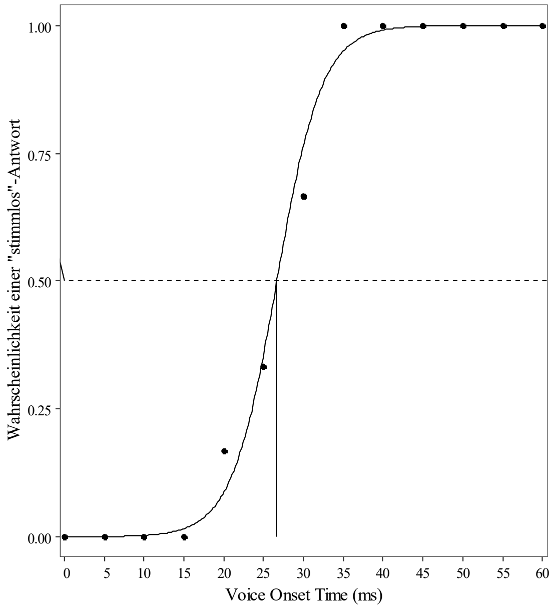


Abb.1: Beispiel für eine psychometrische Kurve. Die Wahrscheinlichkeit einer ‘stimmlos’-Antwort wird als Funktion der VOT-Dauer des präsentierten Stimulus dargestellt.

Die prognostizierten Kategoriengrenzen dienten in weiterer Folge als abhängige Variable in sechs einfaktoriellen ANOVAs (separat durchgeführt für jedes der sechs Kontinua) mit dem Zwischensubjektfaktor Hörer*innengruppe (mit den drei Ausprägungen Österreicher, Süddeutsche und Norddeutsche).

Ergebnisse

Für fünf der sechs getesteten Kontinua konnte keine statistisch signifikant verschobene Kategoriengrenze bei den österreichischen Hörer*innen im Vergleich zu den deutschen Hörer*innen festgestellt werden. Dies deutet darauf hin, dass die VOT für österreichische Hörer*innen im Allgemeinen die gleiche Dauer aufweisen muss wie für deutsche Hörer*innen, um ein stimmloses Perzept auszulösen.

Die einzige Ausnahme stellte das *Kasse*-Kontinuum dar, bei dem die österreichischen Hörer*innen im Durchschnitt eine längere VOT benötigten, um ‘stimmlos’ zu antworten, als die süddeutschen Hörer*innen (21 ms vs. 15,7 ms, p < 0,001) und als die norddeutschen Hörer*innen (21 ms vs. 17,1 ms, p=0,040).

Dass ein Gruppenunterschied beim *Kasse*-Kontinuum (und nur bei diesem!) festgestellt wurde, ist etwas überraschend. Schließlich werden wortinitale prävokalische Velarplosive in der Umgebungssprache der getesteten österreichischen Versuchspersonen üblicherweise nicht deaspiriert [5] und würden sich demnach nicht von der in Deutschland üblichen Aussprache unterscheiden. Allerdings wurde für österreichische Velarplosive eine Tendenz zur Affrizierung attestiert – wenn auch insbesondere vor Vorderzungenvokalen [2]. Es könnte also sein, dass österreichische Hörer*innen generell eine *längere* VOT bei velaren *fortis*-Plosiven gewohnt sind als deutsche Hörer*innen und deshalb bei diesem Kontinuum eine etwas nach rechts verschobene Kategoriengrenze zeigen, d.h. Stimuli erst bei einer längeren VOT als ‘stimmlos’ wahrnehmen.

Da für die anderen Kontinua keine statistisch signifikanten Gruppenunterschiede festgestellt werden konnten, werden die Ergebnisse in weitere Folge über alle drei Hörer*innengruppen zusammengefasst. Dabei wurden jene elf Proband*innen, die für eines der Kontinua keinen 50%-Schwellenwert erreichten, in der Analyse nicht mehr weiter berücksichtigt.

In weiterer Folge wurde eine 3x2 ANOVA für abhängige Stichproben durchgeführt mit den Faktoren Artikulationsort (bilabial, alveolar, velar) und Burstidentität (*lenis*, *fortis*).

Über alle drei Hörer*innengruppen hinweg ergab sich ein signifikanter Effekt des Artikulationsortes auf die Kategoriengrenze (F(2,136) =

56,47, p < 0,001). Bei velaren Stimuli ist die Kategoriengrenze nach rechts verschoben im Vergleich zu den bilabialen und alveolaren. Das bedeutet, dass eine längere VOT notwendig ist, um einen velaren Plosiv als stimmlos wahrzunehmen. Dies stimmt mit den Gegebenheiten in der Produktion überein. Auch hier weisen Velare üblicherweise eine längere VOT auf als Bilabiale oder Alveolare [8].

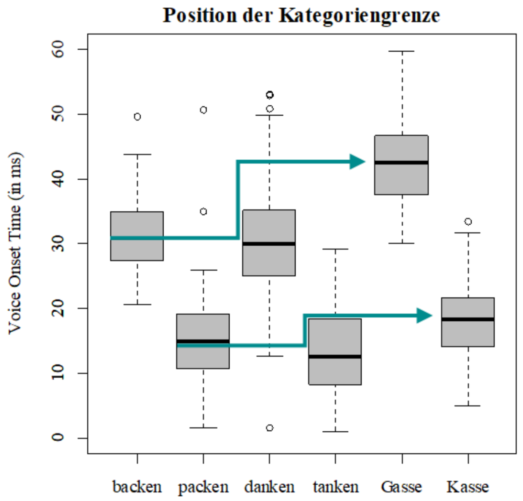


Abb.2: Einfluss des Artikulationsortes auf die Position der Kategoriengrenze.

Des Weiteren wurden *fortis*-Stimuli (d.h. jene Stimuli, die von Wörtern mit ursprünglichen initialen *fortis*-Plosiven generiert wurden) bei einer kürzeren VOT als stimmlos identifiziert als *lenis*-Stimuli (F(1, 68) = 1536,872, p < 0,001).

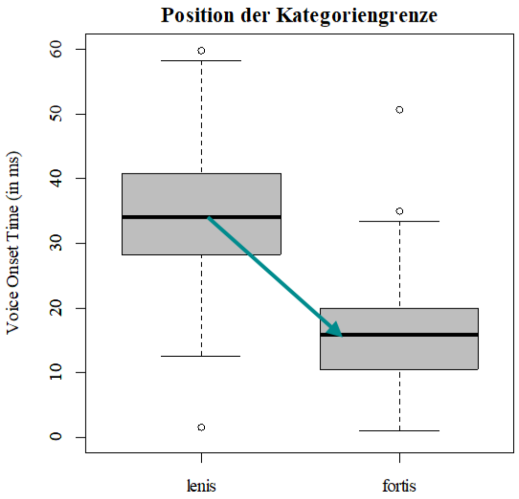


Abb.3: Einfluss von lenis/fortis *auf die Position der Kategoriengrenze.*

Bereits bei VOT-Werten weit unter 20 Millisekunden wurden diese Stimuli (die ein *fortis*-Lösungsgeräusch und einen Folgevokal aufweisen, der koartikulatorisch dem nach einem *fortis*-Konsonanten entspricht) als 'stimmlos' perzipiert. Dieses Resultat deutet darauf hin, dass VOT nicht der einzige relevante akustisch-auditive Parameter bei der perzeptuellen Unterscheidung von /b, d, g/ vs. /p, t, k/ für deutschsprachige Hörer*innen ist.

Was das Muster der Identifikationsleistungen betrifft, kann dieses bei den meisten der getesteten Versuchspersonen eindeutig als kategorial bezeichnet werden. Die kontinuierlich zunehmende VOT-Dauer löst nicht eine gleichermaßen kontinuierliche Zunahme an 'stimmlos'-Antworten aus, sondern die Perzeption 'kippt' an der Kategoriengrenze plötzlich von 'stimmhaft' zu 'stimmlos' 'um'. Die modellierten Identifikationskurven weisen daher in der Regel eine charakteristische sigmoidale Form auf (Abb.1).

Es konnten hier allerdings starke individuelle Unterschiede festgestellt werden. Wie bereits erwähnt, gab es einige wenige Teilnehmer*innen, die im präsentierten VOT-Bereich den 50%-Schwellenwert nicht erreichten. Auch gab es Teilnehmer*innen, die den 25%- und/oder 75%-Schwellenwert nicht erreichten. D.h. diese Hörer*innen perzipierten keinen der präsentierten Stimuli als eindeutig einer der beiden Stimmtonkategorien zugehörig.

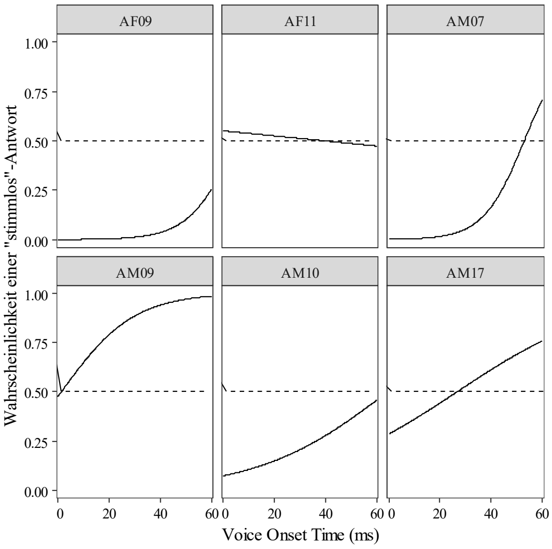


*Abb.4: Identifikationskurven jener sechs Proband*innen, die für die* danken-*Stimuli ein atypisches Identifikationsmuster zeigten.*

Interessanterweise gab es in der österreichischen Hörer*innengruppe tendenziell mehr Personen mit atypischen Identifikationskurven als in den beiden deutschen Hörer*innengruppen. So zeigten beispielsweise für das *backen*-Kontinuum und das *danken*-Kontinuum ausschließlich österreichische Versuchspersonen nicht-kategoriale Identifikationskurven (vier Hörer*innen für *backen* und sechs Hörer*innen für *danken*, vgl. Abb.4 für letztere). Beim *Gasse*-Kontinuum war es ausgeglichen mit zwei Österreichern, zwei Norddeutschen und drei Süddeutschen.

Fazit

Im Allgemeinen scheint die reduzierte Aspiration in der Aussprache keine Auswirkungen auf die Identifikation von VOT bei österreichischen Hörer*innen zu haben. Die Position der Kategoriengrenzen in den präsentierten Kontinua unterschied sich nicht signifikant von jener deutscher Hörer*innen. Die einzige Ausnahme stellte das *Kasse*-Kontinuum dar, für das die österreichischen Hörer*innen eine nach rechts verschobene Kategoriengrenze aufwiesen. Ein eventuell möglicher Erklärungsansatz hierfür wäre die Tendenz zur Affrizierung von Velarplosiven bei österreichischen Sprecher*innen [2].

Neben dem Artikulationsort wirkten sich auch die Charakteristika des Lösungsgeräuschs (und/oder des Folgevokals) statistisch signifikant auf die Kategoriengrenze aus. Wurde ein Stimulus von einem Wort mit einem initialen *fortis*-Plosiv generiert (d.h. von *packen*, *tanken* oder *Kasse*), wurde dieser bereits bei einer viel kürzeren VOT-Dauer als ‘stimmlos’ identifiziert als Stimuli, die von *backen*, *danken* oder *Gasse* generiert wurden. Dies legt den Schluss nahe, dass VOT nicht der einzige relevante Faktor bei der Perzeption der Stimmhaftigkeitsopposition im Deutschen ist.

Was die Kategorialität der Perzeption von VOT betrifft, konnten individuelle Unterschiede festgestellt werden. Zwar zeigten die meisten Versuchspersonen eine kategoriale Wahrnehmung der präsentierten Stimuli, doch einige wenige Hörer*innen wiesen atypische Identifikationskurven auf (beim *backen*- und *danken*-Kontinuum waren dies interessanterweise ausschließlich österreichische Hörer*innen).

Da nicht-kategoriale Identifikationsmuster bei Konsonanten von einigen Autoren als Indiz für eine reduzierte Fähigkeit zur phonemischen Kategorisierung gewertet werden ([9-10]), könnte dies darauf hindeuten, dass im phonologischen System dieser Hörer*innen die Unterscheidung zwischen ‘stimmlos’ und ‘stimmhaft’ auf Basis der VOT nicht so fest verankert ist wie bei den restlichen Proband*innen.

Referenzen

[1] Jessen, M., *Phonetics and phonology of tense and lax obstruents in German*, Amsterdam/Philadelphia: Benjamins, 1998.

[2] Moosmüller, S. & C. Ringen, “Voice and aspiration in Austrian German plosives,” *Folia Linguistica*, 38, pp. 43–62, 2004.

[3] Lisker, L. & A. S. Abramson, “A cross-language study of voicing in initial stops: acoustical measurements,” *Word*, 20, pp. 384–422, 1964.

[4] Muhr, R., “Varietäten des Österreichischen Deutsch,” *Revue belge de philologie et d’histoire*, 79, pp. 779–803, 2001.

[5] Krech, E.-M. et al. (Eds.), *Deutsches Aussprachewörterbuch*, Berlin/New York: de Gruyter, 2009.

[6] Boersma, P. & D. Weenink, *Praat: Doing phonetics by computer*, 2017. Version 6.0.28, http://www.praat.org/.

[7] Linares, D. & J. López-Moliner, “quickpsy: An R package to fit psychometric functions for multiple groups,” *The R Journal*, 8(1), pp. 122–131, 2016.

[8] Fischer-Jørgensen, E., “Acoustic analysis of stop consonants,” *Miscellanea Phonetica*, 2, pp. 42–59, 1954.

[9] Hazan, V. et al., “Speech pattern audiometry for clinical use,” *European Journal of Disorders of Communication*, 30, pp. 116–123, 1995.

[10] Joanisse, M. F. et al., “Language deficits in dyslexic children: Speech perception, phonology, and morphology,” *Journal of Experimental Child Psychology*, 77(1), pp. 30–60, 2000.

Annotation of Haitian Creole prosody and intonation

Alexander Teixeira Kalkhoff[1]

[1]Albert-Ludwigs-Universität Freiburg i.Br.

alexander@teixeirakalkhoff.eu

Abstract

The present paper suggests and comments on an annotation hierarchy with sets of labels for analysing prosodic features and drawing first generalisations about the intonational phonology of Haitian Creole. For phonetic test settings, I adapted the IVTS (Intonational Variation Transcription System) model and annotation system. Besides the collection of more detailed information about the Haitian Creole prosodic system, the core issue of my approach is to correlate speech acts, identified and labelled by the linguist, with local and global intonational event. The goal is to collect (recurrent) intonational patterns that could serve as a starting point for a corpus-based and data-driven intonational phonology of Haitian Creole.

Introduction

The description of Haitian Creole (henceforth HC) prosody and intonation for the most part still remains a desideratum [1]. By starting prosodic and intonational research from scratch, the methodological question arises, how to obtain first observations and generalisations about the prosodic system, i.e. suprasegmentals, and intonational phonology, i.e. the link between illocution and recurrent nuclear pitch accent patterns, from speech data within experimental phonetic settings. Unfortunately, there is a considerable methodological gap within prosodic and intonational research because, due to its strong phonological claim, the generally-accepted and well-established ToBI (Tones and Break Indices) annotation system [2] appears to be inappropriate for the test stage. ToBI-annotated data sets represent the final stage of the analytic process and are readable only in light of the intonational phonology of a given language.

To fill this gap and to annotate intonational variation of still unknown phonological systems of English varieties, the IViE (Intonation Variation in English) annotation system was developed at Cambridge University under the auspices of Francis Nolan [3]. Brechtje Post and Elisabeth Delais-Roussarie expanded the language-specific IViE model to the language-non-specific IVTS (Intonational Variation Transcription System) [4]. The IVTS model encodes orthographic, prosodic, and intonational information on six annotation levels, i.e. a comment, a phonological (or tonal), a global phonetic, a local phonetic, a rhythmic (or prominence), and an orthographic tier. A starting point for the annotation process is the identification of prominent syllables and a narrow phonetic annotation of local intonational events. Intonation at the discourse level is annotated phonetically on the global phonetic tier. First phonological assumptions can also be annotated on the phonological tier. The IViE/IVTS set of labels is transparent, easy to manage, and re-uses well-established ToBI symbols, such as L, H, and % [5].

Currently, there are only a few studies on HC prosody and literally nothing on HC intonation. Anne-Marie Brousseau's [6–7] and my own research [8] on HC prosody has found, according to the hybrid system hypothesis, i.e. the HC prosodic system has been shaped by both West African substrate languages and the French superstrate. First insights include: (i) HC has a stress accent system and (ii) word accent, (iii) the phonological word is the domain of accentuation, (iv) HC uses grammatical and pragmatic tones, and (v) in some prosodic contexts it lengthens penultimate syllables (penultimate lengthening). Dominique Fattier [9] describes the system of HC interrogative morphemes, however, without giving further explanation on the shape of HC interrogation intonational patterns.

In view of the still largely unknown HC prosodic and intonational system, I adopted the IVTS model and modified the original IVTS annotation hierarchy. I added six annotation tiers: three to collect more prosodic information about HC penultimate lengthening, syllable structure, and prosodic constituency; two to correlate pragmatics and information structure with local and global intonational events; and one to correlate specific word classes with tonal movements. The annotation hierarchy can easily be implemented in Praat [10].

The annotation hierarchy and its sets of annotation labels are tested on the basis of an annotated HC sample taken from the APiCS (*The Atlas of Pidgin and Creole Language Structures*) online [11]. My contribution should be seen within the greater discussion on data-driven and corpus-based annotation systems for prosody and intonation of spontaneous speech [12–13].

The design of the annotation hierarchy

Figure 1 shows the suggested annotation hierarchy.

(1) Comment tier = COM
(2) Segmental tier = SEG
(3) Syllable tier = SYL
(4) Phonological tier = IPH
(5) Pragmatic tier = PRA
(6) Information structure tier = INF
(7) Micro-prosodic (phonetic) tier = MIP
(8) Rhythmic/Prominence tier = PRO
(9) Prosodic constituency tier = PCO
(10) Orthographic tier = ORT
(11) Glossed tier = GLO

Fig. 1: Annotation hierarchy to annotate prosodic and intonational HC corpus data.

On the comment tier, penultimate lengthening (PL), breathing, or pauses that are perceived by the phonologist are annotated. On the segmental tier, segments are encoded using the IPA symbols. This segment-sized segmentation allows durational measurements of vocalic segments to assess HC penultimate lengthening. On the syllable tier, auditorily perceived syllable structures are annotated to assess HC syllabic patterns and complexity.

The following three annotation tiers, i.e. the phonological, the pragmatic, and the information structure tiers, are tightly connected. Nevertheless, the pragmatic and the information structure tiers are a priority for the annotation process. On the pragmatic tier, speech acts are annotated and on the information structure tier focus accents (FA) are highlighted to hypothesize about HC pragmatic tones. Only on the basis of recurrent pitch patterns relative to specific speech acts and focus marking, can first hypotheses about the intonational phonology of HC be annotated on the phonological tier by parsimonious use of the set of ToBI labels [2].

At the moment, this path of knowledge from the pragmatic and information structure information of spontaneous speech corpus data to the recognition of recurrent intonational patterns seems to be the only practical way to move forward. Despite the imperfection of this mixed data-driven and knowledge-driven path, it tends to evade fully knowledge-driven approaches to intonation, such as the Discourse Completion Task [14]. Unfortunately, at this juncture, no reasonably manageable data-driven intonation analysis tools are available [15–16].

Figure 2 shows the labels for the seven speech acts found in our short HC test sample taken from the APiCS online. This short list reflects the core speech act nomenclature of speech act theory and speech act annotation [17–20]:

Statement = STA
Enumeration = ENU
Yes-no-question =YNQ
WH-question = WHQ
Indirect question = INQ
Command = COM
Request = REQ

Fig. 2: An extendible set of speech act labels.

The micro-prosodic tier serves to phonetically annotate the realized pitch movement of the range of prominent syllables or implementation domain (ID) using the set of IViE labels [5]. “Each ID contains the preaccentual syllable, the accented syllable, all following syllables (if any) up to the next accented syllable” [5]. The upper-case labels H, M, and L indicate high, mid, and low pitch levels (e.g. L) or pitch glides, and on accented syllables (e.g. LH). The lower-case labels h, m, and l indicate the pitch levels of the unstressed syllable preceding the strong syllable and any unstressed syllables following the strong syllable. The label “-” is used to connect the final pair of phonetic labels in an ID.

As the pitch of the female HC speaker of the APiCS sample ranges from 150 to 300 Hz, I distributed the tone labels as follows: H and h labels were assigned to the two uppermost quarters of her pitch range, i.e. from 230 to 300 Hz; L and l labels were assigned to the lowermost quarter of her pitch range, i.e. from 150 to185 Hz; and M and m labels were assigned to the range in between, i.e. from 185 to 230 Hz.

On the rhythmic or prominence tier, auditorily perceived stress and pitch-accented prominent syllables (P), rhythmic boundaries (%), and hesitations or speech errors (#) are annotated. The identification and annotation of prominences and rhythmic boundaries are always the starting point for the prosodic annotation of speech data [4, 12]. Accented syllables and rhythmic boundaries are landmarks within the whole annotation process.

HC prosodic constituents are indicated on the prosodic constituency tier using ToBI break indices (BI) ranging from 0 to 4: the lowest Index, BI 0, indicates junctions between content and function words; BI 1 indicates phonological word (ω) boundaries; BI 2 indicates clitic group boundaries; BI 3 indicates intermediate phrase (ip) boundaries; and the highest break index, BI 4, indicates intonational phrase (IP) boundaries [21]. The identification of the HC prosodic constituents was important in order to understand that the prosodic domain of accent assignment is the phonological word. Knowing that the French superstrate has a totally different accentuation pattern, namely a fixed phrasal accent on the rightmost full syllable of the intonational phrase [22], that is a very interesting insight. In contrast to French, HC is a right-accented word accent language.

The remaining two annotation tiers, the orthographic and glossed tier, serve the readability of the corpus data. On the orthographic tier, the speech signal is orthographically transcribed and on the glossed tier, HC lexical and grammatical items are glossed interlinearly according to the *Leipzig Glossing Rules* [23]. The glossing also serves to evaluate my prior observation of salient pitch movements on the HC function word class DET (determiner). These tonal movements may be a substrate trace of West African grammatical and pragmatic tones [24–25].

Figure 3 shows the annotation hierarchy implemented in Praat with a fully annotated section of the APiCS HC test sample.

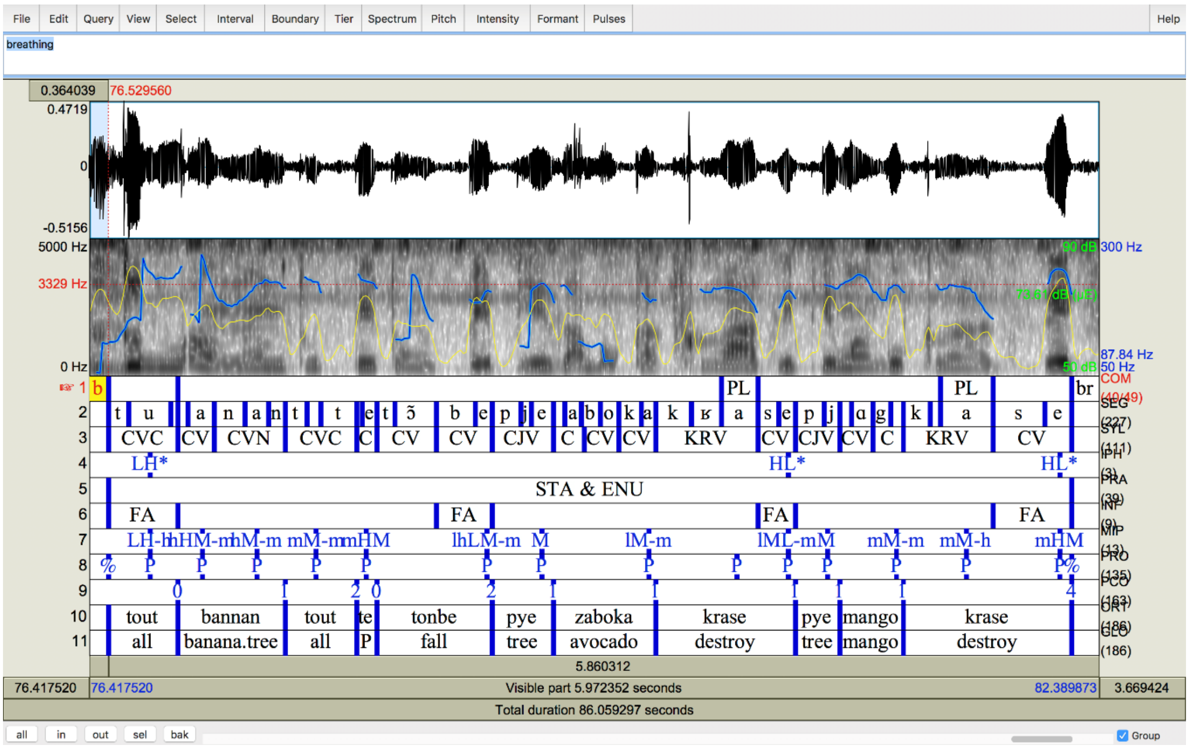

File
Edit
Query
View
Select
Interval
Boundary
Tier
Spectrum
Pitch
Intensity
Formant
Pulses
Help
breathing
0.364039
76.529560
0.4719
0
-0.5156
5000 Hz
3329 Hz
0 Hz
300 Hz
87.84 Hz
50 Hz
1 b PL PL br
2 t u a n a n t t e t ʒ b e p j e a b o k a k ʁ a s e p j a g k a s e
3 CVC CV CVN CVC C CV CV CJV C CV CV KRV CV CJV CV C KRV CV
4 LH* HL* HL*
5 STA & ENU
6 FA FA FA FA
7 LH-hhHM-mhM-m mM-mmHM lhLM-m M lM-m lML-mM mM-m mM-h mHM
8 % P P P P P P P P P P P P P P%
9 0 2 0 2 4
10 tout bannan tout te tonbe pye zaboka krase pye mango krase
11 all banana.tree all P fall tree avocado destroy tree mango destroy
5.860312
76.417520
76.417520
Visible part 5.972352 seconds
82.389873
3.669424
Total duration 86.059297 seconds
all
in
out
sel
bak
Group
COM (4049)
SEG (227)
SYL (111)
IPH (3)
PRA (39)
INT (9)
MIP (13)
PRO (135)
PCO (163)
ORT (186)
GLO (186)

Fig. 3: The annotation hierarchy implemented in Praat.

Conclusion

The approach to the HC prosodic system and in-tonational phonology presented here is a com-promise. Even by following the ideal of data-driven and corpus-based annotation and analy-sis of spontaneous speech, I cannot deny many knowledge-based aspects, such as the identifi-cation of speech acts by the phonologist.

Unfortunately, easily manageable data-driven prosody and intonation analysis tools still remain desiderata.

At this point in time, the path of knowledge towards first hypotheses about HC intonational phonology, moving from the speech act and focus accent annotation to the recognition of recurrent intonational patterns, seems to me to be a realistic and promising one. The annotation hierarchy suggests that this is a work in progress. Further testing and re-evaluation in light of new insights will modify it.

Acknowledgments

I am very grateful to Elisabeth Delais-Roussarie with whom I discussed my corpus-based approach to the annotation of HC prosody and intonation.

References

[1] Teixeira Kalkhoff, A.,"Éléments et structures suprasegmentaux", *Manuel des langues créoles à base française*, in Krämer, P. Mutz, K. & P. Stein (Eds.), Berlin: De Gruyter. In press.

[2] Silverman, K. et al.,"ToBI: A standard for labeling English prosody", *Proceedings from the 2nd International Conference on Spoken Language Processing (ICSLP 92), Banff, Canada*, pp. 867–870, 1992.

[3] Grabe, E. & B. Post, "Intonational variation in the British Isles", *Corpus Linguistics: Readings in a widening discipline*, in Sampson, G. & D. McCarthy (Eds.), London: Continuum International, pp. 474–481, 2004.

[4] Post B. & E. Delais-Roussarie, "Transcribing intonational variation at different levels of analysis", *ISCAM Archives, Speech Prosody 2006, Dresden, Ger-many, May 2-5*, 2006.

[5] *The IViE Labelling Guide*, version 3, copyright by Esther Grabe, 2001 (http://www.phon.ox.ac.uk/files/apps/IViE/guide.html) [12.10.2018].

[6] Brousseau, A.-M., "The accentual system of Haitian Creole: The role of transfer and markedness values", *Phonology and Morphology of Creole Languages*, in Plag, I. (Ed.), Tübingen: Niemeyer, pp. 123–145, 2003.

[7] Brousseau, A.-M. & E. Nikiema, "From Gbe to Haitian: The multi-stage evolution of syllable structure", *L2 Acquisition and Creole Genesis*, in Lefebvre, C., White, L. & C. Jourdan (Eds.), Amsterdam: Benjamins, pp. 295–330, 2006.

[8] Teixeira Kalkhoff, A., "Using corpus data for testing a working hypothesis about Haitian Creole prosody", 2018 (http://alexander.teixeirakalkhoff.eu/wp-content/uploads/2019/01/Teixeira-Kalkhoff_HC-prosody-working-hypothesis-1.pdf) [12.10.2018].

[9] Fattier, D., "Remarques sur l'interrogation en créole haïtien", *La Linguistique*, 41(1), pp. 41–55, 2005.

[10] Boersma, P. & D. Weenink, *Praat: Doing phonetics by computer* [Computer program], version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/.

[11] Fattier, D., "Haitian Creole structure dataset", *Atlas of Pidgin and Creole Language Structures Online*, in Michaelis, S., Maurer, P., Haspelmath, M. & M. Huber (Eds.), Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013 (Available online at http://apics-online.info/contributions/49) [12.10.2018].

[12] Delais-Roussarie, E. & B. Post, "Corpus annotation: methodology and transcription systems", *The Oxford Handbook of Corpus Phonology*, in Durand, J., Gut, U. & G. Kristoffersen (Eds.), Oxford: OUP, pp. 46–88, 2014.

[13] Brierley, C. & E. Atwell, "ProPOSEC: A prosody and POS annotated spoken English corpus", *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2010, Valletta, Malta, pp. 1266–1270, 2019.

[14] Prieto, P., "L'entonació dialectal del català: el cas de les frases interrogatives absolutes", *Actes del novè colloqui d'estudis catalans a Nord-Amèrica*, in Bover, A., Lloret, M.-R. & M. Vidal-Tibbits (Eds.), Barcelona: Publicacions de l'Abadia de Monserrat, pp. 347–377, 2001.

[15] Reichel U. D., *Datenbasierte und linguistisch interpretierbare Intonationsmodellierung*, München: Dissertation, LMU München, 2010.

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

[17] Bach, K. & Harnish R. M., *Linguistic Communication and Speech Acts*, Cambridge, Mass.: MIT Press, 1979.

[18] Leech, G. & M. Weisser, *The SPAADIA annotation scheme*, 2014 (http://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf) [12.10.2018].

[19] Prieto, P. et al. (Eds.), *Interactive Atlas of Romance Intonation*, 2010–2014. (http://prosodia.upf.edu/iari/) [12.10.2018].

[20] Frota, S. & P. Prieto (Eds.), *Intonation in Romance*, Oxford: OUP, 2015.

[21] Frota, S., "Prosodic structure, constituents and their implementation", *The Oxford Handbook in Laboratory Phonology*, in

Cohn, A. C. Fougeron, C. & M. K. Huffman (Eds.), Oxford: OUP, pp. 255–265, 2012.

[22] Van der Hulst H. et al., *A Survey of World Accentual Patterns in the Languages of the World*, Berlin: De Gruyter, 460, 2010.

[23] *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*, 2015. (https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf) [12.10.2018].

[24] Stahlke, H. F. W., *Topics in Ewe phonology*, Dissertation, LA: University of California, 1971 (URL: http://linguistics.ucla.edu/images/stories/Stahlke.1971.pdf) [12.10.2018].

[25] Fiedler I. & S. Jannedy S. (2013). “Prosody of focus marking in Ewe”, *Journal of African Languages and Linguistics*, 34(1), 1–46.

Notions of similarity in non-native vowel perception

Nadja Kerschhofer-Puhalo[1]

[1]Universität of Vienna, Austria

nadja.kerschhofer@univie.ac.at

Abstract

Similarity is one of the central concepts in many models of second language speech perception and acquisition [1–3, 11–12]. Based on empirical data from a large-scale vowel identification experiment [9], this paper will discuss empirically grounded ways to operationalize similarity in L2 by integrating contributions from L2 research, experimental phonetics and cognitive psychology. A method to visualize perceptual similarity of L2 categories by Multidimensional Scaling is presented. Rather than predicting perceptual similarity directly from acoustic-phonetic properties, this approach avoids the a priori *selection of specific phonetic parameters and offers a better account of 'psychological similarity' between L2 sound categories as established by the learners.*

Introduction

Vowels are complex linguistic structures that vary within and across languages according to several articulatory and acoustic phonetic parameters. In articulation, vowels differ in several parameters such as tongue location, lip rounding, larynx lowering, palatal narrowing or tongue blade elevation. Acoustically, vowels are described in terms of several continuous metric variables such as formants, formant distances, overall spectral shape, bandwidth, amplitude or dynamic spectral change. Rather than one specific distinctive feature, multiple co-occurring articulatory gestures and acoustic cues that are partly context-specific, contribute to accurate vowel identification. In speech perception, however, listeners show an astounding capacity to discriminate and identify vowel sounds correctly in a language they are familiar with due to relatively stable mental representations and lexically and morphologically driven top-down processes. However, no one-to-one mapping can be assumed between the speech signal (phonetic substance) and the category perceived by the listener (phonological function). Vocalic form-function pairings vary language and context-specifically, and even individually, as a result of a listener's language experience in L1 and L2.

Considering these conceptual and practical difficulties to define 'similarity' in L2 learning, central findings of a study on the perceptual similarity of German vowels in L2 acquisition based on a vowel identification experiment with L2 and L1 German listeners [9] will be presented here. The collected data was used to find empirically grounded solutions for the operationalization of intra-lingual perceptual similarity in L2 German. The study is based on two central distinctions: (1) phonetic vs. perceptual/ psychological similarity referring to the non-linear and language-specific relation between articulatory or acoustic properties of vowels, and (2) the necessity to distinguish (a) *inter-lingual* differences and similarities between sounds in L1 and L2 from (b) *intra-lingual* perceptual similarity and confusion between categories *within* the L2 target language.

Categorial or continuous? – Visualizing the 'vowel space'

Research on vowel sounds, vowel inventories and vowel perception is strongly determined by the co-existence of two theoretical paradigms: the *categorization* paradigm and the paradigm of the *continuous vowel space*.

Phonetically, vowels are described as sounds varying along several continuous articulatory or acoustic parameters, while vowel perception apparently is determined by discrete category boundaries. Experimental phonetic studies on vowel perception and vowel identification are strongly influenced by the categorial view. Phonetic characteristics are commonly understood as gradients varying along a continuum, whereas category boundaries are conceived as abrupt changes in the listeners' perception.

Speaking of vowel systems, reference is often made to a system of discrete categories, while phonetically-oriented approaches rather focus on the construct of the continuous vowel space. The articulatory-acoustic vowel space refers to an abstraction of the total set of physiologically possible articulatory constellations and their acoustic resonances in the vocal tract. However, the idea of the 'vowel space' as a spatial constellation is a metaphor. We find this spatial metaphor in two very common types of representations: the IPA vowel quadrilateral for articulation and F1xF2-charts in acoustics.

A comparison of vowels and vowel systems in terms of two-dimensional F1xF2 charts, a wide-spread practice, is of course too simplistic to reflect the many phonetic dimensions and the continuous, dynamic character of vowel sounds. Dynamic spectral change, transitions to adjacent consonants, duration and fundamental frequency provide substantial acoustic cues in vowel identification [5]. Similarly, the conventional IPA vowel quadrilateral, which is an abstraction [6: 12] that seems to correspond to the F1xF2-view, is often misunderstood as directly reflecting the tongue position, e.g. suggesting a higher degree of backness of *u* and *o* vowels compared to *a* vowels which is not confirmed by articulatory data.

Phonetic and phonological description of German vowels

For the study on perceptual similarity of vowels in L2 [9] an articulatory-acoustic model of the German vowel system based on a classification in terms of the major internal narrowing in the vocal tract served as a starting point. [18–19] distinguishes (1) the lower pharyngeal region for [ɑ-a-æ]-like *pharyngeal* vowels, (2) the upper pharyngeal region for [o-ɔ] and [ɤ]-like *uvular* vowels, (3) the vicinity of the soft palate for [u-ʊ] and [ɨ]-like *velar* vowels, and (4) the hard palate for [i-ɛ]-like and [y-ø]-like *palatal* vowels. Vowel sounds located in these regions are perceptually and acoustically relatively unambiguous. For palatal vowels, language-specific tendencies to either pre-palatal or mid-palatal tongue positions are reported [21]. Languages contrasting [i] and [y] such as German seem to prefer the pre-palatal position for both [18–19; for German 13–14].

As an alternative visual representation of the German vowel system corresponding to the classification of [18–19], I propose an elliptic representation [9] to distinguish German vowels by constriction location: (1) *pharyngeal* /ɑ ɑ:/, (2) *uvular* /ɔ o:/, (3) *velar* /ʊ u:/, (4) *mid-palatal* /ɛ ɛ: e: œ ø:/ and (5) *pre-palatal* /ɪ i: ʏ y:/ qualities [20]. These are further differentiated into mandibular aperture, rounding, degree of constriction and phonemic length.

In a next step, German vowel categories were compared on the basis of acoustic measurements of the experiment's input stimuli and predictions on the relative similarity and difficulty of categories and contrasts were derived from a Hierarchical Cluster analysis. Different HC solutions were calculated as a function of the selected phonetic parameters (spectral peaks/formant frequencies or formant distances, duration, Hertz or Bark etc.) in order to group the full set of German vowels into 'phonetically similar' sub-groups according to their acoustic properties and to describe possible areas of difficulty and confusion in L2 due to acoustic similarity of the input stimuli.

Experiment

173 adult L2 learners of German (beginners and advanced) and 18 German natives (control group) participated in the experiment. The sample consisted of ten L1 sub-samples (Albanian, Arabic, English, Farsi, Hungarian, Mandarin, Polish, Romanian, SerBoCroatian, Turkish). The participants performed a category identification task: 15 German vowel phonemes and were presented in 12 different consonantal prevocalic and postvocalic contexts in non-words embedded in two carrier sentences: *CVtə bedeutet auch nichts.* ('CVtə means also nothing') or *Er meint heute CVC.* ('He means/says today CVC'). 15 German full vowels and three diphthongs were offered as response options.

Data Analysis and Results

The 45,794 valid L2 responses (270 items x 173 participants) and 4,860 L1 German responses were subject to further analysis. The listeners' wrong and correct responses were summarized in a confusion matrix. Confusion matrices indicate the frequency with which certain categories are selected as response option (*id_V scores*). 'Wrong' refers to cases where stimulus *i* is identified as belonging to category *J*.

As expected, the results clearly show that the same acoustic input is perceived according to the listeners' language backgrounds. Patterns of confusion in a given L1 sub-group are considered to reflect language-specific difficulties in L2 perception. The sample shows language-specific differences as well as more common substitution patterns that are found in several sub-samples.

Four major areas of *confusion* were observed in all sub-samples: confusion between (1) prepalatal and midpalatal non-round *i* and *e* vowels /i: ɪ e: ɛ ɛ:/, (2) velar *u* and uvular *o* qualities /u: ʊ o: ɔ/, and (3) prepalatal and midpalatal rounded *ü* and *ö* qualities /y: ʏ ø: œ/, which in some samples are insufficiently differentiated from *u* and *o* vowels; (4) *a* vowels were not sufficiently differentiated with respect to length.

While it is of course worthwhile to look at the language-specific results in detail, the focus of the present paper is to discuss ways of deriving information about perceptual similarity on a higher scale level (ordinal or metric) than a simple descriptive statistical analysis (nominal) could offer.

Similarity scores for calculating the relative similarity of two given categories *i* and *j* were derived from *id_scores* in the confusion matrix by Shepard's method [12, 8]. These were transformed into values of perceptual distance by taking the negative of the natural logarithm of similarity. A *Multidimensional Scaling* analysis (MDS) was then calculated to map the L2 listeners' perceptual vowel space. These perceptual vowel maps visualize psychological relationships of distance and proximity between German vowels according to similarities as perceived by L2 learners.

MDS – Mapping the perceptual vowel space

Multidimensional Scaling [12, 13] is a statistical procedure to reduce a higher number of parameters in which stimuli may vary (e.g. phonetic dimensions) to a low-dimensional representation of the psychological space. This representation is constructed either from proximity data such as perceived similarities or from distances. It offers a visualization of perceived intra-lingual similarity *within* the L2 vowel system. MDS solutions for the perceptual L2 vowel space were calculated for each of the ten language sub-groups (for a detailed language-specific discussion see [9]). Just to give an illustration of this visualizing technique, a two-dimensional MDS representation for the full L2 sample (N=173, RSQ .905) is presented in Fig. 1. Four major perceptual clusters of German vowels are identified: (1) *pharyngeal a*-vowels, (2) *back rounded u* and *o* qualities, where /ɔ/ is differentiated from other back rounded qualities, (3) a slightly more differentiated cluster of *front rounded* vowels, and (4) a cluster for *palatal* vowels that is divided into three sub-clusters: (a) *i* qualities, (b) *ɛ* qualities, and (c) /e:/ in an intermediate position between *i* and *ɛ* qualities. Mapping phoneme types onto a geometric two-dimensional or three-dimensional MDS space has been used in several previous studies on perceptual similarity of phonemes [e.g. 4, 7, 10, 17]. Some studies have postulated a direct correspondence of distances in MDS solutions and specific acoustic-phonetic properties (e.g. vowel formants). However, it has to be emphasized that MDS dimensions do not directly correspond to phonetic dimensions such as F1 or F2 [16].

For a qualitative interpretation of a given MDS solution it is necessary to consider the relative distances between items in the spatial representation rather than the absolute position of a specific item in the space. Distances in the MDS solutions reflect the relative similarity of items in the perceptual space. The position of a vowel type and its distance from all other categories in the system are determined by the L2 listeners' response behavior and vary systematically with the listeners' L1 background and L2 experience.

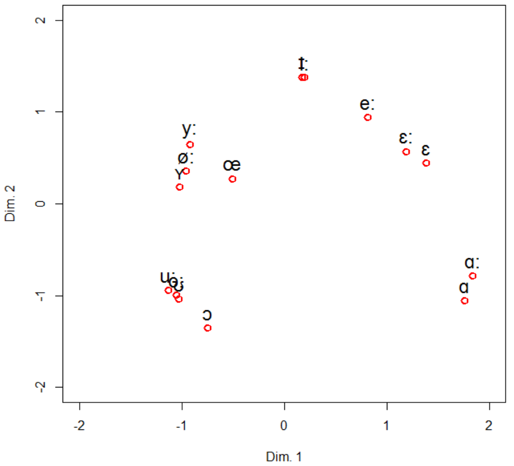


Figure 1: Two-dimensional non-rotated MDS solution for all 173 L2 listeners [8: 255].

Conclusion

To summarize, *perceptual similarity* has to be distinguished from *phonetic similarity*. It is, however, not directly predictable from physical properties of the acoustic signal but is determined by the listener's attentional tuning to specific cues and dimensions and by language-specific and more general physical and cognitive biases associated with stimuli as well as responses.

Rather than predicting perceptual similarity directly from acoustic-phonetic properties, MDS offers an approach to similarity as perceived by the listeners, which is independent of *a priori* selected phonetic parameters. This approach to 'psychological similarity' avoids the difficulty of selecting specific phonetic parameters, which may not correspond to the listeners' attentional weighting of cues in the signal.

For perceptual similarity between two items *i* an *j* the formula $s_{ij} = p_{ij} * b_i * b_j$ is proposed, referring to perceived similarity s_{ij}, which is conditioned by the interaction of phonetic distance or proximity p_{ij}, stimuli bias b_i and response bias b_j. Not only acoustic properties of the stimuli but also the listeners' individual considerations, expectations and (temporary) hypotheses influence the participants' responses. Differential biases in perception must be distinguished in terms of (a) *stimulus*-related biases b_i vs. *response*-related biases b_j, (b) *signal*-related vs. *listener*-related biases, and (c) set-dependent vs. set-independent biases (for discussion, see [9]). Or to put it very simply: Perceptual similarity is, after all, a concept of the listener, not the phonetician.

References

[1] Best, C., "A direct realist perspective on cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, W. Strange (Ed.), Timonium, MD: York Press, pp. 171–204, 1995.

[2] Flege, J., "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification," *JPhon* 15, pp. 47–65, 1987.

[3] Flege, J., "Second-language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, Strange, W. (Ed.), Timonium, MD: York Press, pp. 233–277, 1995.

[4] Fox, R., Flege, J. & M. Munro, "The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis," *JASA* 97, pp. 2540–2551, 1995.

[5] Hillenbrand, J., "Static and dynamic approaches to vowel perception," in *Vowel inherent Spectral Change. Modern Acoustics and Signal Processing*, G. Morrison & P. Assmann (Eds.), Berlin/Heidelberg: Springer, pp. 9–30, 2013.

[6] IPA [International Phonetic Association] (Ed.), *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge: CUP, 1999.

[7] Iverson, P. & P. Kuhl, "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling," *JASA* 97, pp. 553–562, 1995.

[8] Johnson, K., *Acoustic and Auditory phonetics*, Wiley-Blackwell, ^{3}2012.

[9] Kerschhofer-Puhalo, N., *Similarity, Cross-linguistic Influence and Preferences in Non-Native Vowel Perception. An experimental cross-language comparison of German vowel identification by non-native listeners*, Dissertation, University of Vienna, (permalink: http://phaidra.univie.ac.at/o:902493), 2014.

[10] Kewley-Port, D. & B. Atal, "Perceptual distances between vowels located in a limited phonetic space," *JASA* 85, pp. 1726–1740, 1989.

[11] Kuhl, P., "Speech prototypes: studies in the nature, function, ontogeny and phylogeny of the 'centers' of speech categories," in *Speech Perception, Production and Linguistic Structure*, Tohkura, Y., Vatikiotis-Bateson, W. & Y. Sagisaka (Eds.), Tokyo: Ohmsha, pp. 197–219, 1992.

[12] Kuhl, P., "An examination of the 'perceptual magnet' effect," *JASA* 93, 2423, 1993.

[13] Moosmüller, S. *Vowels in Standard Austrian German. An Acoustic Phonetic and Phonological Analysis*, Habilitationsschrift, Universität Wien, 2007.

[14] Moosmüller, S., Schmid, C. & J. Brandstätter, "Standard Austrian German," *JIPA* 45(3), pp. 339–348, 2015.

[15] Shepard, R., "Psychological representation of speech sounds," in *Human Communication: A Unified View*, David, E. & P. Denes (Eds.), New York: McGraw-Hill, pp. 67–113, 1972.

[16] Shepard, R., "Multidimensional Scaling, Tree-Fitting, and Clustering," *Science* 210(4468), pp. 390–398, 1980.

[17] Terbeek, D., *A Cross-Language Multi-Dimensional Scaling Study of Vowel Perception*. PhD Dissertation. University of California, Los Angeles, 1977.

[18] Wood, S., "A radiographic analysis of constriction location for vowels," *JPhon* 7, pp. 25–43, 1986.

[19] Wood, S., "The acoustical significance of tongue, lip, and larynx manoeuvers in rounded palatal vowels," *JASA* 80(2), pp. 391–401, 1986.

Implicit dialect association and accommodation: First results from a pilot study on German Bavarian and Austrian Tyrolean

Felicitas Kleber[1], *Katharina Mittelhammer*[1]

[1]Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München

kleber@phonetik.uni-muenchen.de, katharina.mittelhammer@gmx.de

Abstract

The paper presents first results from an imitation study with German and Austrian participants.

Introduction

Automatic and subconscious accommodation in face-to-face conversation has been proposed as the underlying mechanism for dialect mixture [1] in adult speakers [2]. Besides convergence towards an interlocutor (with linguistic difference decreasing during interaction), speakers may also diverge from the interlocutor, i.e. they may enhance the linguistic difference. Evidence for phonetic convergence resulting from imitation has been presented in studies on laboratory [e.g. 3] and spontaneous speech [e.g. 4], but phonetic convergence is also highly variable, and speakers who are found to converge on one acoustic dimension may simultaneously diverge on another [5].

Babel [6], in her investigation of New Zealand speakers' productions prior to, during, and after listening to an Australian model talker, found that there was a general trend towards convergence in form of small changes below the level of consciousness, but also that social factors played a role. More precisely, greater convergence was found in speakers who were more pro-Australian.

This pilot study is a replication of Babel's study [6] and extends it by looking at bidirectional accommodation in speakers of Southern-Central Bavarian, a regional variety of German that is spoken across the Austrian-German state border. More precisely, the two regions under investigation are the Tegernsee area in the German federal state of Bavaria and Kitzbühel, located in Tyrol, Austria. The two regions share many dialectal similarities, but they also display some differences: for example, only in Kitzbühel are standard German (SG) palatal fricatives velarized (i.e. /ç/ → [x]) [7].

The specific goals were firstly to quantify acoustically vocalic and fricative differences between the two areas before measuring phonetic accommodation between the speakers of the two regions in an imitation study. In a last step we combined the phonetic accommodation results with a measurement of participants' implicit association with the other region. We were particularly interested in whether the groups differ in terms of convergence or divergence [4] and whether implicit association predicts accommodation.

Method

Eight female speakers from Kitzbühel (henceforth Tyroleans) and eight females from the Tegernsee region (henceforth Bavarians) participated in a two-part study comprising a speech production and an implicit association task [8]. Participants were aged between 35 and 61 years and had lived in the respective dialect area for all or most of their lives. They were all speakers of the respective local variety of Southern-Central Bavarian.

Prior to the experiments we recorded one female model talker from each region (aged 44 and 51 years, respectively) producing the same speech materials used in the experiment.

Speech Production

Recordings were made using the SpeechRecorder software package [9]. First, speakers had to name a total of 21 different target words presented as pictures three times each, in randomized order. A picture naming task was chosen to elicit more natural productions of the target words and to avoid potential influences from the orthography. All target words were selected based on the description of dialect differences between the two regions in Wiesinger [7]. These words may not be part of the dialectal core vocabulary but were uttered by the speakers in their regional accent. This paper only reports on findings for the following three target words: *Becher* (SG: /bˈɛçɐ/, ‘cup’), *Dichter* (SG: /dˈɪçtɐ/, ‘poet’), and *Rechen* (SG: /rˈɛçn̩/, ‘rake’). That is, a total of 144 baseline productions were available for analysis (3 words × 3 repetitions × 16 speakers).

In a second step participants then completed a shadowing task, where they heard productions of the same target words by the model talker from the other region (i.e. Tyroleans repeated words from the Bavarian model talker and *vice versa*). Participants were not informed about the regional background of the model talker they were about to hear. Upon presentation of a target word speakers were instructed to repeat loudly the word they had perceived. Each word was again repeated three times and all words were presented in randomized order. To test for longer-term accommodation effects, all speakers participated in a posttest which was identical to the pretest except for all target words being produced only once.

All recordings were then automatically segmented using MAUS [10] and stored as an EMU speech database [11]. Prior to data analysis each segment boundary was checked and corrected manually if necessary, using the EMU-webApp.

All further analyses were done using the R programming language. Between-group vowel-quality differences were assessed by analyzing the first two formant frequencies extracted at the temporal midpoint of the stressed vowel. To determine velarization of the palatal fricative (e.g. Bavarian /ç/ vs. Tyrolean /x/) we applied Discrete Cosine Transformation (DCT) to capture differences in the spectral shape [12], again at the temporal midpoint of the fricative. Only the 1st (slope) and 2nd DCT (curvature) coefficients (henceforth DCT1 and DCT2) were analyzed since they sufficiently separate the German fricatives acoustically [12].

Acoustic differences in a two-dimensional acoustic space between participants and model talkers were quantified by calculating separately for vowels and fricatives the Euclidean distance between each speaker’s average word realization (i.e. aggregated across repetition) per condition (i.e. pretest, shadowing, posttest) to the respective model talker’s value of the same word leaving us with a single distance measure per speaker, category (i.e. vowel or fricative), word, and condition. Finally, we computed the difference between the distance values from the shadowing task and the posttest, respectively, to the corresponding distance values from the pretest. Negative difference-in-distance values are indicative of phonetic convergence, as they reflect a decreasing acoustic distance between a participant and the model talker. Positive values indicate an increase in acoustic distance and hence phonetic divergence.

Implicit Association Task

Following the speech production task all participants completed an implicit association task. In this test reaction times in judgments of combined primes of concepts and semantically good or bad attributes are measured, longer reaction times being interpreted as reflecting weaker association strengths [8].

Tab. 1 shows the target concepts and associated attributes for the corresponding stimuli used in the current study. We selected the country names *Austria* and *Germany* as superordinate target concepts for the two regions but stimuli that represent more local properties.

	Target concept		Associated Attributes	
	Austria	Germany	good	bad
Stimuli	Salzburg	Regensburg	rainbow	pain
	Sisi	F. J. Strauß	health	hell
	Innsbruck	Augsburg	sunshine	war
	Groß-glockner	Bavaria	laughter	hate
	Steiermark	Weißwurst	peace	poison
	Falco	König Ludwig	love	death
	Mozart	Beethoven	flowers	accident

Tab. 1: Stimuli used in the IAT representing one of two concepts and attributes. All stimuli were equally easy to categorize for Bavarians and Tyroleans.

The test was conducted using PsychoPy [13]. Following the methodology in [8], in training blocks participants first categorized each stimulus as AUSTRIA or GERMANY and as GOOD or BAD. In subsequent blocks participants again categorized stimuli (e.g. *Salzburg*) but this time with the target concepts and associated attributes combined. That is, *Salzburg*, for example, had to be categorized as AUSTRIA while once *Austria* was paired with *rainbow* and *Germany* with *pain* and once *Austria* with *pain* and *Germany* with *rainbow*. The test contained all combinations and was counter-balanced [see 8 for details].

Participants' responses were then postprocessed following the method in [8], which in a nutshell quantifies for each participant the strength of association by (1) calculating for each training and test block the mean reaction times of correct responses and their pooled standard deviations, (2) computing the difference between means of training blocks, on the one hand, and between means of test blocks, on the other, and (3) dividing these difference values by the respective pooled standard deviations (i.e. the test block difference by the test block standard deviation, etc.), which gives us the IAT score. While a pro-Austrian bias is reflected by a negative IAT score, a pro-German bias is indicated by a positive IAT score.

Results

As shown in Fig. 1, Bavarians' baseline productions of <ch> were more /ç/-like than those of Tyroleans, though not as /ç/-like as those of standard German speakers. Tyroleans, on the other hand, velarized the fricatives to a greater extent than Bavarians, which is in line with auditory-based descriptions of the variety as reported in the literature [7].

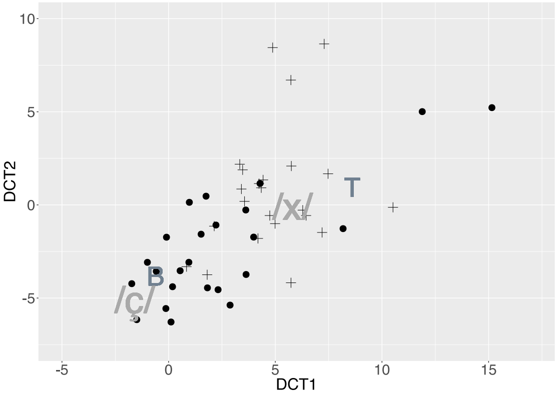


Fig. 1: Baseline realization of <ch> by Bavarian (•) and Tyrolean (+) speakers as well as the Bavarian (B) and Tyrolean (T) model talkers within the DCT1 × DCT2 space. /ç/ *and* /x/ *represent mean values aggregated across several* /vˈɪçn̩/ *and* /vˈaxn̩/ *realizations from 20 younger female speakers of standard German taken from a reference corpus.*

Speakers from the two regions also differed in the quality of /ɛ/, with Bavarians' baseline realizations less open than those of Tyroleans (cf. Fig. 2). This difference may have been enhanced because of the differences in the following fricative, although no vocalic difference was found for /ɪ/ in *Dichter*.

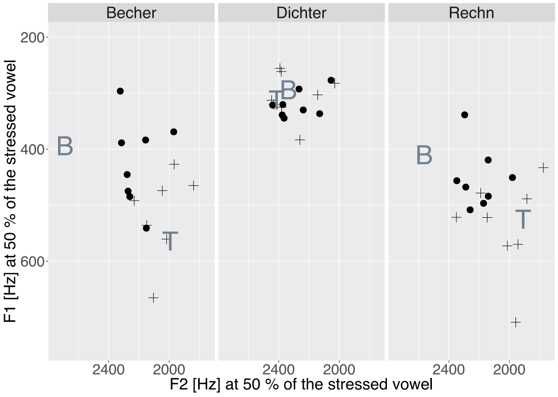


Fig. 2: F1 and F2 measured at the temporal midpoint of the stressed vowel separately for the three target words produced in the pretest by Bavarians (•) and Tyroleans (+) as well as the Bavarian (B) and Tyrolean (T) model talkers.

The results from this baseline analysis suggest that speakers from the two regions differed in the selected target words in both the vocalic and fricative dimension. The two model talkers were found to pattern with their peers in both dimensions (cf. Fig. 1 and 2).

Fig. 3 shows separately for vowels and fricatives the word-aggregated difference-in-distance values from the shadowing task for each speaker plotted against their IAT scores. The IAT scores of Bavarian participants were all positive and those of Tyroleans all negative. This finding suggests that all participants had a pro-home region bias. This bias, however, did not predict the direction of accommodation.

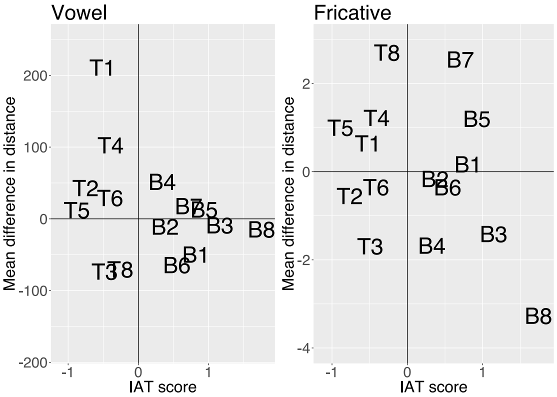


Fig. 3: Each Bavarian's (B) and Tyrolean's (T) averaged short-term difference-in-distance value (shadowing–pretest) plotted against their IAT score separately for vowels (left) and fricatives (right).

With respect to vowels, a greater number of Bavarians than Tyroleans show negative difference values, which indicate phonetic convergence towards the Tyrolean model talker. Most Tyroleans, on the other hand, showed positive values indicating phonetic divergence from the Bavarian model talker. Regarding the acoustic difference of fricatives, the findings are more variable in that convergers and divergers were found equally in both groups. As stated above, accommodation did not correlate with the IAT score, since speakers with a strong pro-home bias like B8 and B3 were nevertheless found to converge phonetically, while speakers with a less pronounced bias like B4 diverged.

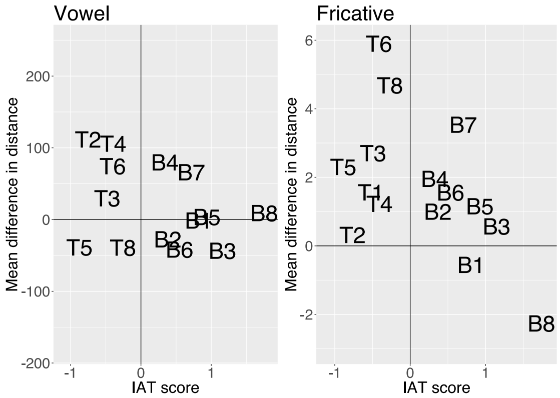


Fig. 4: Each Bavarian's (B) and Tyrolean's (T) averaged long-term difference-in-distance value (posttest–pretest) plotted against their IAT score separately for vowels (left) and fricatives (right).

Fig. 4 is structured like Fig. 3 but shows the difference-in-distance values from the post test. With respect to fricatives, most speakers showed positive values, pointing to phonetic divergence. Convergence appears to be longer lasting in vowels, given the greater number of speakers showing negative values.

Discussion and Conclusion

The main observations arising from this first analysis are threefold: (1) speakers from the two regions differed acoustically in vowel quality and velarization, (2) both groups contained convergers and divergers, although the relation appears to be asymmetric in that Austrians tend towards more divergence than Bavarians, and (3) the IAT scores did not correlate with accommodation. Where convergence is present then it appears to be stronger in vowels than in fricatives.

The findings of (1) both divergence and convergence as well as (2) no correlation between accommodation and a speaker's implicit association towards a region stand in contrast to the findings described in Babel [6]. This suggests that convergence may not occur automatically and that the role of social factors might be weaker than previously assumed. In contrast with Babel [6], our participants were not informed about their model talker's regional background, although some ventured a guess after the experiment.

The tendency towards asymmetric accommodation effects depending on group [cf. 4 for similar results] suggests that dialect mixture may not necessarily be the ultimate result of dialect contact. The fact that Bavarians tended to converge towards Tyroleans' more open /ɛ/ vowel and /x/ realizations than *vice versa* may reflect a more general tendency to converge towards more dominant or standard variants: thus /ɛ/ and not /e/ corresponds to the standard German pronunciation norm; /x/, on the other hand, does not reflect a standard German pronunciation. However, Bavarians already tended towards a more /x/-like realization in the pretest.

Apart from analyzing the entire corpus, the next steps will be to quantify statistically the accommodation trends described here and to carry out perception tests to substantiate the accommodation findings as described in Pardo [5].

References

[1] Trudgill, P., “Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation,” *Language in Society*, 37, pp. 241–254, 2008.

[2] Babel, M et al., “Novelty and social preference in phonetic accommodation,” *Journal of Laboratory Phonology*, 5(1), pp. 123–150, 2014.

[3] Goldinger, S. D., “Echoes of echoes? An episodic theory of lexical access”, *Psychological Review*, 105, pp. 251–279, 1998.

[4] Ruch, H., Vowel convergence and divergence between two Swiss German dialects. *Proceedings 18th ICPhS, Glasgow,* 2015.

[5] Pardo, J. S., “Measuring phonetic convergence in speech production,” *Frontiers in psychology*, 4, article 559, 2013.

[6] Babel, M., “Dialect divergence and convergence in New Zealand English,” *Language in Society,* 39(4), pp. 437–456, 2010.

[7] Wiesinger, P., “The Central and Southern Bavarian Dialects in Bavaria and Austria,” in *The Dialects of modern German: a linguistic survey*, Russ, C. V. J. (Ed.), London: Routledge, pp. 438–519, 1990.

[8] Greenwald, A. G., Nosek, B. A. & M. R. Banaji, “Understanding and using the implicit Association Test: 1. An improved scoring algorithm,” *Journal of Personality and Social Psychology*, 85, pp. 197–216, 2003.

[9] Draxler, C. & K. Jänsch, “SpeechRecorder – a universal platform independent multichannel audio recording software”, *Proceedings 4th LREC, Lisbon,* 2004.

[10] Kisler, T., Reichel U. & F. Schiel, “Multilingual processing of speech via web services”, *Computer Speech and Language*, 45, pp. 326–347, 2017.

[11] Winkelmann, R., Harrington, J. & K. Jänsch, “EMU-SDMS: Advanced Speech Database Management and Analysis in R”, *Computer Speech and Language*, 45, pp. 392–410, 2017.

[12] Jannedy, S. & M. Weirich, “The Acoustics of Fricative Contrasts in Two German Dialects,” *Proceedings P&P12, München,* pp. 70–73, 2016.

[13] Peirce, J. W., “PsychoPy – Psychophysics software in Python,” *Journal of Neuroscience Methods*, 162(1), pp. 8–13, 2007.

Arnold Schwarzenegger now and then: A longitudinal pilot investigation into Schwarzenegger's production of plosives in German and English

Lisa Kornder[1], *Ineke Mennen*[1]

[1]Karl-Franzens-University Graz, Department of English Studies

lisa.kornder@uni-graz.at, ineke.mennen@uni-graz.at

Abstract

The present study presents preliminary results of a longitudinal investigation into Arnold Schwarzenegger's realization of voice onset time (VOT) contrast in his first language (L1), Austrian German, and his second language (L2), English. Acoustic analyses of speech samples representing his early (1979–1989) and late (2010–2018) L1 and L2 pronunciation revealed that Schwarzenegger's realization of VOT contrast in his L2 is close to native norms. At the same time, his late L1 productions were observed to move towards the norms of the L2, which is interpreted as evidence for L1 phonetic attrition, that is, changes in L1 phonetic categories due to L2 learning experience [1, 2].

Introduction

Arnold Schwarzenegger (AS) was born in 1947 in a small village near Graz (Austria). To advance his career as a professional bodybuilder he moved to the United States at the age of 21, where he has been living ever since. Within a short period of time, he became *the* bodybuilding legend of the 1970s and gained world-wide fame. However, AS is not only well-known for his career in bodybuilding and the film business, but he also became famous for his accent [3]. Even after more than 50 years of living in an English-speaking country, AS still maintains a detectable Austrian German (AG)[1] accent when speaking his second language (L2) English, a fact that is often commented upon by the general public [3]. In addition, it is frequently discussed whether AS has forgotten or even 'unlearned' his first language (L1) [4], which he is said to speak with an English accent.

The present study is part of a larger research project which aims at investigating AS's L1 and L2 segmental speech production over a period of roughly 49 years, i.e. from the time AS moved to the US in the 1960s up to the 2010s. Based on acoustic investigations of plosives and vowel space, the main objectives of this research are to find out (1) whether speech production in AS's L2 has become more native-like since he moved to an L2-speaking environment, (2) whether there is evidence for L1 phonetic attrition, and (3) if it is possible to establish a relationship between the development of AS's L2 and potential modifications of phonetic categories in his L1.

While research into the factors contributing to foreign-accented L2 speech has made quite some progress in the last decades [5, 6], there are only few studies which examine attrition of pronunciation skills in L1 [7–10]. In order to gain a better understanding of phonetic attrition processes and their interaction with L2 acquisition, the present pilot study set out to examine AS's realization of voice onset time (VOT) contrast in word-initial plosives in his L1 and L2. While English distinguishes between short-lag VOT in voiced plosives, and long-lag VOT in voiceless plosives [11, 12], speakers of AG varieties tend to neutralize VOT contrast in word-initial bilabial and alveolar plosives in conversational, spontaneous speech, i.e. they produce both the voiced and voiceless targets with short-lag VOT and thus do not realize a voicing contrast [13]. Based on these differences between

[1] Austrian German, as defined in the present study, refers to the regional variety spoken in Thal, a municipality which is located four kilometers west of Graz-City. According to Wiesinger (1967), the variety of AG spoken in Graz and surrounding areas can be classified as west-Styrian, which belongs to the larger (South-)Bavarian dialect region.

English and AG, the aim was to determine whether and to what extent AS's realization of VOT contrast in his L1 and L2 has changed from his early to his late speech samples, i.e. whether VOT has moved towards more native-like production norms in his L2, and whether the realization of VOT contrast in his current L1 speech has changed compared to early L1 speech samples.

Materials and procedure

Speech data were taken from Austrian German (AG) and English interviews with AS which were broadcasted on various US and Austrian TV and radio channels. The AG and English corpora were divided into early (1979–1989) and late (2010–2018) speech samples, respectively.

The complete corpus underwent fully automatic phonological segmentation and labelling using WebMausBasic [14], a segmentation tool provided by the Bavarian Archive for Speech Signals (BAS) at the University of Munich. Monosyllabic and bisyllabic tokens containing a pre-vocalic, word-initial plosive in stressed or accented position were selected for analysis. Tokens which were affected by noise, speaker hesitation or speaker overlaps were discarded. As the sound recordings differed in length and quality, an unequal number of tokens was included representing AS's late and early L1 and L2 production of plosives (Tab. 1). Using oscillogram and sound spectrogram displays in the EMU-webApp [15], VOT was manually labelled from the burst of the plosive to the onset of periodicity indicating the start of the following vowel. Calculations of VOT durations were carried out in emuR [15].

	EN					
Early1979	p	b	t	d	k	g
Total: 242	31	32	56	40	39	44
Late2016	p	b	t	d	k	g
Total: 184	20	26	51	25	31	30
	AG					
Early1982/85/86	p	b	t	d	k	g
Total: 57	2	5	8	5	18	19
Late2012/13/17	p	b	t	d	k	g
Total: 60	–	17	8	20	5	10

Tab. 1: Overview of the number of tokens included for early and late English (EN) and Austrian German (AG) samples.

Results

The aim of the present study was to examine potential changes in Arnold Schwarzenegger's realization of VOT contrast by comparing his early and late AG and English speech.

Fig. 1 displays the mean VOT durations of plosives in AS's early English productions compared to his late L2 productions. The findings show that the mean VOT duration for all plosives in his L2 hardly changed from early to late productions, i.e. he realized and still realizes the voiceless targets with long-lag VOT and the voiced counterparts with short-lag values. As can be seen in Fig. 1, the mean VOT durations measured for the velar targets are longer than those measured for the bilabial and alveolar targets. This is according to expectation as VOT duration is influenced by place of articulation, i.e. the further back in the oral cavity the plosive is produced, the longer the VOT duration [11].

Furthermore, a comparison of his early and late VOT to native American English speaker values reported by [12] reveals that his mean VOT values are within native speaker norms (Fig. 2), even in his early data.

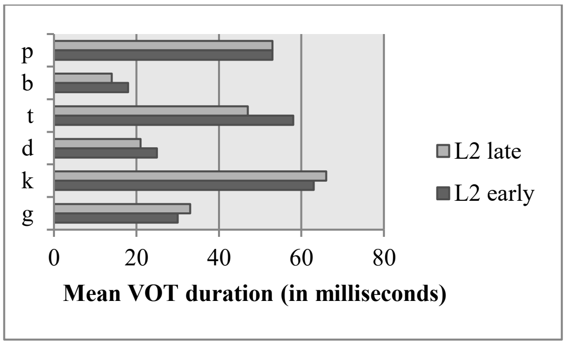


Fig. 1: Mean VOT durations in AS's early and late L2 productions.

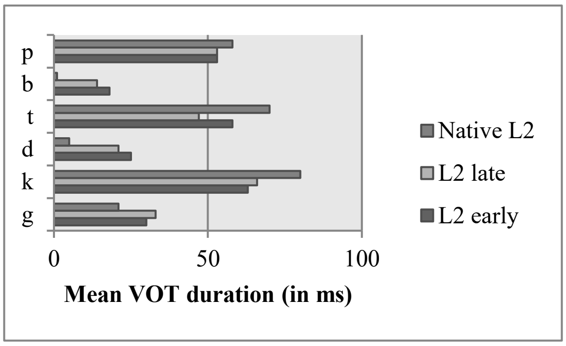


Fig. 2: Mean VOT durations in AS's early and late L2 productions compared to native speaker values [12].

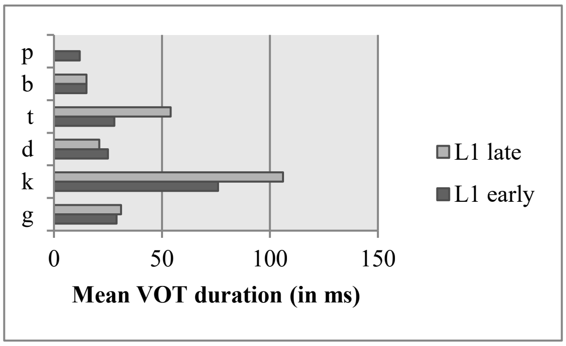


Fig. 3: Mean VOT durations in AS's early and late L1 productions.

In terms of AS's realization of VOT contrast in his L1, the findings indicate a change in mean VOT duration between early and late L1 productions. As depicted in Fig. 3, AS did not realize a VOT contrast in the bilabial and alveolar targets in his early AG and produced both the voiceless and the voiced plosives with short-lag VOT below 30 milliseconds (ms). For his late production of the voiceless bilabial plosive no values could be included at this point as there were no targets available in the speech material used for this pilot investigation.

When comparing the realization of VOT contrast in his early L1 productions of the voiced and voiceless alveolar plosive with his late productions, a change in mean VOT duration for the voiceless target can be observed. In his current L1 speech, AS produces the voiceless alveolar with long-lag VOT (54 ms on average) and the voiced counterpart with short-lag values (21 ms on average), i.e. he now realizes a voicing contrast. A comparison of his L1 VOT durations with L2 native speaker values [12] reveals that his VOT values have moved towards the norms of his L2 (Fig. 4) when it comes to the production of the alveolar targets.

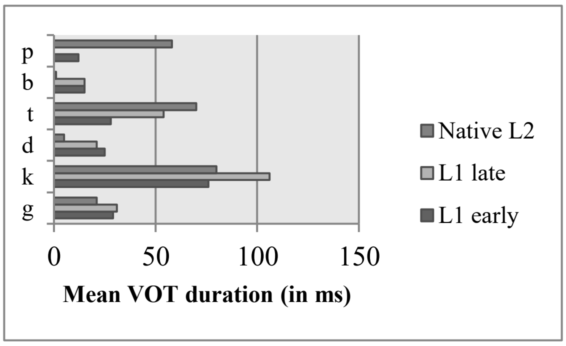


Fig. 4: Mean VOT durations in AS's early and late L1 productions compared to native speaker values [12].

As regards the realization of the velars, it can be seen that in his late L1 speech he lengthens the mean VOT duration for the voiceless velar target and, thus, produces this target with a higher degree of aspiration compared to his early productions.

Discussion

The main objective of this study was to determine if and to what extent AS's realization of VOT contrast in his L1 and L2 has changed since he moved to an L2-speaking environment in the 1960s. For this purpose, VOT durations in word-initial, pre-vocalic plosives were examined in early (1979–1989) and late (2010–2018) speech samples taken from English and AG interviews with AS.

A comparison of mean VOT durations in AS's early and late English plosives revealed that he realized and still realizes a VOT contrast in all his plosives. This observation is quite surprising, given that native AG speakers tend to eliminate VOT contrast in bilabial and alveolar plosives [13]. Thus, AS would have been expected to transfer the L1-characteristic of neutralizing VOT contrast to his L2 pronunciation

[16], at least at the early stages of his residency in the US. A possible explanation for his near native-like production of VOT could be that the speech sample representing his early English pronunciation is taken from an interview which was conducted in 1979, almost ten years after he moved to the US. Thus, it can be assumed that this ten-year-immersion in an English-speaking environment and regular contact with native L2 speakers improved his L2 pronunciation in terms of the realization of VOT contrast. As no further improvement can be observed between AS's early and late productions of VOT in his L2, it can be argued that this specific L2-feature has fossilized, i.e. he might have reached an end state of L2 acquisition [17]. In order to find out whether AS was able to produce native-like VOT in his L2 from the onset of L2 learning or whether this ability has developed over time and ceased to improve at one point, it will be necessary to find speech samples of AS's L2 pronunciation in the early 1960s.

When it comes to AS's L1 the findings show differences in terms of mean VOT durations in his early and late productions. As is typical for native AG speakers [13], he was observed to neutralize VOT contrast in his early alveolar and bilabial plosives and produced both the voiceless and voiced targets within a short-lag range. An investigation of his late AG productions, however, revealed that he now realizes a VOT contrast in the alveolar plosives by lengthening the mean VOT duration for the voiceless target. Even though no values could be included representing his production of the voiceless bilabial plosive, it can be assumed that AS also produces a VOT contrast in the bilabial targets in his late L1. Further investigations including more tokens will reveal whether this assumption proves to be valid. What can be concluded from these findings is that AS's L1 VOT-categories for the voiceless alveolar plosive – and presumably also for the voiceless bilabial target – have moved towards the norms of his L2, which can be interpreted as evidence for L1 attrition due to L2 learning experience. In fact, previous research into processes of L2 acquisition and L1 attrition has shown that long-term immersion in an L2-speaking environment is beneficial when it comes to acquiring native-like pronunciation skills in the L2, but this immersion might also contribute to the decline of pronunciation abilities in the L1 with regard to various segmental [8, 10] and suprasegmental [7, 9] features.

Further investigations into AS's realization of VOT contrast will focus on potential differences in mean VOT duration depending on the number of syllables (monosyllabic vs. bisyllabic) in the target items and on the quality of the vowel following the target plosive. However, in order to gain a more profound understanding of bidirectional L1-L2 influences operating in a late consecutive bilingual and to characterize the relationship between L2 acquisition and L1 attrition over an extended period of time, it will be necessary to examine further segmental and also suprasegmental features of AS's pronunciation in both of his languages.

References

[1] De Bot, K., "The psycholinguistics of language loss," in *Bilingualism and migration*, Extra, G. & L. Verhoeven (Eds.), Berlin: De Gruyter, pp. 345–361, 1998.

[2] Köpke, B. & M.S. Schmid, "Language attrition: The next phase," in *First Language Attrition: Interdisciplinary perspectives on methodological issues*, Schmid, M.S., Keijzer, M. & L. Weilemar (Eds.), Amsterdam: John Benjamins, pp. 1–46, 2004.

[3] Daily Mail UK, "Arnold Schwarzenegger reveals he can speak perfect English," *Daily Mail online*, 2015, available from: https://www.dailymail.co.uk/tvshowbiz/article-3141778/Arnold-Schwarzenegger-reveals-speak-perfect-English-keeps-talking-accent-fans-expect-it.html [02.10.2019].

[4] Quora, "Can Arnold Schwarzenegger speak German?," 2017, available from: https://www.quora.com/Can-Arnold-Schwarzenegger-speak-German [02.10.2019].

[5] Edwards, J.G. & M.L. Zampini (Eds.), *Phonology and second language acquisition*, Philadelphia, PA: John Benjamins, 2008.

[6] Piske, T., MacKay, I.R.A. & J.E. Flege, "Factors affecting the degree of foreign accent in an L2: a review," *Journal of Phonetics*, 29(2), pp. 191–215, 2001.

[7] De Leeuw, E., Mennen I. & J.M. Scobbie, “Singing a different tune in your native language: first language attrition of prosody,” *International Journal of Bilingualism*, 16(1), pp. 101–116, 2012.

[8] Mayr, R., Price S. & I. Mennen, “First language attrition in the speech of Dutch-English bilinguals: The case of monozygotic twin sisters,” *Bilingualism: Language and Cognition*, 15(4), pp. 687–700, 2012.

[9] Mennen, I., “Bi-directional interference in the intonation of Dutch speakers of Greek,” *Journal of Phonetics*, 32(4), pp. 543–563, 2004.

[10] Sancier, M.L. & C.A. Fowler, “Gestural drift in a bilingual speaker of Brazilian Portuguese and English,” *Journal of Phonetics*, 25(4), pp. 421–436, 1997.

[11] Docherty, G.J., *The Timing of Voicing in British English Obstruents*, Berlin, New York: Foris, 1992.

[12] Lisker, L. & A.S. Abramson, “A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements,” *Word*, 20(3), pp. 384–422, 1964.

[13] Moosmüller, S., Schmid, C. & J. Brandstätter, “Illustrations of the IPA: Standard Austrian German,” *Journal of the International Phonetic Association*, 45(3), pp. 339–348, 2015.

[14] Kisler, T., Reichel, U.D & F. Schiel, “Multilingual processing of speech via web services,” *Computer Speech and Language*, 45, pp. 326–347, 2017.

[15] Winkelmann, R., Harrington J. & K. Jänsch, “EMU-SDMS: advanced speech database management and analysis in R,” *Computer Speech and Language*, 45, pp. 392–410, 2017.

[16] Flege, J.E., “Interactions between the native and second-language phonetic systems,” in *An integrated view of language development: Papers in honor of Henning Wode*, Burmeister, P., Piske, T. & A. Rhode (Eds.), Trier: WVT, pp. 217–243, 2002.

[17] Schachter, J., “On the issue of completeness in second language acquisition,” *Second Language Research*, 6, pp. 93–124, 1990.

Voice quality of ironic utterances in standard Austrian German: preliminary results

Hannah Leykum[1]

[1]Acoustics Research Institute, Austrian Academy of Sciences, Vienna

hannah.leykum@oeaw.ac.at

Abstract

Most listeners are quite good at recognising verbal irony, even if the context is ambiguous. One reason for this is that many speakers apply disambiguating paraverbal cues to highlight being ironic. The present study investigates whether voice quality differences are used to mark irony by speakers of Standard Austrian German. The waveform characteristics of electroglottographic recordings are investigated to obtain objective measurements. Since only a small dataset is analysed in this study, the results are not yet generalisable.

Introduction

Irony is regularly used in everyday communication. To avoid misunderstandings, most speakers apply disambiguating cues to underline that an utterance is meant in an ironic sense. These cues for irony can be verbal, nonverbal, or paraverbal. If both verbal and nonverbal cues are missing in interaction (e.g. ambiguous utterances on the telephone), listeners have to rely on paraverbal cues in order to understand irony. The most prominent paraverbal cues to indicate verbal irony are average fundamental frequency (F0), F0-contour, intensity, and duration.

Acoustic analyses performed on German Germany utterances of sarcasm/ironic criticism showed a lower average F0, smaller F0 variation, a lower mean intensity, and longer segment durations in verbal irony compared to the literal counterparts [1–4]. Concerning voice quality, Niebuhr [2] found a more variable, mainly breathier or tenser, voice quality for ironic utterances. In addition, the study of Schmiedel [3] revealed a higher harmonics-to-noise ratio (HNR) for sarcastic utterances. Until now, no investigations of voice quality in ironic utterances of speakers of Standard Austrian German (SAG) have been done. In order to see whether it is worth investigating in more detail potential voice quality differences between ironic and literal realisations of utterances, the present study analyses a small subset of the available data. Electroglottographic (EGG) recordings are analysed to get objective measurements for voice quality. As described in [5], the shape of the EGG waveform differs between different voice qualities: e.g. the EGG of a modal voice shows a steep increase in the closing phase and an open quotient of about 50%; and a breathy voice is mainly characterised by a long open phase and a reduced amplitude [5].

Method

The present study is part of a larger project investigating on the one hand the acoustic characteristics of verbal irony and on the other hand the recognition of irony by listeners, when context-free stimuli are presented. Therefore, more material is recorded than analysed within this study. During the recording session, 20 speakers (two age groups, balanced for gender) of Standard Austrian German (SAG) [6] are presented with 20 scenarios that evoke ten short utterances either in the literal sense or in an ironic sense (an example is given in).

The speakers are instructed to read the scenarios and the answers to the scenarios while audio, video, and electroglottographic (EGG) recordings are conducted. For the present study, only the data of four speakers (two male, two female) is analysed. Out of the ten utterances, the following were chosen for the present analyses: *Danke!* (‘Thanks’), *Spannend!* (‘Exciting’), *Sehr gut!* (‘Very good’), and *Sehr schön!* (‘Very beautiful’).

ironic	literal
Person A gibt Person B einen dreckigen Putzfetzen und sagt: „Ich habe dir was mitgebracht.“ „Danke schön!“	Person A kommt lächelnd zur Tür rein und sagt: „Ich habe dir Blumen mitgebracht!“ „Danke schön!“
Person A gives person B a dirty floor cloth and says: “I’ve brought you something.” “Thank you!”	Person A smiles while entering the room and says: “I’ve brought you some flowers!” “Thank you!”

Tab. 1: Example scenarios for “Danke schön!”.

To analyse the data, in a first step, the EGG waveform was combined and temporally aligned with the first derivative of the EGG waveform (DEGG) (converted with Praat [7]) and with the acoustic recording (cf. Fig. 1). Afterwards, each glottal cycle was manually segmented into three phases: a closing phase (peak in the derivative), a contact phase + opening, and an open phase (as shown in Fig. 2). Due to the fact that in most recordings no negative peak (=opening instant) was detectable in the EGG derivative, the contact phase and opening were analysed as a single phase, which leads to the disadvantage that the results on the relative duration of the open phase are not comparable to open quotient analyses [e.g. 5] in other studies.

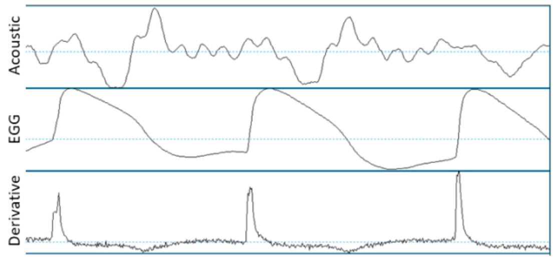


Fig. 1: Acoustic signal + EGG signal + derivative of the EGG signal (DEGG).

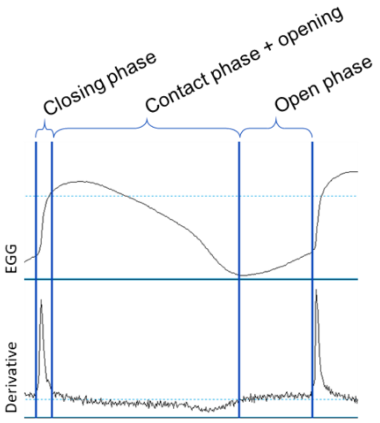


Fig. 2: Phases of one glottal cycle.

In order to see whether voice quality differs between ironic and literal realisations of utterances, the stressed vowel of each utterance was analysed with respect to the following parameters:

- Relative duration of the closing phase.
- Relative duration of the contact phase + opening.
- Relative duration of the open phase (= no-contact phase).
- Intensity of the EGG signal (in dB).
- Peak amplitude of the EGG derivative.
- Measurements of the harmonics-to-noise ratio (acoustic signal).

Results

Since each glottal cycle of the vowels was measured, enough data (1458 measurements) was available to conduct meaningful statistical analyses (by fitting mixed-effects models [8] in R [9]). The vowel duration as well as the number of glottal pulses per vowel were included in the analyses to avoid a confounding influence by these variables.

As regards the relative duration of the closing phase, a significant interaction was found between participant and manner-of-realisation (F(3, 471)=12.874. *p*<0.001). Tukey post-hoc tests showed significant differences in the closing duration for the two male speakers (m01: t(472)=3.824, *p*=0.004; m02: t(472)=4.867, *p*<0.001) and a tendency towards an effect for one female speaker (f01: t(474)=2.979, *p*=0.060). For these speakers, the duration of the closing phase was longer in the stressed vowels of ironic utterances compared to literal utterances. For the second female speaker, the differences between the manners of realisation were not significant (f02: t(472)=–2.591, *p*=0.162). These results are shown in Fig. 3.

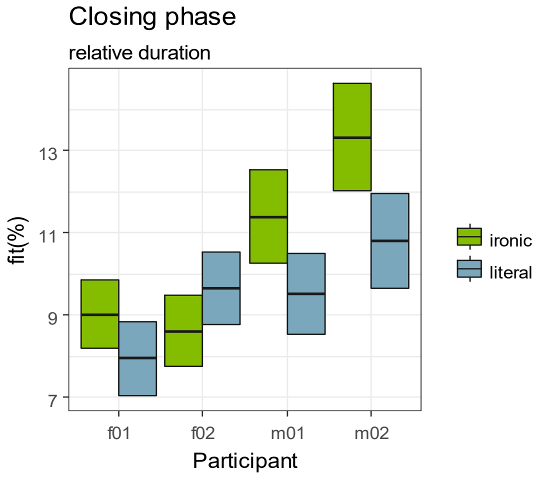


Fig. 3: Relative duration – Closing phase.

The analysis of the relative duration of the contact phase + opening shows a significant participant*manner-of-realisation interaction (F(3, 475)=4.258, *p*=0.006). Post-hoc analyses revealed a significant difference between ironic and literal realisation of the utterances only for speaker f01 (t(476)=3.936, *p*=0.002). For her, the duration of the contact phase + opening was longer in ironic utterances, as can be seen in Fig. 4.

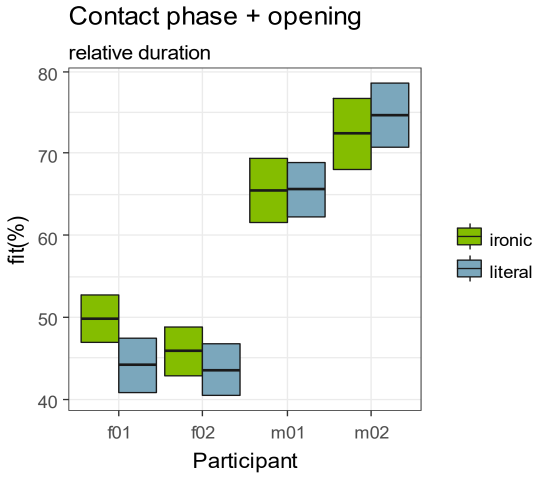


Fig. 4: Relative duration – Contact phase + opening.

Regarding the open phase, a significant participant*manner-of-realisation interaction (F(3, 474)=3.536, *p*=0.015) revealed in post-hoc tests a significant gender effect (all comparisons *p*<0.001, see Fig. 5); and significantly shorter durations of the open phases in literal realisation by participant f01 compared to her ironic reali-sations (t(473)=–4.708, *p*<0.001).

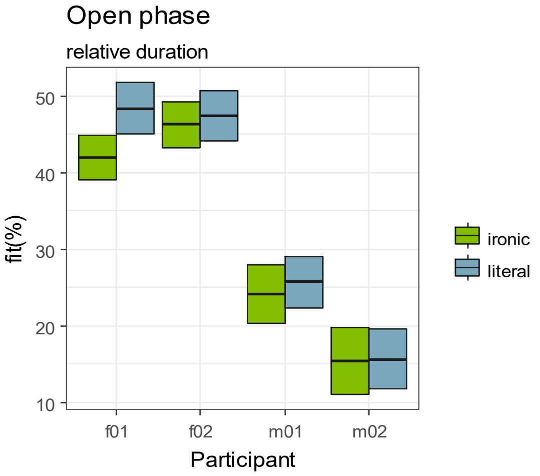


Fig. 5: Relative duration – Open phase.

For the mean intensity of the EGG signal, a significant gender*manner-of-realisation interaction (F(1, 28)=8.274, *p*=0.008) revealed neither for male speakers (t(25)=–2.161, *p*=0.162), nor for female speakers (t(25)=1.907, *p*=0.251) any statistically significant differences between ironic and literal realisations of the target utter-ances. However (as can also be seen in Fig. 6), the gender difference is larger for literal utterances (t(25)=–8.241, *p*<0.001) than for ironic ones (t(25)=–4.173, *p*=0.002).

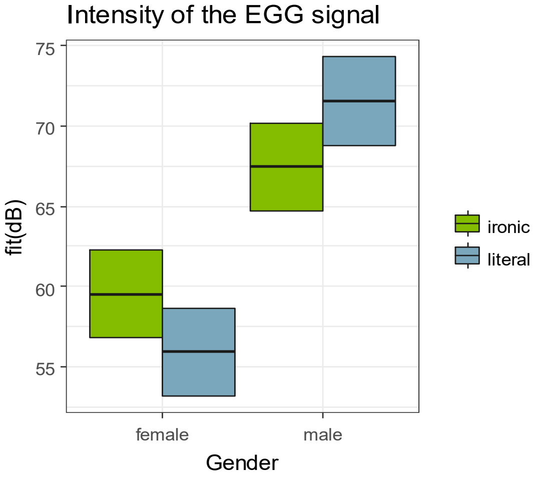


Fig. 6: Intensity of the EGG signal.

With respect to the amplitude of the peak in the derivative of the EGG signal (DEGG peak amplitude), the analyses showed a significant interaction between gender and manner of realisation (F(1, 545)=74.516, *p*<0.001). The results of Tukey post-hoc analyses indicated no differences in the DEGG amplitude between the two manners of realisation by the female speakers (t(500)=0.093, *p*=1.000), but a significant effect for the male speakers (t(447)=10.225, *p*<0.001), with a higher amplitude in literal utterances compared to ironic utterances (cf. Fig. 7).

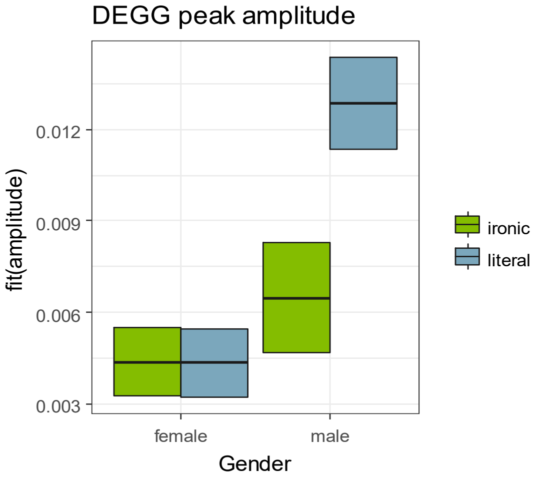


Fig. 7: DEGG amplitude.

Moreover, the harmonics-to-noise ratio was measured for each stressed vowel. Here, only differences between the individual speakers reached significance (F(3,23)=6.380, *p*=0.003). For none of the speakers did significant differences between literal and ironic realisations of the utterances emerge (cf. Fig. 8).

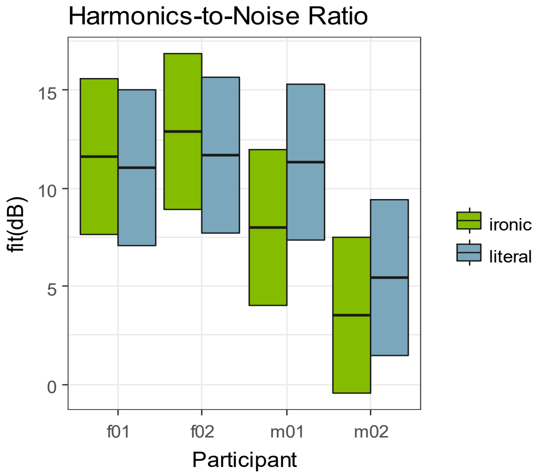


Fig. 8: Harmonics-to-noise ratio.

Summary and discussion

Differences between ironic and literal realisations of utterances were found in the relative duration of the closing phase, with a longer closing phase for ironic utterances when realised by male speakers. The contact phase + opening was longer in ironic utterances than in literal utterances only in the realisations of one of the speakers. For the other speakers, no significant differences occurred with respect to the manner of realisation. A longer closing phase results from a less steep increase in the EGG signal, indicating a slower closure of the vocal folds. This could point to a slightly tenser voice quality [5].

The relative duration of the no-contact phase (open phase) revealed a gender effect confirming that the voices of the female speakers were more breathy than those of the male speakers. Moreover, one female speaker showed a significant difference between literal and ironic realisations, indicating breathier realisation of the literal utterances.

Since the intensity values of the EGG signal were not normalised, the gender difference could have occurred solely because of differences in the recording level. However, the larger magnitude of the differences in the literal realisations of the utterances is interpretable. A more breathy voice quality results in a lower amplitude of the EGG signal. Therefore, these results point to a gender difference in the breathiness of literal realisations.

Concerning the amplitude of the DEGG signal, a significant difference between the manners of realisation was found only for male speakers. Here, the literal realisations had a higher peak compared to the ironic realisations. The DEGG peak amplitude gives information about the steepness of the closing phase (= closing speed of the vocal folds).

Unlike the results of Schmiedel [3], in the present study no significant differences occurred in the harmonics-to-noise ratio of ironic and literal utterances.

Outlook

Since voice quality characteristics of vocal fold movement consist of more than one single EGG waveform feature, the data also needs to be analysed with respect to combinations of several features in order to draw conclusions from the measurements. Moreover, more data (more speakers and more utterances) will be analysed to reduce the influence of speaker-specific differences.

References

[1] Scharrer, R. & U. Christmann, "Voice Modulations in German Ironic Speech," *Language and Speech*, 54(4), pp. 435–465, 2011.

[2] Niebuhr, O., "A little more ironic – Voice quality and segmental reduction differences between sarcastic and neutral utterances," *7th International Conference of Speech Prosody*, pp. 608–612, 2014.

[3] Schmiedel, A., P*honetik ironischer Sprechweise: Produktion und Perzeption sarkastisch ironischer und freundlich ironischer Äußerungen*, Berlin: Frank & Timme, 2017.

[4] Nauke, A. & A. Braun, "The production and perception of irony in short context-free utterances," *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*, pp. 1450–1453, 2011.

[5] Marasek, K., *Electroglottographic Description of Voice Quality*, Habilitation, University of Stuttgart, Stuttgart, 1997.

[6] Moosmüller, S., Schmid, C. & J. Brandstätter, "Standard Austrian German," *Journal of the International Phonetic Association*, 45(03), pp. 339–348, 2015.

[7] Boersma, P. & D. Weenink, Praa*t: doing phonetics by computer* [Computer program]. Version 6.0.43, retrieved 8 September 2018 from http://www.praat.org/ [04.02.2019].

[8] Bates, D. et al., "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67(1), pp. 1–48, 2015.

[9] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria; http://www.R-project.org/, 2015.

Wanted! – Assessing effects of distraction on working memory in speech perception

Matthias Wilke[1], Daniel Duran[2], Natalie Lewandowski[3]

[1]University of Stuttgart, [2]University of Freiburg, [3]High Performance Computing Center Stuttgart

matthias.wilke@ims.uni-stuttgart.de, daniel.duran@germanistik.uni-freiburg.de, natalie.lewandowski@hlrs.de

Abstract

With the technological advance of recent years and widespread social acceptance for video games, numerous studies have confirmed the usefulness of digital games in psychological experiments. A frequent commonality in previous studies is the attention-based test framework, applying distraction in some way in order to interfere with participants' cognitive capacities. Adopting this idea, we have developed the state-of-the-art first-person 3D video game 'Wanted!' *to examine the interactions between attention, cognitive load, working memory and speech processing. In a free recall task with auditorily presented items and visual recall, we apply varying combinations of auditory (speech and non-speech) and visual distraction during memorization and recall. We present the results of a pilot study with 16 German native speakers playing the game, where auditory distraction during memorization and recall was found to impair working memory performance by over 11% each. Visual distraction only had significant impact during memorization, with a performance deterioration of 7%, yet a tendency of visual distraction during recall affecting working memory was still apparent. Finally, we make numerous suggestions for harvesting the full potential of our game prototype in the future.*

Introduction

Among the copious advantages of games for psychological research is their interactivity, distinguishing them from other media such as film, still images or audio records. Some previous experiments using video games apply tasks "of alertness, orienting, and executive control" [1] or reaction tests [2] within virtual environments. There is, however, a lack of research that combines computerized games with memory tests. We therefore present a novel game framework to measure effects of distraction on memorization and recall of conversationally gained information. A comparison of these distraction modalities is conducted to identify visual effects analogous to the *irrelevant sound* or *unattended speech* effect [3], which describes the phenomenon of working memory deterioration during exposure to acoustic background noise.

Baddeley & Hitch's *working memory* model is presumed, dividing short-term memory into several sub-systems.

Generally, working memory is assumed to passively hold items for 10–15 seconds, which can be prolonged by active rehearsal [5]. As far as capacity is concerned, Miller [6] speculated on a working memory limit of seven, plus or minus two, items. This rule of thumb is also known as *Miller's law* among psychologists.

When conducting experiments about distraction, effects of selective attention need to be anticipated: While the irrelevant sound effect is known to impair recall efficiency quite reliably under various testing conditions [7], "selective focused attention allows people to successfully ignore irrelevant distractions" [8], leading to inattentional blindness. While focussing on a single, resource-intensive task, peripheral perception can be nearly eliminated [9].

Additionally, the universally accepted *two modes of thinking* [10] can play a role in experiments about attention: Highly automated tasks such as comprehending simple sentences are effortlessly attended to by system 1. System 2, on the other hand, requires active allocation of attention and involves complex decision-making tasks. Listening to a single person in noisy environments is an example for usage of system 2.

Another restraining factor of working memory is word length. However, the impact of word length on working memory capacity also depends on phonological complexity when items are presented auditorily [11].

Finally, serial recall performance can be greatly affected by *serial position* effects, facilitating memorization for initial and terminal elements of a list that is to be remembered.

The Game

In the game, players assume the role of the identikit expert's holiday replacement in a police station during the course of nine in-game days (each day resembles a *Trial*). It is a linear

3D game, viewed from a first-person perspective. The game's spoken and textual language was set to German. One playthrough consists of three phases:

Instruction and data acquisition

On an in-game computer screen, text windows provide advice on the game's controls, goals and story, and personal information has to be filled in.

Trials

Training

Players start each day by playing an arcade game similar to the classic arcade game *Space Invaders*, with the purpose of assessing participants' gaming proficiency.

Memorization

After the minigame, players can move freely within the police station. The first task is to meet up with a person who witnessed a crime. Upon player interaction, the witness starts talking and reveals details of a crime, during which the player is unable to walk. The interaction also triggers the first distraction. There are seven randomly allocated crime features that players must memorize.

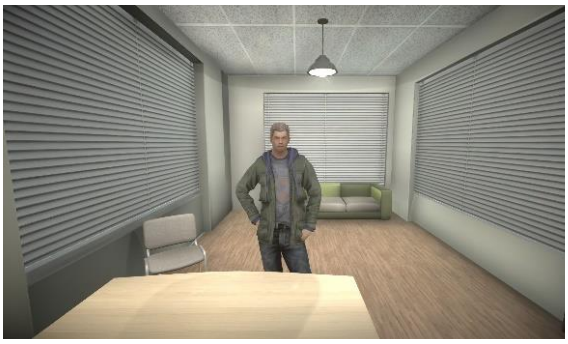


Fig. 1: Memorization without visual distraction.

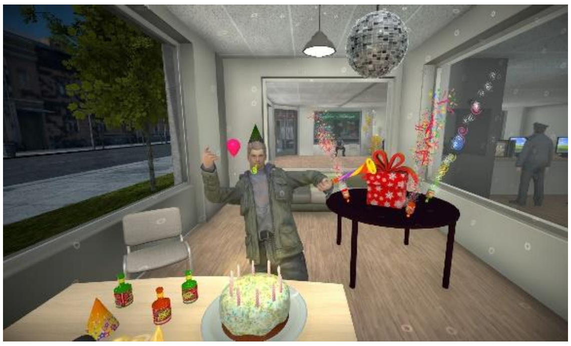
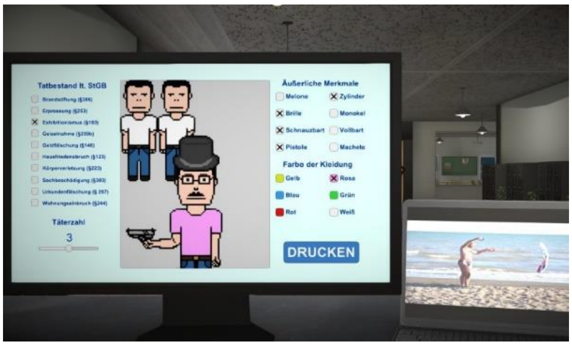


Fig. 2: Memorization with visual distraction.

Recall

As soon as the witness stops talking, distractions cease, and players need to walk to a computer. Interacting with the computer triggers the second distraction. Players report the memorized crime information in a multiple-choice test on the computer screen.



Tatbestand lt. StGB

- ☐ Brandstiftung (§306)
- ☐ Erpressung (§253)
- ☒ Exhibitionismus (§183)
- ☐ Geiselnahme (§239b)
- ☐ Geldfälschung (§146)
- ☐ Hausfriedensbruch (§123)
- ☐ Körperverletzung (§223)
- ☐ Sachbeschädigung (§303)
- ☐ Urkundenfälschung (§ 267)
- ☐ Wohnungseinbruch (§244)

Täterzahl

3

Äußerliche Merkmale

☐ Melone ☒ Zylinder
☒ Brille ☐ Monokel
☒ Schnauzbart ☐ Vollbart
☒ Pistole ☐ Machete

Farbe der Kleidung

☐ Gelb ☒ Rosa
☐ Blau ☐ Grün
☐ Rot ☐ Weiß

Fig. 3: Recall with visual distraction.

Intermission

Players end the wanted-poster creation by clicking the PRINT button on the virtual computer screen. The wanted poster then needs to be fetched from the printer room and pinned to a bulletin board, where the Trial score is shown.

Conclusion

After completing all Trials, a final score summary and a short story-related conclusion are given.

Materials and Tools

Our game was developed using the state-of-the-art game creation platform *Unity* [12] in its 2017 version, with all scripts written in C#. Many of our 3D models were retrieved royalty-free from *Free3D*, *CGTrader* and *TurboSquid*.

The majority of game sounds are based on files from *Freesound*. For recording the witness voice samples, we engaged an actor, who read out the sentences in an anechoic chamber. After each sentence, we inserted a one-second pause for the purpose of standardizing the memorization process.

We constructed *XML* files containing all session data such as mouse movements, scores and time measurements on the fly throughout the game. In order to skip time-consuming digitization of each player's feedback form, we created an online form to fill out after the game is finished. The feedback form contains questions about the game's design and level of difficulty, as well as the participant's subjective perception of distractions.

Evaluation

As a proof of concept, we had 16 German native speakers play the game (age range 19 to 32; 12 male and 4 female; none with reported hearing impairments).

For our evaluation, we recorded participants' training performance, Trial times, Trial score (correctly recalled crime features), and the answers from the online questionnaire.

Habituation effects and gaming experience were only found to affect training, but not Trial performance itself.

The error rate discrepancy between first and intermediate list items was not significant (*p*=0.262), but error rates for the last item deviated significantly from those of the first (*p*=0.031) and intermediate (*p*=0.008) items.

We identified a strong positive correlation of word length (in characters) and error rate (*r*=0.696, *p*=0.082) and a moderate positive correlation between syllable count and error rates (*r*=0.401, *p*=0.373). Even though both relations were insignificant, a tendency for word length (in characters) to affect error rates is apparent.

Phase	Distraction	Impairment Rating	Number of Ratings	Mean Trial Score
Memorization	auditory	<=3	10	6.57
		>3	6	5.89
	visual	<=3	8	6.42
		>3	8	5.91
Recall	auditory	<=3	8	6.58
		>3	8	5.79
	visual	<=3	14	6.40
		>3	2	5.67

Tab. 1: Trial score vs. subjective distraction ratings.

The effects of distraction on Trial performance were highly dependent on how strong the distractions were subjectively perceived by participants: Visual distractions with subjective impairment ratings >3 only had significant impact during memorization (deterioration: 7.85%, *p*=0.040), but not during recall (deterioration: 10.53%, *p*=0.145). The influence of auditory distraction with impairment ratings >3 was significant both during memorization (deterioration: 11.17%, *p*=0.005) and recall (deterioration: 12.03%, *p*=0.005).

Recall duration did not have significant effects on Trial score. However, the time that participants focused on the witness in the screen center during his monologue was decreased significantly when the subjects were distracted. The disparity between distraction types "none" (92.0%) and "visual" (80.7%) was highly significant (*p*=0.0098), whereas the mean values for distraction types "auditory" (88.0%) and "none" (*p*=0.149) as well as "auditory" and "visual" (*p*=0.082) did not differ significantly.

We measured a mean offset time of 25.48 seconds between memorization and recall, with a standard deviation of 10.78 seconds, meaning that active rehearsal was required to preserve the remembered items in working memory.

Discussion

Our game prototype was acclaimed for its understandability and player guidance, while maintaining a suitable level of difficulty for both beginners and experienced players. The rich game world, the detailed visuals and the imaginative implementation of distractions were highlighted positively by our players.

A plausible point of criticism is a lack of variability regarding the crime features to be remembered.

A crucial misconception of the game was the lack of a predefined structure for the offset phase between memorization and recall, observable in a slight correlation between offset duration and Trial score. This can be overcome in the future by standardizing the offset duration.

In two sessions, we encountered limitations of our video game: One participant closed her eyes as a way of concentration enhancement [13], ignoring any visual distraction in the process. Another player took his hands off the mouse, cancelling any interaction with the game.

Beginners and experts achieved very similar scores in the Alien Invasion minigame, but beginners outperformed advanced and expert players as regards total Trial score. The two cognitive systems mentioned above might serve as a possible explanation for this: People with no gaming experience do not interact with video games naturally, activating the more powerful system 2. Due to gaming experts' practice and experience, many video game related tasks can routinely be solved by system 1, which was manifested in high Alien Invasion scores. It seems that advanced players failed to activate system 2 rapidly enough when they encountered the memorization task.

When identifying which distraction stimuli were being perceived particularly intensively, *cross-modal interference* seemed to play a role. During the auditory memorization procedure, visual distraction was rated as more disruptive than auditory stimuli, and while visually occupied with recall, auditory distraction was perceived as more obstructive.

Apart from that, the distractions were varyingly invasive, causing inattentional blindness in some situations.

Conclusion

In conclusion, we have succeeded in reinforcing the irrelevant sound effect theory [3] in a virtual environment. In addition to that, we have managed to find evidence for effects of visual stimuli on sequence learning during memorization, but not during recall.

When focussing on word-length effects in future experiments, the list items should be chosen more uniformly in terms of semantic, syntactic and phonological complexity. The comparative examination of distraction modalities would likely benefit from the introduction of Trials with both auditory and visual distraction at the same time. Due to the modular nature of the game framework, distractions could even be swapped out effortlessly to examine specific matters such as the irrelevant sound effect for certain sound categories or languages. Even experiments with no distractions at all, solely manipulating the voice samples themselves, can be carried out with very few changes to the game.

In order to compare individual distractions instead of grouping them by modality, studies with a considerably higher number of participants will be required. A larger scale study could also facilitate investigating age effects and reassessing the impact of visual distraction during recall.

For the purpose of further analyzing habituation to the tasks in *'Wanted!'*, we propose daily or weekly repetition of the experiment.

Without any changes to the game itself, an eye tracker could be installed to monitor the focus of our players' pupils. A more costly option would be the use of virtual reality gear, masking out extrinsic influences on gameplay and thus maximizing immersion. Yet the effort and cost for a VR version of the game need to be weighed carefully, considering the convenience and affordability of eye trackers.

References

[1] Berger, A. et al., "Computerized games to study the development of attention in childhood," *Behavior Research Methods, Instruments, & Computers*, 32, pp. 297–303, 2000.

[2] Harbluk, J. L. et al., "Using the lane-change test (lct) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces," *Proceedings of the Fourth International Driving Symposium*, University of Iowa, 2007: 16-22.

[3] Farley, L. A. et al., "Irrelevant speech effects and sequence learning," *Memory & Cognition* 35, pp. 156–165, 2007.

[4] McLeod, S., *Working memory*. 2012. URL https://www.simplypsychology.org/working%20memory.html.

[5] Goldstein, E. B., *Cognitive psychology: Connecting mind, research and everyday experience*, Nelson Education, 2014.

[6] Miller, G. A., "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, 101, pp. 343–352, 1956.

[7] Jones, D. M. & W. J. Macken, "Organizational factors in the effect of irrelevant speech: The role of spatial location and timing," *Memory & Cognition*, 23, pp. 192–200, 1995.

[8] Konstantinou, N. et al., "Working memory load and distraction: dissociable effects of visual maintenance and cognitive control," *Attention, perception &psychophysics*, 76, pp. 1985–1997, 2014.

[9] Légal, J.-B. et al., "Beware of the gorilla: Effect of goal priming on inattentional blindness," *Consciousness & Cognition*, 55, pp. 165–171, 2017.

[10] Kahneman, D., *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, 2011.

[11] Service, E., "The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration," *The Quarterly Journal of Experimental Psychology*, 51(2), pp. 283–304, 1998.

[12] Unity Technologies SF. Unity Engine, October 2018. URL https://unity3d.com/.

[13] Vredeveldt, A., Hitch, G. J. & A. D. Baddeley, "Eye closure helps memory by reducing cognitive load and enhancing visualisation," *Memory & Cognition*, 39, pp. 1253–1263, 2011.

AUTOR*INNENVERZEICHNIS