MARTIN-LUTHER-UNIVERSITÄT
HALLE-WITTENBERG

20. Jahrestreffen für Phonetik und Phonologie
im deutschsprachigen Raum

**Sven Grawunder & Neda Mousavi (Hrsg.)**

# P&P20 – 2024
## in Halle (Saale)

# Tagungsband

Abteilung für Sprechwissenschaft und Phonetik
Institut für Musik-, Medien- und Sprechwissenschaften

**Hallesche Phonetische Studien 01**

# P&P20 - 2024

# Tagungsband

zum

*20. Jahrestreffen*
*Phonetik und Phonologie im deutschsprachigen Raum*

1. & 2. Oktober 2024 in Halle (Saale)

Sven Grawunder & Neda Mousavi (Hrsg.)


Martin-Luther-Universität Halle-Wittenberg
Hallesche Phonetische Studien 01

**Redaktion:**
Sven Grawunder & Neda Mousavi
Abt. Sprechwissenschaft und Phonetik
Institut für Musik-, Medien- und Sprechwissenschaften
Martin-Luther-Universität Halle-Wittenberg
Emil-Abderhalden-Str. 26/27
06108 Halle (Saale)

**Co-Review:**
Ines Bose
Alexandra Ebel
Anna Wessel
Julia Werth

**Publikation** durch die
Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB)

http://dx.doi.org/10.25673/116710



# Danksagung

Liebe Freundinnen und Freunde der P&P-Tagungen,
liebe Kolleginnen und Kollegen,

Das *20. Jahrestreffen für Phonetik und Phonologie im deutschsprachigen Raum* - die P&P-Tagung - fand im Oktober 2024 zum ersten Mal an der Martin-Luther-Universität in Halle an der Saale statt. Die Abteilung Sprechwissenschaft und Phonetik fühlte sich sehr geehrt, diese so einzigartige und angenehme Tagung der phonetischen Community beherbergen zu dürfen.
Wie gewohnt thematisch breitgefächert und insbesondere an den wissenschaftlichen Nachwuchs gerichtet, wurden Fragen zu einzelnen Segmenten oder prosodischen Eigenschaften in diversen sozio-demographischen, dialektalen, lexikalischen, semantischen und funktionalen Kontexten der Sprech- und Sprachproduktion und -perzeption von Gruppen und Individuen bearbeitet. Der vorliegende Band bildet eine (Selbst-)Auswahl von 27 der insgesamt angenommenen 65 Tagungsbeiträge.
Mein besonderer Dank gilt Neda Mousavi, die maßgeblich im Vorfeld zur Organisation und Planung der Tagung beigetragen hat und daher auch Mitherausgeberin dieses Bandes ist. Desweiteren haben Anna Wessel, Julia Werth und Babett Taubert wertvolle und tatkräftige Hilfestellung bei der Vorbereitung der Tagung geleistet.
Zudem möchten wir auch die studentischen Hilfskräfte Lara Höffner, Julia Skilandat, Eva Ritzer und Lucienne Wilkes nennen, die zum Gelingen der Konferenz beigetragen haben und denen wir dafür von ganzem Herzen danken.
Schließlich danke ich allen Beitragenden zu diesem Band für ihre Geduld, war doch das Erscheinen viel früher geplant.

Sven Grawunder
Halle/Saale, Oktober 2025

# Inhaltsverzeichnis

# Beiträge

# THE PROSODIC DEVELOPMENT OF THE *SORT OF*/ *KIND OF*/ *TYPE OF*-CONSTRUCTION IN ENGLISH

*Nicole Benker*[1,2]

[1]*Ludwig-Maximilians-Universität München, Department für Anglistik und Amerikanistik*
[2]*Graduate School Language and Literature Munich*
*Nicole.Benker@campus.lmu.de*

**Abstract:** The present study represents a diachronic look at the relationship between prosody, specifically prosodic prominence patterns, and grammaticalization. This is done by investigating the example of the *sort of*/ *kind of*/ *type of*-construction (SKT-construction), whose prosodic realization has so far only been investigated synchronically in a diachrony-in-synchrony approach [1]. The basis for the present investigation is the *parallel reduction theory* [2], which claims that semantic bleaching, as is typical of grammaticalization, goes hand in hand with phonetic reduction. The present study aims to (re-)investigate the question if a higher degree of grammaticalization correlates with a reduction in prosodic prominence not only synchronically but also diachronically by comparing data from two corpora that cover two different timeframes, the original *London-Lund Corpus* (LLC1, [3]), which comprises data (mostly) from the 1960s and 1970s, and the *London-Lund Corpus 2* (LLC2, [4]), with data from the 2010s. The investigation revealed that the SKT-construction as a whole grammaticalized further between the two data sets, a change that mostly affected *kind of*. However, this higher degree of grammaticalization does not correlate with a reduction in prosodic prominence as was predicted by the parallel reduction theory. It remains unclear why this is the case.

## 1 Introduction

The *sort of*/ *kind of*/ *type of*-construction (SKT-construction) is highly frequent in the English language and has the structure *SKT of X*. While the construction is now most often used as a pragmatic marker (e.g., *This song is kind of weird*.), it originally denoted class membership (e.g., *Rye is a kind of grain*.). Because of its frequency, the SKT-construction has been studied from a variety of viewpoints: There are synchronic studies [e.g., 1, 5, 6, 7], diachronic studies [e.g., 8, 9-11], cross-linguistic studies [e.g., 12, 13], investigations into the construction's structure [e.g., 8, 9, 14, 15-18], its pragmatic function [e.g., 5, 19], and some prosodic descriptions [e.g., 1, 5, 20, 21].

The grammaticalization cline (cf. Table 1) of the construction is commonly described as follows (terminology following Denison [18]): The first stage, the binominal stage, denotes class membership and comprises two noun phrases (NPs) connected by the preposition *of* (NP [of NP]), with *of NP* acting as a postmodifier of the first NP, i.e., the SKT-element [1, 11, 16, 20]. The second stage of development, the qualifying stage, involves reanalysis of the construction's underlying structure, in which the elements are rebracketed [22] in such a way that the preposition *of* becomes more closely associated with the SKT-element than the second NP ([NP of] NP). Functionally, the qualifying stage is no longer used to describe a hyponymic relationship but rather functions as a hedge or downtowner with the meaning 'almost' or 'nearly' [1, 11, 20], e.g. *He is kind of a friend*. The qualifying stage is often subdivided based on structural criteria, i.e., use of determiners in e.g., *He is a kind of a friend* or lack of number agreement between the NPs, e.g., *these sort of rules* [13, 20, 21]. Following Denison [18], these examples will be categorized as instances of the qualifying stage without further subdivisions. The third and last stage that will be discussed in the present study, the adverbial stage, is characterized

by host-class expansion [23]. This means that *SKT of* no longer only collocates with nouns but also with other parts-of-speech, such as verbs, adjectives or adverbs. Functionally, the adverbial stage is used as a pragmatic marker expressing doubt or uncertainty [1, 5, 11, 12]. Because the second NP is no longer considered an NP at this stage, the second element will be referred to as X in the rest of the study. Traugott [11] identifies a stage of grammaticalization beyond the adverbial stage, in which *SKT of* is used as a free adjunct, either right-peripherally, e.g., *I know what you mean, sort of*, or by itself, e.g., *Yes, sort of*. This stage will not be discussed in the present study. As is common for grammaticalization processes, older forms do not disappear completely but rather coexist with newer ones, a principle known as *layering* [24], which forms the basis of the diachrony-in-synchrony principle.

**Table 1** - Summary of the stages of grammaticalization of the SKT-construction [1, 11, 18] and their prosodic realization [1, 20].

| **Stage** | **Binominal stage** | **Qualifying stage** | **Adverbial stage** |
|---|---|---|---|
| **Form** | SKT [of NP] | [SKT of] NP | [SKT of] X |
| **Meaning/ Function** | group membership | 'almost'<br>downtowning, hedging | expressing doubt<br>hedging |
| **Prosody** | Stress on SKT<br>Stress on both | Stress on NP | unstressed<br>Stress on X |
| **Example** | *Rye is a kind of grain.* | *He is kind of a friend.* | *It is kind of strange.* |

These stages of grammaticalization have been correlated with patterns of prosodic prominence by both Keizer [20], in a qualitative manner, and Dehé & Stathi [1], in a synchronic quantitative study (cf. Table 1). Dehé & Stathi [1] use the *parallel reduction theory* [2] as the basis for their investigation, which describes the tendency for grammatical items (i.e., more semantically bleached items) to be shorter (i.e., more phonetically reduced) than lexical items. Both semantic bleaching and phonetic reduction tend to increase as items become more frequent [e.g., 2, 25]. Using this theory as the background, they showed that binominal uses of the SKT-construction, which are also its rarest variant, are associated with prominence on the SKT-element (~30% of attestations) or prominence on both SKT and X (~50%). The qualifying stage is associated with prominence marking on X (~90%) and the adverbial stage, which was the most frequent stage in [1], is either entirely unstressed (~40%) or exhibits prominence on X (~60%).

The present study constitutes a replication study of [1], which established a correlation between the stage of grammaticalization of the SKT-construction and prosodic prominence of its elements synchronically under the assumption of the diachrony-in-synchrony principle. The present study aims to investigate this correlation in a truly diachronic study, comparing data from the 1960s/70s with data from the 2010s. It is expected that if the SKT-construction increases in frequency from the LLC1 to the LLC2, this will correlate with higher degrees of grammaticalization, i.e., more adverbial uses, and a reduction of prosodic prominence.

## 2 Material and Method

The material that was used for the present study comprises two spoken corpora of southern British English, the original *London-Lund Corpus* (LLC1) [3] and the *London-Lund Corpus* 2 (LLC2) [4]. The LLC1 is a prosodically annotated corpus with material recorded from 1961 to 1984 (though the later 13 texts are supplements and were not included in the original release

and the majority of texts are from the 1960s and 1970s). The present study focuses on the face-to-face-conversation portion of the LLC1, which consists of 38 texts (~190,000 words). The conversations were recorded at the University of London and mainly feature conversations between students and/or university staff. This means that conversation topics tend to be of a more academic nature and because of that the language used might be more formal than in more casual face-to-face conversations. Unfortunately, the audio recordings for the LLC1 are not commonly accessible, so only the prosodic tags for pitch (slashes for tonics; underscores, colons and exclamation points for varying degrees of pitch increases) and intensity (quotation marks) were used in the present study. Sections that did not contain prosodic tags were excluded from further analysis

The LLC2 was created specifically to allow for comparison between the LLC1 and LLC2, i.e., the LLC2 has a similar sociodemographic composition and similar text genres as the original corpus. The material featured in the LLC2 was recorded between 2014 and 2019 at the University of London. At the time the present study was conducted, it was possible to analyze 29 full or partial texts of the face-to-face-conversation portion of the corpus. Each text contains about 5,000 words each (~95,000 words in total were available at the time of the study). In cases where a text was split into two, e.g., because it contained conversations by the same speakers at a different time, the two texts combined contain 5,000 words. The LLC2 is not prosodically annotated but instead is composed of .xml files that are aligned with .wav audio files, which were used for acoustic and auditory analyses. Acoustic analyses were conducted in Praat [26] and comprised measurements of maximum pitch and intensity of the two NPs in their entirety (including determiners and/or premodifiers). It was not possible to use all texts because of excessive background noise in some of the files. Only counting the texts that contained usable material, the LLC2 data used for the prosodic analysis comprised ~77,500 words/ 465 mins.

The analysis was conducted in two parts: For the first part, instances of the SKT-construction were extracted using AntConc [27] and then categorized as one of the three stages of grammaticalization based on the criteria laid out in [18]. SKT-elements with completely different meaning (i.e., the verbs *to sort* and *to type* or the adjective *kind*), uses as free adjuncts, or cases in which X was filled with *thing* or *stuff* (which is a special subform of the SKT-construction that will not be discussed in the present paper) were excluded. This left 733 attestations in the LLC1 (15 for *type of*, 99 for *kind of* and 619 for *sort of*) and 400 attestations in the LLC2 (6 for *type of*, 252 for *kind of* and 142 for *sort of*). The second part of the analysis comprised using the prosodic tags and measurements to identify the prominence patterns (as identified in [1]) with which the individual tokens were produced. For this analysis only those tokens from texts with sufficient audio quality or prosodic tags could be used, which left 534 tokens from the LLC1 (15 for *type of*, 76 for *kind of* and 443 for *sort of*) and 146 tokens from the LLC2 (2 for *type of*, 78 for *kind of* and 66 for *sort of*).

## 3 Results

### 3.1 Frequency of use and degree of grammaticalization

In the LLC1, *sort of* is the overwhelmingly most frequent SKT-element in the SKT-of-X-construction (32.6 attestations per 10,000 words), followed by *kind of* (5.2) and then *type of* (0.8), cf. Table 2 (yellow columns). *Sort of* is also the most grammaticalized SKT-element, i.e., most attestations are used adverbially, while *kind of* is mostly used binominally and *type of* is only used binominally. In the LLC2, *kind of* is the most frequent SKT-element (32.5 attestations per 10,000 words), followed by *sort of* (18.3) and *type of* (0.8), cf. Table 2 (orange columns). *Kind of* and *sort of* are most often used adverbially in the LLC2 followed by qualifying and then binominal uses, while *type of* is only used binominally.

**Table 2** - Comparison of frequency of use in the LLC1 (yellow) and LLC2 (orange) per stage of grammaticalization and SKT-element. Numbers per 10,000 words.

| | Binominal stage | | Qualifying stage | | Adverbial stage | | total | |
|---|---|---|---|---|---|---|---|---|
| **Corpus** | **LLC1** | **LLC2** | **LLC1** | **LLC2** | **LLC1** | **LLC2** | **LLC1** | **LLC2** |
| *type* | 0.8 | 0.8 | - | - | - | - | 0.8 | 0.8 |
| *kind* | 3.1 | 5.7 | 1.5 | 9.8 | 0.6 | 17.0 | 5.2 | 32.5 |
| *sort* | 5.2 | 2.1 | 10.9 | 5.4 | 16.5 | 10.8 | 32.6 | 18.3 |
| **total** | 9.1 | 8.6 | 12.4 | 15.2 | 17.1 | 27.8 | 38.6 | 51.6 |

The SKT-construction overall is more frequent in the LLC2 (51.6 attestations per 10,000 words) than in the LLC1 (38.6) and also shows a higher degree of grammaticalization, cf. Table 2. There are slightly fewer attestations in the binominal category (from 9.1 to 8.6) but more in the qualifying (12.4 to 15.2) and especially adverbial category (17.1 to 27.8). These higher frequencies are not distributed equally across all SKT-elements but are mostly due to *kind of* being massively more frequent in the LLC2 (from 5.2 in the LLC1 to 32.5), while the frequency of *sort of* is cut almost in half (from 32.6 to 18.3). The frequency of *type of* does not differ between the data sets.

Moreover, *kind of* shows a higher degree of grammaticalization in the LLC2 compared to the LLC1. In the LLC1, *kind of* is mostly used in the binominal stage (3.1 attestations per 10,000 words, compared to 1.5 in the qualifying stage and 0.6 in the adverbial stage), whereas it is mostly used in the adverbial stage in the LLC2 (17.0 attestations compared to 5.7 in the binominal stage and 9.8 in the qualifying stage). The use of *sort of* does not significantly differ between the two corpora but is used slightly more often in the adverbial stage and less often in the binominal stage in the LLC2 compared to the LLC1. While this indicates that *sort of* has grammaticalized further, this does not constitute a large functional change, since *sort of* was already used mostly adverbially in the LLC1. *Type of* shows no signs of grammaticalization between the two corpora and is only used binominally.

## 3.2 Prosodic prominence and grammaticalization

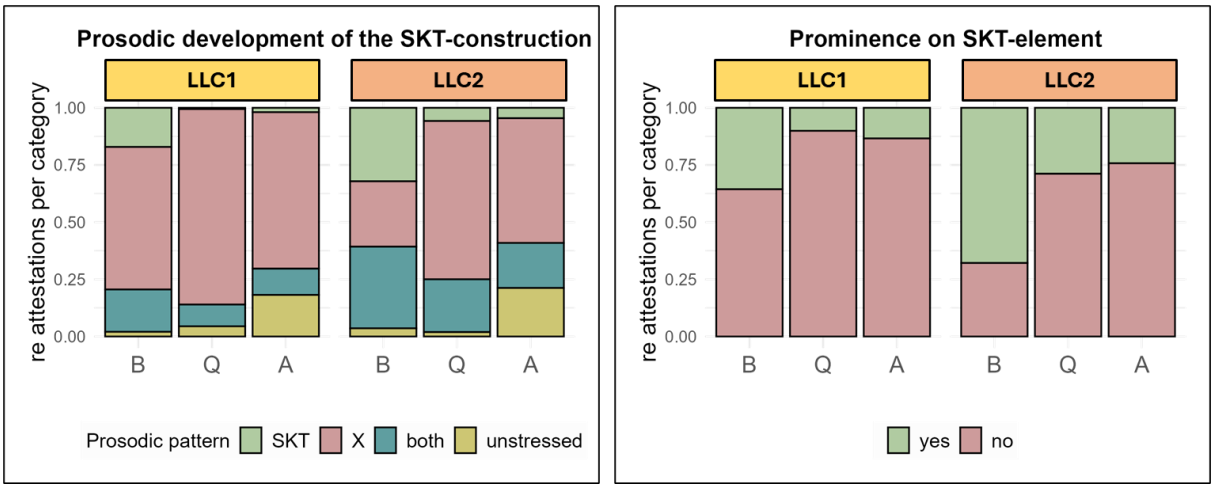



**Figure 1** - Prosodic patterns (as described in [1]) and stages of grammaticalization (B = binominal, Q = qualifying, A = adverbial) for all SKT-nouns (left panel) and correlation of prominence on the SKT-element and stage of grammaticalization (right panel) for all SKT-nouns in the LLC1 (yellow) and the LLC2 (orange).

In the next step of the analysis, the attestations of the SKT-construction were classified according to their stress patterns identified in [1], based on whether the SKT-element and/or the second element X were produced with prosodic prominence. Then, the stages of grammaticalization were correlated with these prosodic patterns, cf. Figure 1 (left panel). Overall, stress on X is the most common prosodic pattern and thus strongly correlates with all stages of grammaticalization. Since English favors end-focus, this result is unsurprising. Stress on the SKT-element and stress on both elements most strongly correlate with the binominal stage, while unstressed realizations correlate most strongly with the adverbial stage. These correlations can be found in both data sets. A comparison between the two corpora revealed that the attestations taken from the LLC2 were more prosodically prominent than those from the LLC1, which does not match the expectations, as higher degrees of grammaticalization should correlate with less prosodic prominence. This increase in prosodic prominence is especially apparent for the binominal stage, in which the SKT-element is more prominent in the LLC2 than in the LLC1 (cf. Figure 1, right panel). This increase in prominence can also be observed for the qualifying and the adverbial stage but to a lesser degree. Furthermore, X is less prominent in the LLC2 than in the LLC1. This is again most apparent for the binominal stage, but is also apparent for the qualifying and adverbial stages, albeit to a lesser degree (cf. Figure 1, left panel, comparing the sum of the blue and pink bars).

### 3.3 Development of individual nouns

The higher levels of prominence in the LLC 2 identified in the previous section are not equally distributed among all SKT-elements. Because *type of* is generally very rare in the SKT-construction and does not grammaticalize (i.e., only occurs in the binominal construction), it will not be discussed in this section. *Sort of*, which was the most common SKT-element in the LLC1 and also the most grammaticalized one, does not significantly differ in terms of prosodic realization between the two corpora, cf. Figure 2 (left panel). If anything, the attestations in the LLC2 show slightly less prosodic prominence which correlates with the slightly higher degree of grammaticalization mentioned in section 3.1.

The SKT-element that shows the most significant differences between the LLC1 and LLC2, and is thus responsible for the higher levels of prosodic prominence in Figure 1, is *kind of*. Even though *kind of* is much more frequent in the LLC2 (cf. section 3.1), it is much more prosodically prominent in the LLC2 than the LLC1 (cf. Figure 2, right panel). Possible reasons for this unexpected result will be discussed in the next section.

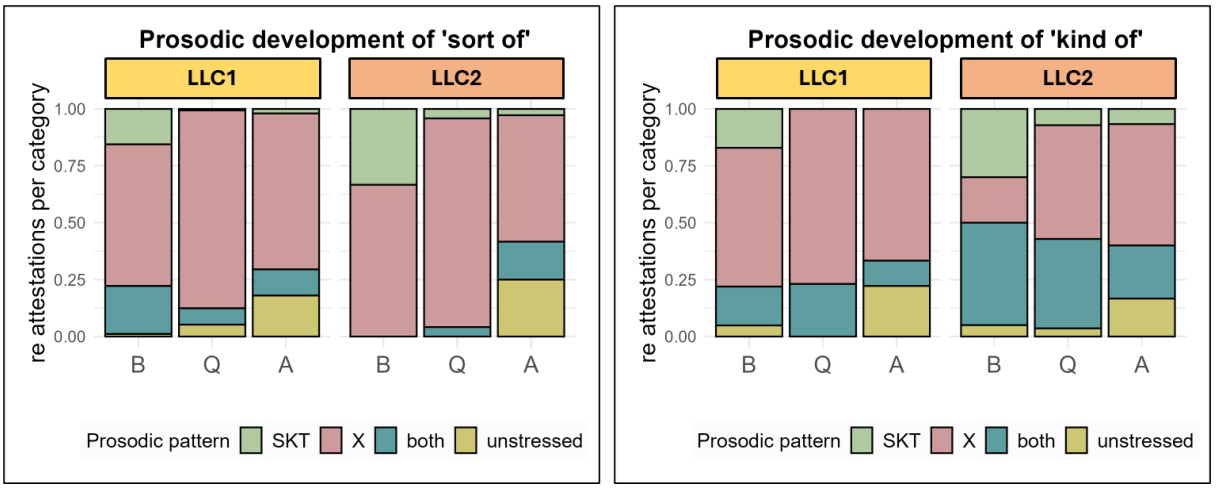



**Figure 2** - Prosodic development of the individual SKT-elements proportionally re stage of grammaticalization (B = binominal, Q = qualifying, A = adverbial): *sort of* (left panel) and *kind of* (right panel) in the LLC1 (yellow) and the LLC2 (orange).

## 4 Discussion

The present study investigated the correlation between stage of grammaticalization and prosodic prominence patterns of the SKT-construction in southern British English on the basis of the parallel reduction theory [2] in two comparable data sets covering different time frames: the LLC1, which contains data mostly from the 1960s and 1970s, and the LLC2, which contains data from the 2010s. This was done to confirm findings from previous work [1, 20] not only synchronically, according to the diachrony-in-synchrony principle, but also diachronically.

Between the LLC1 and the LLC2, the SKT-construction as a whole increased in frequency of use and also further grammaticalized, which is principally true for *kind of*, which is used much more frequently in the LLC2. Furthermore, although stress on X is overall much more common in the present study, the correlations between the prosodic patterns and stages of grammaticalization identified in [1, 20] could also be found in the two LLC corpora. This is especially true when the two corpora are regarded separately in a diachrony-in-synchrony approach: binominals are more prominent than adverbials in the LLC1 as well as the LLC2. However, the comparison between the two data sets calls into question the veracity of this approach. According to parallel reduction theory the higher frequencies and correlated higher degrees of grammaticalization in the LLC2 should mean less prosodic prominence, which was not found. In fact, unexpected higher levels of prominence for *kind of* in the LLC2 could be observed.

This increase in prominence is not observable for *sort of*, which indicates that the results for *kind of* are unlikely to be the result of technical differences between data sets and analyses (i.e., the fact that the basis for the prosodic analyses of the LLC1 data are prosodically annotated transcripts, not the actual audio recordings). Technical differences between data sets should affect all analyzed tokens, not only *kind of*. This means that the difference between *kind of* and *sort of* possibly has something to do with aspects on a structural or a functional level that were not investigated in the present study (e.g., use of demonstrative determiners, which are more often emphasized, compared to indefinite articles, which are often unaccented, or use in different pragmatic contexts). However, since these aspects were not analyzed in the present study, it cannot be said for certain what the cause of the discrepancy between the expected and actual results is. It is also possible that prosodic realization and degree of grammaticalization have a less linear relationship than previously thought. However, this claim warrants further investigation.

## 5 Conclusion and Outlook

In sum, the results of the present study is partially inconsistent with previous findings as regards the parallel reduction of semantic and phonetic bulk during the grammaticalization of the SKT-construction. However, previous studies were either qualitative diachronic accounts [20] or operated under the diachrony-in-synchrony principle [1]. Furthermore, there are several aspects that neither previous studies nor the present study have taken into account. Previous work comprised a fairly broad functional and formal analysis. But perhaps a more detailed functional analysis, including the usage context and a more detailed structural analysis might be necessary to explain why the results of the present study seemingly do not match the expected results. These more fine-grained analyses need not only take into account the situational and pragmatic context (as far as the information is available) but also the structural composition of the NPs in the construction.

## Acknowledgements

Many thanks to Nele Põldvere for access to the *London-Lund Corpus 2*.

## List of references


[1] DEHÉ, N. AND K. STATHI, "Grammaticalization and prosody: The case of English sort/kind/type of constructions," *Language*, vol. 92, no. 4, pp. 911–946, 2016, doi: 10.1353/lan.2016.0077.

[2] BYBEE, J., R. PERKINS, AND W. PAGLIUCA, *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press, 1994.

[3] SVARTVIK, J., R. QUIRK, S. GREENBAUM, AND K. HOFLAND. *London-Lund Corpus of Spoken English*, University College London.

[4] PÕLDVERE, N., V. JOHANSSON, AND C. PARADIS. *The London-Lund Corpus 2 of Spoken British English*, Lund University.

[5] AIJMER, K., *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins, 2002.

[6] FETZER, A., "Hedges in context: Form and function of 'sort of' and 'kind of'," in *New Approaches to Hedging*, G. KALTENBÖCK, W. MIHATSCH, AND S. SCHNEIDER Eds. Leiden: Brill, 2010, pp. 49–72.

[7] MARGERIE, H., "On the rise of (inter)subjective meaning in the grammaticalization of kind of/kinda," in *Subjectification, intersubjectification and grammaticalization*, K. DAVIDSE, L. VANDELANOTTE, AND H. CUYCKENS Eds. Berlin: DeGruyter, 2010, pp. 315–348.

[8] BREMS, L., *Layering of Size and Type Noun Constructions in English*. Berlin: DeGruyter, 2012.

[9] BREMS, L. AND K. DAVIDSE, "The grammaticalisation of nominal type noun constructions with 'kind/sort of'," *English Studies*, vol. 91, no. 2, pp. 180–202, 2010, doi: https://doi.org/10.1080/00138380903355023.

[10] DAVIDSE, K., "Complete and sort of: From identifying to intensifying?," *Transactions of the Philological Society*, vol. 107, pp. 262–292, 2009, doi: https://doi.org/10.1111/j.1467-968X.2009.01225.x.

[11] TRAUGOTT, E. C., "Grammaticalization, constructions and the incremental development of language: Suggestions from the development of degree modifiers in English," in *Variation, selection, development: Probing the evolutionary model of language change*, R. ECKARDT, G. JÄGER, AND T. VEENSTRA Eds. Berlin: DeGruyter, 2008, pp. 219–250.

[12] AIJMER, K., "Sort of and kind of from an English-Swedish perspective," in *Voices Past and Present: Studies of Involved, Speech-Related and Spoken Texts*, E. JONSSON AND T. JONSSON Eds. Amsterdam: John Benjamins, 2020.

[13] JANEBOVÁ, M. AND M. MARTINKOVÁ, "NP-internal kind of and sort of: Evidence from an English-Czech parallel translation corpus," in *Contrasting English and Other Languages through Corpora*, M. JANEBOVÁ, E. LAPSHINOVA-KOLTUNSKI, AND M. MARTINKOVÁ Eds. Newcastle upon Tyne: Cambridge Scholars, 2017.

[14] AARTS, B., "Binominal noun phrases in English," *Transactions of the Philological Society*, vol. 96, no. 1, pp. 117–158, 1998.

[15] BREMS, L., "Size noun constructions as collocationally constrained constructions: lexical and grammaticalized uses," *English Language and Linguistics*, vol. 14, no. 1, pp. 83–109, 2010, doi: https://doi.org/10.1017/S1360674309990372.

[16] DENISON, D., "History of the 'sort of' construction family," presented at the 2nd International Conference on Construction Grammar, Helsinki, Finland, 2002.

[17] DENISON, D., "The grammaticalisations of sort of, kind of and type of in English," presented at the New Reflections on Grammaticalization 3, Santiago de Compostela, 2005.

[18] DENISON, D., "The construction of SKT," presented at the Second Vigo-Newcastle-Santiago-Leuven international workshop on the structure of the noun phrase in English, Newcastle upon Tyne, 2011.

[19] GRIES, S. T. AND C. DAVID, "This is kind of/sort of interesting: Variation in hedging in English," presented at the Studies in Variation, Contacts and Change in English: Towards Multimedia in Corpus Studies, Helsinki, Finland, 2007.

[20] KEIZER, E., *The English Noun Phrase: The Nature of Linguistic Categorization*. Cambridge: Cambridge University Press, 2007.

[21] TIZÓN-COUTO, D. AND D. LORENZ, "Grammaticalization, reduction and prosodic prominence: the 'sort/kind/type of X' construction in spoken American English," presented at the LVII Congresso internazionale della Società di Linguistica Italiana, Catania, Italy, 2024.

[22] DETGES, U. AND R. WALTEREIT, "Grammaticalization vs. reanalysis: a semantic-pragmatic account of functional change in grammar," *Zeitschrift für Sprachwissenschaft,* vol. 21, no. 2, pp. 151–195, 2002.

[23] HIMMELMANN, N., "Lexicalization and grammaticization: Opposite or orthogonal?," in *What Makes Grammaticalization: A Look from its Components and its Fringes*, W. BISANG, N. HIMMELMANN, AND B. WIEMER Eds. Berlin: DeGruyter, 2004.

[24] HOPPER, P., "On some principles of grammaticalization," in *Approaches to Grammaticalization - Volume I. Theoretical and methodological issues*, E. C. TRAUGOTT AND B. HEINE Eds. Amsterdam: John Benjamins, 1991, pp. 17–36.

[25] BYBEE, J., *Language, Usage and Cognition*. Cambridge: Cambridge University Press., 2010.

[26] BOERSMA, P. AND D. WEENINK, *Praat - Doing Phonetics by Computer (Version 6.4.13)*. 2024.

[27] ANTHONY, L., *AntConc (Version 4.2.4)*. Tokyo: Waseda University, 2023.

# DO YOU HEAR [S]? THE ROLE OF PHONETIC EDUCATION AND F0 CONTEXT ON THE PERCEPTION OF GERMAN SIBILANTS

*Stephanie Berger*[1], *Marlene Böttcher*[1]

[1]*Linguistics and Phonetics, ISFAS, Kiel University*
*{sberger,mboettcher}@isfas.uni-kiel.de*

**Abstract:** In this replication study of [1] we set out to investigate the influence of the intonatory context on the categorical perception of sibilants in two groups of students enrolled in a linguistics and phonetics study program. Prior research has shown that the intonatory context influences both the production and the perception of sibilants. Sibilants produced in the context of a high boundary tone are produced with higher CoG compared to productions in the context of a low boundary tone [2, 3]. In the context of a high boundary tone, ambiguous sibilants are perceived as candidates with lower CoG while in the context of a low boundary tone, ambiguous sibilants are perceived as candidates with higher CoG [1]. We used a subset of the stimuli of the [1] study including syntactically unambiguous interrogative sentences produced with different boundary tones and sibilants of a naturally produced continuum form /s/ to /ʃ/. Sibilants were presented within target words in sentence-final position. Participants were two groups of students of the same study program in different semesters and with different amounts of exposure to phonetic training. Our results partly replicate the results of the original study and an effect of boundary tone on the perception of sibilants while highlighting important differences between groups. Students in the first semester show a different category boundary compared to students in the sixth semester, while the intonatory context was only relevant in the group of sixth semester students.

## 1 Introduction

Different speech sounds tend to be perceived in terms of categories despite having highly variable productions. This categorical perception is based on linguistic knowledge. In particular, it has been suggested that the ability to discriminate similar sounds and place them in different categories is knowledge-driven and can be acquired and developed over time [4].

The voiceless alveolar fricative /s/ and the voiceless post-alveolar fricative /ʃ/ in German are produced in fairly close proximity to each other in the vocal tract and can be produced on a continuum. Since they are voiceless, they have no pitch in the traditional sense, but voiceless fricative can have an intrinsic pitch or a so-called "sibilant pitch" ([5], see also [6]). One acoustic feature that correlates with the intrinsic pitch of sibilant is the spectral center of gravity (CoG), the mean frequency of the spectrum of a signal [7].

Previous research on the production of different fricatives in German found that the center of gravity of fricatives was influenced by the height of boundary tones. The center of gravity of fricatives was naturally lower in the context of falling final contours or low boundary tones, and it was higher with rising final contours or high boundary tones [2, 3].

Additionally, Niebuhr [1] found that the categorical perception of sibilants was also influenced by the boundary tones of an utterance. He investigated the influence on rising and falling final pitch contours on utterance final sibilants. These sibilants were produced naturally

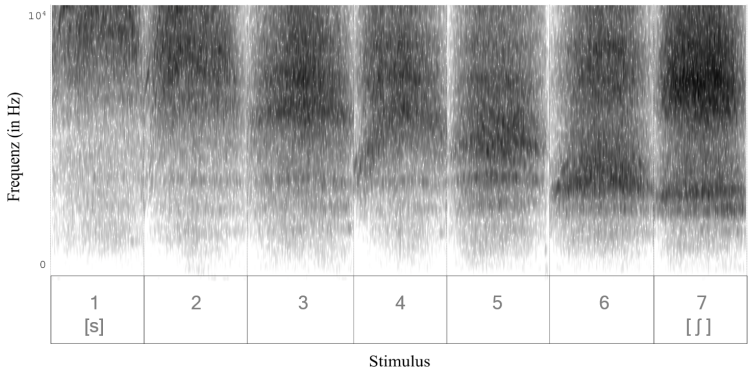



**Figure 1** – Spectrograms of the sibilants in the naturally-produced fricative continuum, from /s/-like (stimulus 1, left) to /ʃ/-like (stimulus 7, right).

on a continuum from [s] to [ʃ] and spliced into the utterances after the offset of voicing (see [1], p. 7ff. for further details). The stimuli were presented to students of linguistics and phonetics who were in the early stages of their studies and had not yet taken courses on prosody and speech acoustics ([1], p. 11). The results of this study showed that /ʃ/ was more easily recognized as in the context of a rising boundary tone, while /s/ was more easily recognized in the context of a falling boundary tone. That means that ambiguous voiceless sibilants were perceived as having lower center of gravity in the context of a higher fundamental frequency, which resulted in more /ʃ-responses, and vice versa.

Considering all these findings from previous research, especially Niebuhr [1] together with the suggestion from Liberman et al. [4] that discrimination between segments is a skill that is acquired, we are investigating the following research question: What influence do the phonetic education and the F0 context have on the categorical perception of sibilants, in particular /s/ and /ʃ/? We predict that participants with more extensive phonetic education (i.e., enrollment in more phonetics-based courses over a longer period of time) will differentiate more strongly between /s/ and /ʃ/ depending on the F0 context of the utterance.

## 2 Methods

### 2.1 Speech material

This study uses a subset of the speech materials of [1] and is therefore partly replicating that investigation. Niebuhr [1] used a sibilant continuum that originally consisted of 14 sibilants and was then reduced to 10 (see p. 7). We reduced this amount further to seven sibilants.

The recordings are exactly the same. Speaker SB (female, German L1, trained phonetician, age: 22 at the time of recording in 2014) recorded naturally-produced sibilant continua ranging from clear [s] to clear [ʃ]. The individual sibilants within the continua were produced in isolation and repeatedly, varying in tongue position and shape, as well as degree of lip rounding ([1]; see p. 7 for more detailed information). The sibilants as part of the stimulus continuum were chosen on an auditory basis to provide transitions that were as smooth as possible between the different fricatives, while still being auditorily different. This was done both for the original investigation [1], as well as the reduction in sibilants for the present study. Figure 1 provides a spectrogram representation of the fricatives in the stimulus continuum in this study, though the stimulus numbers do not coincide with those in [1]. Table 1 provides mean center-of-gravity measurements across each sibilant segment using ProsodyPro ([8], version 5.7.8.7; note that the measurement of stimulus 7 appears to be an outlier caused by two regions of high energy, see Figure 1).

**Table 1** – Overview of the mean center of gravity (CoG) of the sibilants in the fricative continuum.

| Stim. | CoG (Hz) | Stim. | CoG (Hz) | Stim. | CoG (Hz) | Stim. | CoG (Hz) |
|---|---|---|---|---|---|---|---|
| 1 | 12037.90 | 3 | 7623.40 | 5 | 5820.10 | 7 | 6986.21 |
| 2 | 9201.77 | 4 | 7642.97 | 6 | 5108.97 | | |

Speaker SB also recorded utterances that the sibilants were then embedded in. Three pairs of questions (syntactically marked) were recorded ending in monosyllabic nouns that are minimal pairs with /s/ and /ʃ/, and differ between the pairs in their vowel (/ɪ/, /a/, and /ʊ/). Examples 1 through 3 provide the different utterances.

1. Siehst du den Bus/Busch? (*Do you see the bus/bush?*)
2. Hast du einen Pass/Pasch? (*Do you have a passport/double (while rolling dice)?*)
3. Kaufst du noch Viss/Fisch? (*Do you still buy Viss (cleaning supply)/fish?*)

The intonation contours of the utterances were kept as constant as possible during production. For that, the verb in initial position of the utterance received a prenuclear pitch accent with a high tone, and then a fall to a lower F0 point. This was a transitional dip, followed by the nuclear pitch accent, which was again realized as a high tone on the target word in utterance-final position. The nuclear pitch accent also included a fall to a mid-low level at the end of the utterance. The contours were then manipulated in Praat ([9]; then version 5.4) after the pre-nuclear pitch accent into a) a deep dip followed by a steep rise for the rising contour, and b) a rise until before the target word and then a steep fall within for the falling contour (see [1], p. 10f. for further information).

The final sibilants (intended /s/ and /ʃ/) were cut from the utterances at the point where frication was already present, at which point the voicing from the preceding vowel had also ended. The seven sibilants were then spliced into the utterance in the experiment software, see below.

### 2.2 Experiment design

The experiment in this study consisted of 84 stimuli (7 sibilants $\times$ 6 utterances $\times$ 2 F0 contours). The stimuli were presented once without repetition in an identification-task design to participants using ExperimentMFC [10, 11] in Praat [9]. The utterances were included without a final sibilant, and the sibilants were spliced in during the experiment by including two stimuli and setting the stimulus medial silence duration in ExperimentMFC to zero seconds for a smooth transition. The stimuli were paired with images of the sentence-final nouns (see Figure 2 for examples; see also [1], p. 12) as the response buttons, and participants were instructed to click on the image that corresponds to what they heard, but were reassured that there were no right or wrong answers. The utterance/sibilant combinations were randomized for each participant using the <PermuteBalancedNoDoublets> option in ExperimentMFC. However, the image corresponding to the /s/-target was always on the left of the screen, and the image corresponding to the /ʃ/-target was always on the right.

The experiments took place in a phonetics class with participants wearing headphones. Participating in the experiment was part of the course (see below for further information), but there was no grade, extra credit, or other positive or negative incentive offered. All participants provided informed consent for subsequent data usage.

### 2.3 Participants

Thirty students participated in the experiment. Of these students, 16 were in the first semester of their Bachelor studies “Empirical Linguistics”, which includes both linguistics and phonetics


**(a)** **(b)** **(c)**

**Figure 2** – Example of a rating screens. Possible stimuli: a) Siehst du den Bus/Busch? (engl.: *Do you see the bus/bush?*), b) Hast du einen Pass/Pasch? (*Do you have a passport/double (while rolling dice)?*), c) Kaufst du noch Viss/Fisch? (*Do you still buy Viss (cleaning supply)/fish?*). Participants were asked to click on the image corresponding to what they heard.

courses. The participation of these students was part of their "Introduction to Phonetics" seminar ("Grundlagen der Phonetik"), which runs alongside a lecture and marks their first contact with phonetics. Three of the first-semester students self-identified as male, and 13 as female. All are German-L1 speakers with two bilingual speakers in the group (Arabic and Russian), between the age of 18 and 25, and none have prior perception or production experiment experience. None reported auditive impairment at the time of the study, and all had normal or corrected to normal vision. Ten participants report playing musical instruments, and of these, four participants are also singing.

There were also 14 students in the sixth and final semester of the Bachelor program (three self-identified as male, ten as female, one as diverse). These students have undergone extensive phonetic training during their program, and participated—aside from the introductory lecture and seminar—in courses teaching a) to identify and produce the entire IPA, b) how to annotate and segment speech signals, and c) prosody, and the experiment is conducted as part of a seminar on experimental phonetics near the end of the course. Thirteen participants were German L1 speakers (three are bilinguals with Low German, Danish, and Polish), and one participant was a Russian L1 speaker. They were between the ages of 21 and 29. Again, none reported auditive impairments, and all had normal or corrected to normal vision. Only one out of the 14 participants had previously participated in another perception experiment, the others had no prior experience. Eight participants reported playing musical instruments, and three participants are singers (two of which are also included in the group playing musical instruments).

### 2.4 Statistical analyses

Statistical analyses were run in R Statistical Software (version 4.3.1; [12, 13]) using the `lmer()` function of the `lme4` package [14] and the `emmeans()` function of the `emmeans` package [15].

## 3 Results

Figure 3 shows the results from the identification experiment for both student groups. In both groups, the stimuli one and two are fair /s/-candidates (identified as /s/ in 98 to 100% of instances), while the stimuli five, six and seven are fair /ʃ/-candidates (identified as /ʃ/ in 90 to 100% of instances). Stimuli three and four are less clearly identified. Stimulus three is more frequently identified as a /s/-candidate by both students in the first (78%) and sixth semester (61%), while stimulus four is less frequently identified as a /s/-candidate by both students in the first (59%) and sixth semester (38%). These stimuli are considered ambiguous candidates.

The identification of these ambiguous candidates is affected by the intonation contour. In the group of students from the sixth semester, the third stimulus is more frequently identified as a /s/-candidate in both contexts, while a rising boundary tone leads to more [ʃ] responses (43%)

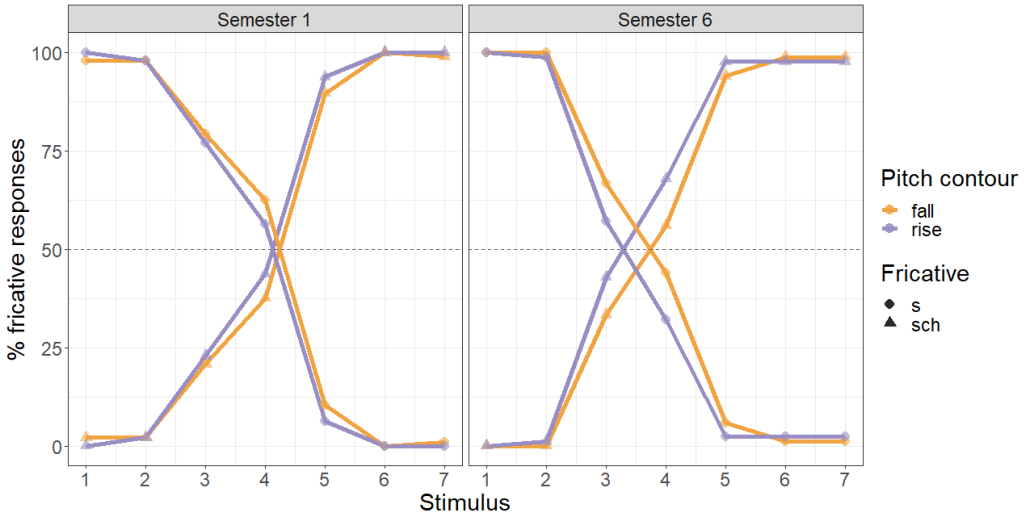



**Figure 3** – Results of the identification experiment, separated by students taking part in a course of the first or sixth semester of the linguistics program. (Stimulus 1: /s/-like, stimulus 7: /ʃ/-like)

**Table 2** – Type III Analysis of Variance Table with Satterthwaite's method for the /s/-Response Ratio.

| | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
|---|---|---|---|---|
| Semester | 787 | 787 | 4.29 | .0477* |
| Stimulus | 707354 | 117892 | 642.07 | <.0001* |
| Contour | 663 | 663 | 3.61 | .0581 |
| Semester:Stimulus | 8293 | 1382 | 7.53 | <.0001* |
| Semester:Contour | 81 | 81 | 0.44 | .5060 |
| Stimulus:Contour | 1321 | 220 | 1.20 | .3062 |
| Semester:Stimulus:Contour | 291 | 49 | 0.26 | .9531 |

compared to a falling boundary tone (33%). The fourth stimulus is more frequently identified as a /ʃ/-candidate with a similar tendency of more [ʃ] responses in the context of a rising boundary tone (68%) compared to a falling boundary tone (56%).

In the group of students from the first semester, the third stimulus is equally frequently identified as a /ʃ/-candidate in the two intonatory contexts: 79% in a falling or low f0-context and 77% in a rising or high f0-context. The fourth stimulus, however, shows a similar tendency to the third stimulus in the group of sixth semester students and is more frequently identified as a /s/-candidate preceded by a falling contour (63%) compared to a rising contour (56%).

A linear mixed regression was used to test whether the length of studies (i.e. the *semester*), the specific *stimulus* and the intonation *contour* as well as their interaction and the random intercept of the individual participant significantly predict the proportion of a [s]-response (model syntax: ([s]-response ratio $\sim$ semester * stimulus * contour + (1 | participant)). The model revealed significant main effects for semester and stimulus as well as a significant interaction of semester and stimulus ($p < 0.001$, conditional $R^2 = 0.91$, marginal $R^2 = 0.89$). Table 2 presents the analysis of variance for the dependent variable.

Post-hoc Tukey pairwise comparisons revealed significantly more [s]-responses in the first semester compared to the sixth semester for the third and fourth stimulus, as well as for the third compared to the fourth and the fourth compared to the fifth stimulus in the first and the sixth semester. Table 3 presents the relevant pairwise comparisons.

**Table 3** – Relevant pairwise comparisons of the percentage of [s]-responses, irrespective of the F0 contour. Only comparisons of the stimuli 3 through 5 are mentioned, as this is where the category boundaries are expected to be.

| contrast | estimate | SE | t.ratio | p.value |
|---|---|---|---|---|
| Semester 1, stimulus 3 – semester 6, stimulus 3 | 16.22 | 4.09 | 3.97 | .0079* |
| Semester 1, stimulus 4 – semester 6, stimulus 4 | 21.28 | 4.09 | 5.20 | <.0001* |
| Semester 1, stimulus 5 – semester 6, stimulus 5 | 4.17 | 4.09 | 1.02 | .9991 |
| Semester 1, stimulus 3 – semester 1, stimulus 4 | 18.75 | 3.39 | 5.54 | <.0001* |
| Semester 1, stimulus 4 – semester 1, stimulus 5 | 51.04 | 3.39 | 15.07 | <.0001* |
| Semester 6, stimulus 3 – semester 6, stimulus 4 | 23.81 | 3.62 | 6.58 | <.0001* |
| Semester 6, stimulus 4 – semester 6, stimulus 5 | 33.93 | 3.62 | 9.37 | <.0001* |

# 4 Discussion

The results of this study show that there is evidence for the categorical perception of sibilants, but that the effect of the F0 context is not significant in the current sample—contrary to what would be expected from previous research such as [1].

The phonetic education, meaning the comparison between listeners in their first or last semester of the Bachelor program, had a much stronger and statistically significant effect on the stimulus ratings, as was predicted. For the students with more experience in phonetics, the responses switched to /ʃ/ earlier than for the students at the beginning of their Bachelor program within the sibilant-continuum.

The inferential statistics provide only partial evidence for the hypothesis put forward earlier. We only found a significant effect of semester, yet not of the F0 context. The descriptive analyses suggest that there are still differences in sibilant perception in the expected direction. Students at the end of their Bachelor program and after taking part in several phonetics courses show a sensitivity to the spectral CoG in different intonatory contexts with more [s]-responses in a falling F0 context, and more more [ʃ]-responses in a rising context. That suggests that the expected results from [1] do occur also in our sample, but that the phonetic education of the participants influences the results. The participants in this study in the sixth semester have more experience with prosody than those in Niebuhr's original study while the participants in the first semester have close to no phonetic training. Potentially, the group difference may turn out statistically significant with participants of even longer phonetic training, since an effect was visible.

A consequence of the significant effect of phonetic education on the results of this investigation for future phonetic studies would be to shy away from eliciting perception data from students with phonetic education, unless this is part of the research question. Most perception-based research in phonetics is generally interested in the perception of speech by either a wide range of people or a narrowly defined group of people. Introducing the risk of skewing the data away from untrained perceptions by including phonetics students with acquired knowledge of the general premise of perception experiments and trained ears that may pick up more cues than untrained listeners is something that needs to be considered by researchers before settling on a sample, and potential biases in the results should be pointed out.

In our study, the main difference between the participants was the semester of their Bachelor program and their exposure to phonetics content. Only one of the thirty students had previously participated in another perception experiment—or any phonetic experiment, for that matter. That suggests, that experiment experience was not a factor in our study.

Another factor potentially affecting the different perceptional skills in the two groups of students is musical training. Musical skills affect the abilities in different areas of speech per-

ception, e.g. speech segmentation [16], prosody [17] and lexical tone [18]. Within our groups of participants, a similar distribution of musical abilities was reported. It is therefore assumed to be not relevant for the interpretation of our results. Yet, future research could check the differences of students with and without reported musical training.

It is also possible that the results were affected by word frequency. A word like "Viss", a German cleaning supply, seems to be less known by younger students today than when the stimuli were designed in 2014. Previous research suggests more probable words are more easily identified [19]. The word "Viss" is spelled out on the image and therefore serves as a reminder of the meaning, which means it is less likely to confound the results. But it is still possible that especially the target word pair *Viss/Fisch*, but also to some degree the pair *Pass/Pasch*, may skew the responses, the former one in the direction of the more frequent and therefore probable *Fisch* and its fricative /ʃ/, the latter one towards the more likely *Pass* (shown in the image with its passport meaning, but which can have multiple meanings in German) and its fricative /s/. Potential other effects on the ratings could be the vowel before the fricative and the formant transitions into the fricatives as reported in [1]. A follow-up investigation could check the individual questions (and with it the different vowels) to see if word frequency and probability has an influence on the results as well.

## 5 Conclusion

We found that in the perception of segment categories, especially in different F0 contexts, there is a difference between speakers with more or less phonetic training, which is in line with other research in a similar domain (e.g. [20]). Our study shows that it is necessary for researchers in phonetics to choose their perception experiment participants deliberately and potentially refrain from recruiting students within the linguistics and phonetics departments when the perception of fine phonetic detail is of interest and training could skew its perception.

## Acknowledgements

Thank you to Oliver Niebuhr for providing the original recordings, the participants of the courses "Grundlagen der Phonetik" (2024) and "Experimentelle Phonetik" (2023) at Kiel University, and to the attendants of the "20. Jahrestreffen für Phonetik und Phonologie im deutschsprachigen Raum" conference for the fruitful discussions.

## References

[1] NIEBUHR, O.: *On the perception of "segmental intonation": F0 context effects on sibilant identification in German. EURASIP Journal on Audio, Speech, and Music Processing*, 19(1), pp. 1–20, 2017.

[2] NIEBUHR, O.: *At the edge of intonation: The interplay of utterance-final F0 movements and voiceless fricative sounds. Phonetica*, 69, pp. 7–27, 2012.

[3] RITTER, S. and T. B. ROETTGER: *Speakers modulate noise-induced pitch according to intonational context*. In *Proceedings of the 7th International Conference on Speech Prosody*, pp. 890–893. 2014.

[4] LIBERMAN, A. M., K. S. HARRIS, J. A. KINNEY, and H. LANE: *The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. Journal of Experimental Psychology*, 61(5), pp. 379–388, 1961.

[5] TRAUNMÜLLER, H.: *Some aspects of the sound of speech sounds*. In *The psychophysics of speech perception*, pp. 293–305. Springer, 1987.

[6] JOHNSON, K.: *Acoustic and auditory phonetics (3rd Edition)*. Malden, MA: Wiley-Blackwell, 2012.

[7] BOERSMA, P.: *Spectrum: Get centre of gravity...* 2007. URL https://www.fon.hum.uva.nl/praat/manual/Spectrum__Get_centre_of_gravity___.html.

[8] XU, Y.: *ProsodyPro – A tool for large-scale systematic prosody analysis*. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France*, pp. 7–10. 2013.

[9] BOERSMA, P. and D. WEENINK: *Praat: Doing phonetics by computer (version 6.0.37)*. 2018. URL http://www.praat.org/.

[10] BOERSMA, P.: *ExperimentMFC*. 2013. URL https://www.fon.hum.uva.nl/praat/manual/ExperimentMFC.html.

[11] BOERSMA, P.: *ExperimentMFC 2.1. The experiment file*. 2016. URL https://www.fon.hum.uva.nl/praat/manual/ExperimentMFC_2_1__The_experiment_file.html.

[12] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

[13] POSIT TEAM: *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA, 2024. URL http://www.posit.co/.

[14] BATES, D., M. MÄCHLER, B. BOLKER, and S. WALKER: *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software*, 67(1), pp. 1–48, 2015. doi:10.18637/jss.v067.i01.

[15] LENTH, R. V.: *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023. URL https://CRAN.R-project.org/package=emmeans. R package version 1.8.8.

[16] FRANÇOIS, C., J. CHOBERT, M. BESSON, and D. SCHÖN: *Music Training for the Development of Speech Segmentation*. *Cerebral Cortex*, 23(9), pp. 2038–2043, 2012. doi:10.1093/cercor/bhs180.

[17] CASON, N., M. MARMURSZTEJN, M. D'IMPERIO, and D. SCHÖN: *Rhythmic abilities correlate with L2 prosody imitation abilities in typologically different languages*. *Language and Speech*, 63(1), pp. 149–165, 2019.

[18] DELOGU, F., G. LAMPIS, and M. O. BELARDINELLI: *From melody to lexical tone: Musical ability enhances specific aspects of foreign language perception*. *European Journal of Cognitive Psychology*, 22(1), pp. 46–61, 2010. doi:10.1080/09541440802708136.

[19] ROGERS, C. S. and A. WINGFIELD: *Stimulus-independent semantic bias misdirects word recognition in older adults*. *The Journal of the Acoustical Society of America*, 138(1), pp. EL26–EL30, 2015.

[20] WONG, J. W. S.: *The effects of high and low variability phonetic training on the perception and production of english vowels/e/-/æ/by cantonese esl learners with high and low l2 proficiency levels.* In *Interspeech*, pp. 524–528. 2014.

# WHAT MAKES A CONSONANT A GEMINATE?

*Lia Saki Bučar Shigemori, Felicitas Kleber*

*Institute for Phonetics and Speech Processing, LMU München*
*{lia | kleber}@phonetik.uni-muenchen.de*

**Abstract:** The aim of this cross-linguistic study is to assess the phonetic implementation of duration based phonemic oppositions in consonants of prototypical geminating languages Finnish and Italian, in comparison to languages where the oppositions in consonants have traditionally been described as only secondarily being cued by duration (German varieties). Therefore, absolute closure duration and consonant to vowel ratio were analyzed. Additionally, language specific articulation rates are reported, given their impact on quantity contrasts. Results suggest that even if robust durational differences are observed in non-geminating languages, these are less pronounced than in geminating languages. This is further confirmed in the accuracy rates of the random forest analysis. In addition, while relational measures appear to be more robust cross-linguistically and across speech rates, the stability and the impact of a cue within a language is determined by its phonotactics.

## 1 Introduction

Geminates are long consonants which exist in a phonological opposition to short consonants or singletons [1]. Despite the primary cue signalling this opposition being consonant duration, languages vary in the implementation of this durational contrast. For example, the geminates to singletons ratio in Japanese is about 3:1 [2], while in the languages studied by [3] it ranges from 1.43:1 in the Swiss German varietey spoken in Bern to 2.16:1 for Hungarian. Preceding vowel shortening as a correlate of gemination has been observed in Kannada, Tamil, Telugu, Hausa, Italian, Icelandic, Norwegian, Hungarian, Arabic, Shilha, Amharic, Galla, Dogri, Bengali, Sinhalese, and Rembarrnga [4]. Mixed results regarding this correlate have been presented for Japanese and Finnish with [4] reporting vowel shortening before geminates in Finnish but not in Japanese and [5] presenting evidence for vowel shortening before geminates in Japanese (analogously to the Tashlhiyt and Italian speakers) and vowel lengthening before geminates in Finnish. [6, 7] report further correlates of gemination which can serve as secondary cues, such as f0 or intensity.

In non-geminating languages, too, durational variation is utilized for other phonological contrasts. In Standard German for example, proportional closure durations are relevant cues in addition to voice onset time (VOT) for the voicing contrast, often referred to as fortis/lenis contrast. Vowels preceding lenis stops (/b, d, g/) are produced with a longer duration than before fortis stops (/p, t, k/), i.e. a cue in proportional length that listeners heavily rely on in perception [8].

The aim of this study was to investigate the role of acoustic duration in consonant contrasts by means of a comparison between typologically different geminating and non-geminating languages and language varieties. We specifically compare the phonetic implementation of stop consonant duration in Finnish (FIN), Italian (ITA), German German (D-GER) and Swiss German (CH-GER). FIN and ITA are prototypical geminate languages, but differ in that FIN has

a phonological length contrast for vowels, while in ITA vowel length is allophonic. More precisely, in ITA vowels are short before geminates and long before singletons and the consonant to vowel ratio has been found to be a stable correlate of this complementary length contrast across speech rates [9]. In FIN, on the other hand, both phonemic quantity contrasts can be freely combined, i.e. the vowel length contrast is independent of the following consonant length contrast, which may explain some of the above mentioned opposite implementation patterns as e.g. phonetic vowel lengthening before geminates in FIN [5] and allophonic vowel shortening in ITA. As already mentioned above, D-GER is traditionally described as having a fortis/lenis consonant contrast. It also has a vowel length contrast, which for most vowels is also cued by clear auditory differences in vowel quality. In CH-GER again closure duration plays a far greater role in distinguishing lenis and fortis stops, so that some researchers have suggested to refer to it as a singleton/geminate contrast [10]. The comparison between FIN and ITA enables us to assess whether the relevance of absolute or relative measures depends on the phonotactics of the language and how the utilization of duration differs between prototypical geminate languages. By comparing CH-GER with two phonotactically diverse geminate languages and D-GER, in which consonant duration has been reported to play only a secondary role, we evaluate the phonetic justification of the singleton/geminate terminology for CH-GER. To this end we compared various measures accommodating for duration in true and false geminating languages. The specific hypotheses are given in Section 2.2, alongside with the descriptions of the measures. We will refer to the consonant contrast in D- and CH-GER in our data as singleton/geminate for simplicity reasons.

One of these measures is not a correlate of gemination itself but impacts all quantity contrasts by its very nature. Articulation rate is defined as the number of segments realized per second in an uninterrupted stretch of speech [11]. A higher number is not only an indicator of faster speech but also of shorter segments, regardless their linguistic implication. Underlying long phonemes are more affected than their short counterparts, an observations that holds true for both geminates as opposed to singletons [5] and long compared to short vowels [12]. While speech rate has been recurrently manipulated actively in the study of quantity contrasts (e.g. [9]) in terms of different speech rate conditions (i.e. e.g. fast vs. normal rate), the present study determines the effect of speech rate on quantity contrasts given that languages differ in articulation rate [13].

# 2 Methods

## 2.1 Data

The present analysis was based on data that has been extracted from recordings of previous studies (CH-German: Zurich speakers from [14]; D-German: Standard German speakers from [15], ITA: [16], FIN: [17]). The target words were disyllabic words with lexical stress on the initial syllable. They were produced in language-specific carrier phrases. While for FIN as well as D- and CH-GER the carrier phrase was always the same and the target word carried the nuclear accent, for ITA the carrier phrase differed for each target word. In addition, each target word was presented in two different prosodic contexts; in a left dislocated position and in utterance final position. The target words for each of the languages are shown in Tables 1, 2 and 3. Data from 13 FIN, 10 ITA, 20 CH-GER and 19 D-GER speakers were analyzed. While the data sets for FIN, CH-GER and D-GER included recordings with different speech rates, only the normal speech rate was used in this study. Regarding the ITA data, only recordings in warm ambient temperature was included for the analysis.

**Table 1** – Target words for the FIN set.

| short vowel | | long vowel | |
|---|---|---|---|
| short cons. | long cons. | short cons. | long cons. |
| k**ot**a | k**att**o | k**aat**o | t**aakk**a |
| t**ak**a | m**att**o | k**oot**a | t**uutt**i |
| t**ap**aa | t**akk**a | | |
| m**at**o | t**app**aa | | |
| r**ap**u | t**utt**i | | |
| | r**app**u | | |

**Table 2** – ITA target words.

| ambigue vowel | |
|---|---|
| short cons. | long cons. |
| P**ap**a | p**app**a |
| c**ap**i | t**app**i |
| t**at**a / f**at**a | S**att**a |
| d**at**i | r**att**i |

**Table 3** – Target words for the D-GER and CH-GER set.

| short vowel | | long vowel | |
|---|---|---|---|
| short cons. | long cons. | short cons. | long cons. |
| P**udd**ing | B**utt**er | K**ad**er | K**at**er |
| R**abb**i | C**utt**er | L**up**e | P**ut**e |
| S**upp**e | R**app**e | P**ud**er | |
| | | R**ab**e | |
| | | T**ub**e | |

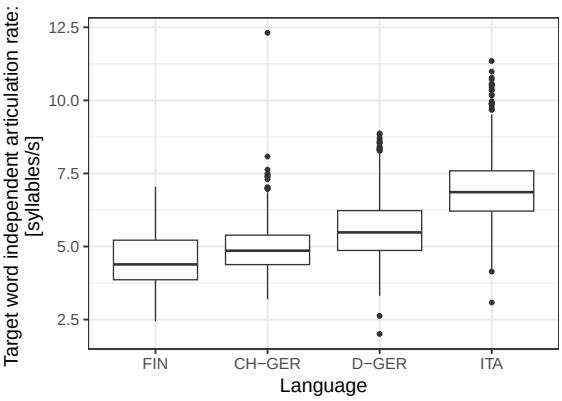
### 2.2 Acoustic measures

The following acoustic measures were determined and compared for the four languages using mixed effects models in R [18] with the package `lmerTest` [19] and `emmeans` [20]. Specific models will be presented in the Results section.

**Target word independent articulation rate:** Following the definition in Section 1, articulation rate was measured as syllables per second, separately for each sentence production. The number of syllables was calculated based on orthography, so that a segment was counted as a syllable even if the vowel was deleted in the actual production. However, if speakers produced a different word with a deviant syllable count, the number of syllables for that sentence was adjusted. Only the number of syllables of the carrier sentence without the target word was used. Given that we only have single sentences of read speech, pauses were not an issue. The duration was measured based on the acoustic segmentation, with the duration of the target word subtracted from the utterance duration.

**Absolute consonant closure durations** Closure duration was measured from the end of the preceding vocalic segment to the burst and does not include the burst and aspiration phase. Consonant closure duration should be significantly longer in geminates than in singletons, but its magnitude can differ between speakers and languages.

**Consonant to vowel ratio** In this analysis consonant to vowel ratio (C/V ratio) was tested as the proportional measure which has been often referred to in studies on geminates and which has been suggested as a rate independent acoustic correlate to the consonant length contrast (cf. Section 1). It is calculated as the closure duration of the singleton or geminate stop, respectively, divided by the preceding vowel duration.




| | estimate | SE | p-value |
|---|---|---|---|
| FIN - CH-GER | -0.35 | 0.29 | 0.64 |
| FIN - D-GER | -1.02 | 0.30 | 0.005 |
| FIN - ITA | -2.45 | 0.34 | <.0001 |
| CH-GER - D-GER | -0.67 | 0.24 | 0.035 |
| CH-GER - ITA | -2.09 | 0.32 | <.0001 |
| GER - ITA | -1.43 | 0.32 | 0.0002 |

**Figure 1** – Articulation rate measured as syllables per second based on the sentence excluding the target word. The table on the right lists the results of the post-hoc pairwise comparison.

### 2.3 Random forest

To determine whether the C/V ratio or the absolute consonant stop duration is more relevant to separate longer and shorter consonants in each of the languages, a random forest model was created separately for each language. The preceding vowel duration was added as an additional parameter. On the one hand, it is known that in general consonant and preceding vowel duration correlate and adding this additional parameter might improve the accuracy of the random forest models, on the other hand it serves as a control as vowel duration is not expected to be the most important predictor for the singleton/geminate contrast. In R, using the `randomForest`-package [21], models were trained on 80% of the available data and then tested on the remaining 20%. The models were used to categorise automatically the tokens into singletons or geminates, based on consonant closure duration, preceding vowel duration and the C/V ratio. The effects of the acoustic features to categorize stop consonants in singletons and geminates was evaluated based on the variable importance scores provided by the random forest models.

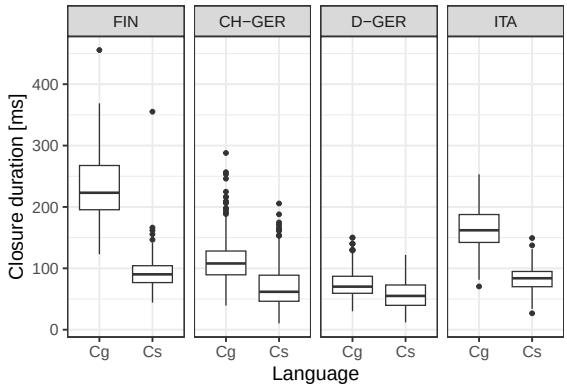
## 3 Results

### 3.1 Target word independent articulation rate

As can be seen in Figure 1, the target word independent articulation rate was fastest for ITA and slowest for FIN with CH-GER and D-GER in between. The linear mixed effects analysis with LANGUAGE as fixed effect and by-subject and by-word intercept as random effects confirmed that indeed, there was a significant relationship between language and articulation rate ($F[3, 72.2] = 20.4$, $p<0.001$). The post-hoc pairwise comparison revealed that only the difference between FIN and CH-GER was not significant. The very fast articulation rate for ITA could be partially related to the generally longer sentences in the ITA data, which consisted of 6 to 16 syllables (target word excluded) compared to 6 syllables in the FIN and 5 syllables in the data of GER varieties.

### 3.2 Stop closure duration

The closure duration for singleton and geminate consonants, separately for the four languages is shown in Figure 2. FIN, which has the slowest articulation rate also has the longest segment durations, especially for geminates. To asses whether consonant quantity had an effect on closure duration in each of the languages, we performed a linear mixed effects analysis with LANGUAGE (four levels) and CONSONANT QUANTITY (singleton/geminate), as well as their interaction as fixed factors and SPEAKER (with by-speaker slope for CONSONANT QUANTITY)




| | estimate | SE | p-value |
|---|---|---|---|
| FIN Cg - Cs | 138.0 | 10.80 | <0.0001 |
| CH-GER Cg - Cs | 42.7 | 11.50 | 0.0032 |
| D-GER Cg - Cs | 16.9 | 11.50 | 1.0000 |
| ITA Cg - Cs | 80.4 | 13.80 | <0.0001 |

**Figure 2** – Closure duration of geminates (Cs) and singletons (Cs), separately for the four languages FIN, CH-GER, D-GER and ITA. The table on the right lists the results of the post-hoc pairwise comparison between geminates and singletons within each of the languages.

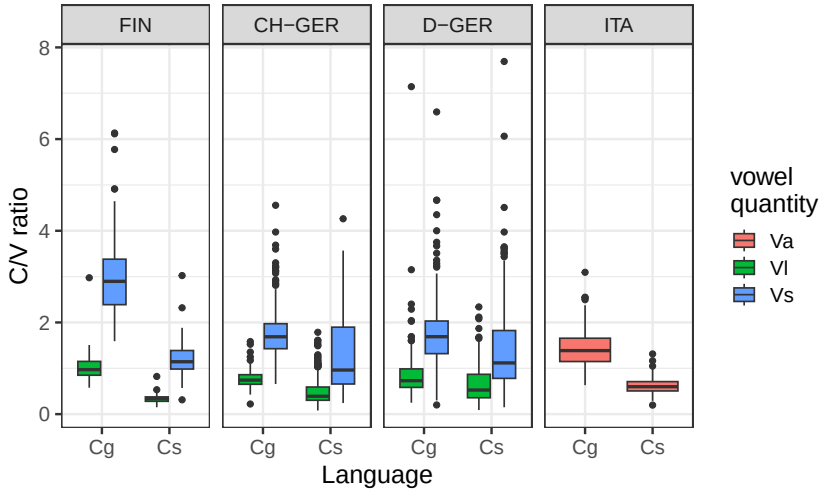



**Figure 3** – Consonant to vowel ratio of geminate (Cg) and singleton (Cs) stops. For FIN, CH-GER and D-GER separately for long (Vl) and short (Vs) vowels, while in ITA the vowel quantity is allophonic (Va).

and WORD as random factors. The main effects of LANGUAGE (F[3,60.7]=36.0, p<0.001) and CONSONANT QUANTITY (F[1,35.6]=95.2, p<0.001) as well as their interaction effect (F[3,42.4]=25.1, p<0.001) were significant. The post-hoc test revealed that the CONSONANT QUANTITY effect was significant for all languages except for D-GER.

### 3.3 C/V ratio

Figure 3 shows the C/V-ratio for singleton and geminate stops, separately for the four languages. For FIN, CH-GER and D-GER the data is additionally categorized for vowel quantity, while in ITA there is no phonological vowel quantity contrast. In general, C/V ratio is higher for geminates than singletons, as expected. However, the contrast is smaller (in terms of a greater overlap between boxes) in CH-GER and D-GER than in FIN and ITA. In FIN, CH-GER and D-GER, which have a phonological vowel quantity contrast, the C/V ratio for geminate stops with long vowels is very close to the C/V ratio of singleton stops with short vowels. If only geminate stops with long vowels and singleton stops with short vowels are compared, the C/V ratio for singletons is slightly higher than for geminates.

According to the linear mixed effects analysis with LANGUAGE, consonant CONSONANT QUANTITY and their interaction as fixed factors and SPEAKER with by-Speaker slope for CON-

SONANT QUANTITY and WORD as random factors, the main effects of CONSONANT QUANTITY (F[1, 34.2]=22.4, p<0.001) and LANGUAGE (F[3, 40.5]=5.6, p=0.003) are significant but not their interaction effect. Because Italian does not have a phonological vowel quantity contrast, VOWEL QUANTITY could not be included in the model. Therefore, we carried out separate analyses for each language.

For FIN, CH-GER and D-GER the mixed effects models included CONSONANT QUANTITY, VOWEL QUANTITY and their interaction as fixed factors with SPEAKER and WORD as random factors. In FIN, the main effects CONSONANT QUANTITY (F[1, 11]=62.6, p<0.001), VOWEL QUANTITY (F[1, 11]=80.7, p<0.001) and their interaction effect (F[1,11]=12.4, p=0.005) were significant. For CH-GER only the main effect of VOWEL QUANTITY was significant (F[1,11]=14.4, p<0.01). For D-GER, too, only the main effect of VOWEL QUANTITY was significant (F[1,11]=14.1, p<0.01).

For ITA, the mixed effects model included CONSONANT QUANTITY as fixed factor with SPEAKER and WORD as random factors. CONSONANT QUANTITY had a significant effect on the C/V ratio (F[1,7.1]=86.0, p<0.001)

### 3.4 Random forest

Random forests were trained separately for each language on a random sample of 80% of the data and tested on the remaining 20% of the data. The predictor variables were CONSONANT CLOSURE DURATION (C2closDur), VOWEL DURATION (V1Dur) and C/V RATIO (CVratio). The classification of the test data resulted in a 99% accuracy rate [95% CI : (0.9624, 0.9998)] for FIN; a 93% accuracy rate [95% CI : (0.8773, 0.9643)] for ITA, 67% accuracy rate [95% CI : (0.6132, 0.7212)] for D-GER, and 71% accuracy rate [95% CI : (0.655, 0.7566)] for D-GER. In Figure 4 the predictors are ranked according to their mean decrease accuracy for each of the languages. The mean decrease accuracy reflects how much the accuracy of the random forest drops, if the measure is excluded. For FIN, consonant closure duration was clearly most important. ITA, D-GER and CH-GER show similar patterns, with C/V ratio ranking highest, however the mean decrease accuracy value of the highest ranking measure does not stand out as much as for FIN.

## 4 Summary and Discussion

We compared stop closure duration and C/V ratio in FIN, ITA, CH-GER and D-GER consonants contrasting in length, for which duration was expected to be a cue, after having accounted for articulation rate. Language specific differences in the latter call for normalization [9, 5]. While CONSONANT QUANTITY had a significant effect on C/V ratio only for FIN and ITA, CONSONANT CLOSURE DURATION was ranked highest in the random forest classifiers for ITA, CH-GER, and D-GER, suggesting it to be a reliable measure. Raw consonant closure duration is also a reliable predictor in all languages except D-GER. Although with faster speech rate particularly long segments have been shown to compress more [5, 12], stop closure durations in ITA were longer than in CH-GER or D-GER.

For ITA and FIN, both the absolute consonant closure duration as well as C/V ratio were significantly affected by phonological consonant quantity. However, based on the random forest analysis, it can be assumed that relational measures such as C/V ratio are more important in fast than in slow geminating languages. FIN may be utilizing slower articulation rate and greater durational differences between consonant categories, because relational measures such as C/V ratio are impeded by the phonotactics including a pure phonological vowel quantity contrast.

Another important observation is that of similarities in cue ranking between ITA and the

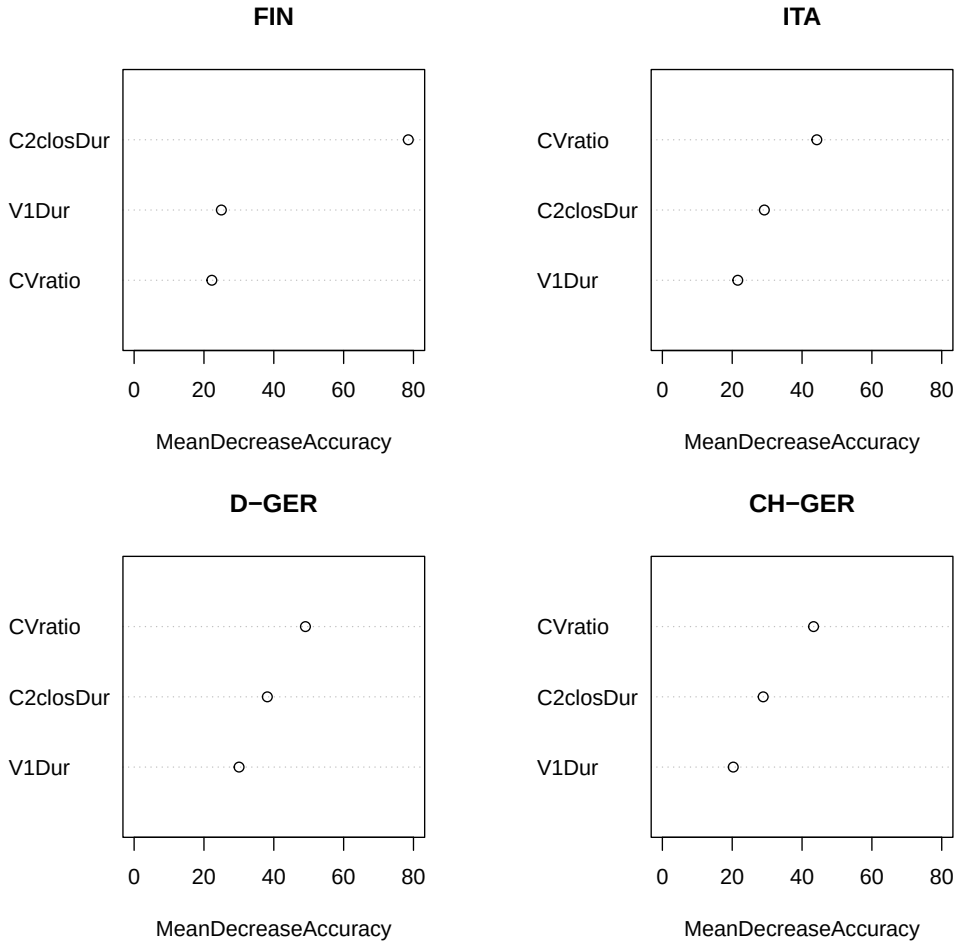



**Figure 4** – The mean decrease accuracy values in the random forest models for the languages FIN, ITA, D-GER and CH-GER.

two GER varieties. Only the greater accuracy rate for ITA suggests that duration has a greater impact in this language. Our current study confirms that duration is more relevant in CH-GER than in D-GER [14]. The absolute closure duration differed significantly for singleton and geminate consonants in CH-GER but not for DE-GER. Further more, the random forest model yielded higher accuracy rates for CH-GER than for D-GER.

## References

[1] DAVIS, S.: *Geminates*. In M. VAN OOSTENDORP, C. J. EWEN, E. HUME, and K. RICE (eds.), *The Blackwell Companion to Phonology*, vol. 2, pp. 837–859. Wiley-Blackwell, Malden, MA & Oxford, 2011.

[2] IDEMARU, K. and S. G. GUION: *Acoustic covariants of length contrast in Japanese stops*. *Journal of the International Phonetic Association*, 38(2), pp. 167–186, 2008.

[3] HAM, W. H.: *Phonetic and phonological aspects of geminate timing*. Routledge, 2001.

[4] MADDIESON, I.: *Phonetic cues to syllabification*. In V. FROMKIN (ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged*, no. 59 in Working papers in phonetics, UCLA. New York: Academic Press, 1984.

[5] HERMES, A., S. TILSEN, and R. RIDOUANE: *Cross-linguistic timing contrast in geminates: A rate-independent perspective*. In *12th International Seminar on Speech Production (ISSP2020)*, pp. 52–55. 2020.

[6] YOSHIDA, K., K. J. DE JONG, J. K. KRUSCHKE, and P.-M. PÄIVIÖ: *Cross-language similarity and difference in quantity categorization of Finnish and Japanese*. *Journal of Phonetics*, 50, pp. 81–98, 2015.

[7] BURRONI, F., R. LAU-PREECHATHAMMARACH, and S. MASPONG: *Unifying initial geminates and fortis consonants via laryngeal specification: Three case studies from Dunan, Pattani Malay, and Salentino*. In *Proceedings of the Annual Meetings on Phonology*. 2021.

[8] KOHLER, K. J.: *Dimensions in the perception of fortis and lenis plosives*. *Phonetica*, 36(4-5), pp. 332–343, 1979.

[9] PICKETT, E. R., S. E. BLUMSTEIN, and M. W. BURTON: *Effects of speaking rate on the singleton/geminate consonant contrast in Italian*. *Phonetica*, 1999.

[10] KRAEHENMANN, A.: *Swiss German stops: geminates all over the word*. *Phonology*, 18(1), pp. 109–145, 2001.

[11] JESSEN, M.: *Forensic reference data on articulation rate in german*. *Science & Justice*, 47(2), pp. 50–67, 2007.

[12] SIDDINS, J., J. HARRINGTON, U. REUBOLD, and F. KLEBER: *Investigating the relationship between accentuation, vowel tensity and compensatory shortening*. In *7th Speech Prosody Conference*. Dublin, Ireland, 2014.

[13] ROACH, P.: *Some languages are spoken more quickly than others*. In L. BAUER and P. TUDGILL (eds.), *Languages myths*, pp. 150–158. Penguin, London, 1998.

[14] ZEBE, F.: *Vowel and consonant quantity in two swiss german dialects and their corresponding varieties of standard german: effects of region, age, and tempo*. *Phonetica*, 80(3-4), pp. 185–223, 2023.

[15] JOCHIM, M. and F. KLEBER: *Fast-speech-induced hypoarticulation does not considerably affect the diachronic reversal of complementary length in Central Bavarian*. *Language and Speech*, 67(2), pp. 463–497, 2024.

[16] BUČAR SHIGEMORI, L. S. and A. VIETTI: *Acoustic analysis of Italian singleton/geminate stop production in two ambient temperature conditions*. In *Proceedings of the 19th ICPhS, Melbourne, Australia 2019*, pp. 3676–3680. 2019.

[17] JOCHIM, M. and F. KLEBER: *What do Finnish and Central Bavarian have in common? Towards an acoustically based quantity typology*. In *Proceedings of Interspeech 2017, Stockholm, Sweden*, pp. 3018–3022. Stockholm, Sweden, 2017.

[18] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

[19] KUZNETSOVA, A., P. B. BROCKHOFF, and R. H. B. CHRISTENSEN: *lmerTest package: Tests in linear mixed effects models*. *Journal of Statistical Software*, 82(13), pp. 1–26, 2017.

[20] LENTH, R. V.: *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024. URL `https://CRAN.R-project.org/package=emmeans`. R package version 1.10.5.

[21] LIAW, A. and M. WIENER: *Classification and regression by randomforest*. *R News*, 2(3), pp. 18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

# GENDER DIFFERENCES IN CHILDREN'S VOICE QUALITY AND THEIR IMPACT ON GENDER PERCEPTION – INSIGHTS FROM THE LOKIS DATABASE

*Riccarda Funk, Melanie Weirich, Adrian P. Simpson*

*Friedrich-Schiller-Universität Jena*
*riccarda.funk@uni-jena.de*

**Abstract:** The present study examines gender differences in voice quality in prepubertal children and their influence on gender perception. Recordings were made of approximately 60 German primary school children from first to third grade (LoKiS database [1] [2]). In two different listening experiments, listeners judged the gender perception on a seven-point scale and evaluated the perceived hoarseness of the five girls/boys who were rated most feminine/masculine. Voice quality (CPPS; smoothed-cepstral-peak-prominence) of girls and boys was compared and LMMs were run to investigate the influence of CPPS on perceived hoarseness and to examine the impact of voice quality on gender perception.

Gender perception ratings of girls and boys differ significantly. LMMs verify a significant influence of gender on perceived hoarseness, whereby boys sound hoarser than girls. The ratings of hoarseness are influenced by CPPS. The acoustic analysis indicates a tendency towards a more modal voice quality in boys with higher values in CPPS. However, boys and girls do not differ significantly in CPPS. Effects of CPPS on gender perception can be found, but are not systematic in all grade levels or genders. The children's voices become more modal over time, characterized by increasing CPPS.

## 1 Introduction

Studies investigating gender differences in voice quality in prepubertal children show inconsistent results. Some studies demonstrate that boys' voices are more modal than girls', exhibiting lower values of Jitter and Shimmer [3] and higher values of harmonics-to-noise-ratio (HNR) [4] and smoothed-cepstral-peak-prominence (CPPS) [5]. Other studies report contrasting results, with higher HNR found in girls [6], or can find barely any gender differences at all [7, 8]. This inconsistency may be due to large variation in children's speech and children's voice quality [4].

Significant differences have been found in the perception of hoarseness, also strong correlations between perceived hoarseness and measured HNR, indicating that voice quality could have an impact on gender perception in children [8].

The present study examines gender differences in voice quality of prepubertal children and their influence on gender perception in a longitudinal study (LoKiS database [1] [2]). In this paper, we focus on measuring CPPS for voice quality analysis, as it is less susceptible to interference from breathing noise or environmental noise, providing more reliable results in realistic speech situations [9]. Our specific research aims are as follows:

**Speech production**:

1. Examine differences in voice quality between prepubertal girls and boys and between unambiguous female- and male-sounding children.
2. Evaluate the development of voice quality over time.

**Speech perception**:

3. Analyze relationships between perceived hoarseness and measured voice quality.
4. Examine the effect of voice quality on gender perception in children's voices.

# 2 Method

## 2.1 Speakers

The ongoing LoKiS database [1] [2] contains acoustic recordings of approximately 60 German-speaking primary school children. In this paper, we analyze the data of the children in 1st, 2nd and 3rd grade (age 6 to 9 years). These longitudinal recordings took place in autumn 2020, 2021 and 2022 in two different primary schools in neighboring East German villages. 62 children (29 f, 33 m) were recorded in 1st grade, 60 children of this group (28 f, 32 m) in 2nd grade and 56 children (26 f, 30 m) in 3rd grade.

## 2.2 Audio recordings

The children were recorded individually in a quiet room at school. The procedure was consistent across all recordings from the first to third grade at both schools. Both spontaneous and read (repeated) speech was elicited. A complete list of the recorded speech material is available in [2].

All recordings were made with a Røde NT-USB microphone, featuring an integrated audio interface, placed within a Marantz Pro Sound Shield to enhance sound quality in the school environment. The microphone was positioned approximately 30 cm from the child's mouth. Each child was instructed to sit up straight in front of the microphone, with adjustments made to their posture if needed. The recordings were captured using Audacity [10], with a 32-bit amplitude resolution and a sampling rate of 44.1 kHz.

In this study, the acoustic analysis and stimuli for the listening experiments are limited to the description of a living room decorated for Christmas (see figure 1), the naming of disyllabic target nouns containing corner vowels (/aː/, /iː/, /uː/, see figure 1) and the repetition of two sentences: "Im Sommer blühen die Blumen" ("Flowers blossom in summer") and "Kannst du die Rose riechen?" ("Can you smell the rose?").

## 2.3 Acoustic analysis

For the acoustic analysis, the audio recordings were segmented and annotated using WebMAUS [11]. The resulting praat [12] text grids were manually reviewed and corrected. All acoustic analyses were conducted in praat. Voice quality was measured in the vowels /ɔ/, /yː/, /uː/ of sentence 1 ("Im Sommer blühen die Blumen"), /a/, /oː/ and /iː/ of sentence 2 ("Kannst du die Rose riechen?"), and /aː/, /iː/, /uː/ in the dysillabic target nouns (see figure 1). To calculate CPPS in these vowels, we used a voice report script in praat (see [13]).

## 2.4 Listening experiment 1 – Gender perception

The first listening experiment aimed to determine how feminine or masculine the children's voices were perceived to be, identifying which children produced robust or ambiguous phonetic correlates of gender (previously published in [1] and [2]). For this experiment, re-

cordings of sentences 1 and 2, along with approximately five to ten seconds of the Christmas picture description (see figure 1), were used.

**Figure 1** - Picture of a Christmas scene for the elicitation of spontaneous speech (right) and target nouns /haːzə/, /bluːmə/, /biːnə/, /taʃə/, /naːzə/, /kuːxən/, /iːgəl/, /tasə/, /vaːzə/, /tiːgɐ/ and /luːpə/ (left).

A pretest was conducted to ensure that the lexical content of the spontaneous speech samples in the listening experiment did not influence the perception of a child's gender. The Christmas picture descriptions were transcribed, and participants rated on a seven-point scale whether the text seemed more like a 'boy' (1) or 'girl' (7). About 70 people participated in each grade level's pretest, conducted online using SoSci Survey [14]. Recordings were excluded from the listening experiment if the lexical content received average ratings between 1-2 or 6-7. This led to the exclusion of three first-grade recordings, with no exclusions in the second and third grade.

167 listeners (119 f, 46 m, 2 d) participated in the listening experiment for the first-grade recordings, ranging in age from 18 to 71 years (M = 33.2, SD = 13.7). For the second-grade recordings, 114 listeners (94 f, 18 m, 2 d) aged 18 to 70 years (M = 33.2, SD = 13.4) participated. The third-grade recordings involved 101 listeners (77 f, 20 m, 4 d), aged 18 to 77 years (M = 33.6, SD = 14.5).

The listening experiment was conducted online using Percy software [15]. Stimuli were presented in a randomized order for each session, and each stimulus could be listened to twice. Listeners rated on a seven-point scale whether they perceived the speaker as a 'boy' (value 1) or a 'girl' (value 7). To keep the experiment duration for each listener to around 15 minutes, the stimuli were divided into three groups. The program was written to ensure that these groups were approximately evenly distributed among the participants.

### 2.5 Listening experiment 2 – Attribute rating

The aim of the second listening experiment was to examine how children with unambiguous gender ratings (see listening experiment 1) are perceived with respect to different attributes related to speech and voice, e. g. hoarse – clear (previously published in [16]). Stimuli consisted of recordings of sentence 1 and 2 from the five children with the most masculine and feminine-sounding voices identified in listening experiment 1.

In this listening experiment, 102 listeners (80 female, 21 male, 1 diverse) took part. They were 18-77 years old (M = 34.8 years, SD = 12.8 years).

The listening experiment was conducted online using SoSci Survey [14]. It included recordings of sentence 1 from first graders and sentence 2 from second graders. Listeners rated each child's voice on a seven-point scale for the attributes 'hoarse – clear.' To minimize the potential influence of gender stereotypes on perception, listeners were divided into two approximately equal priming groups like in [8]. Both groups heard the same stimuli, but the first group was informed they were listening to 'boys' (46 participants), while the second group was informed they were listening to 'girls' (56 participants).

### 2.6 Data analysis

Statistical analyses were performed in R [17], with graphical analyses using ggplot2 [18]. To identify voices perceived as distinctly feminine or masculine, we calculated a mean gender perception index (GPI) across sentences and picture descriptions from the first listening experiment. Children with GPI values below 2 or above 6 were classified as "unambiguous" feminine- or masculine-sounding.

To assess gender differences in CPPS (research aim 1), we used independent t-tests for normally distributed data and Wilcoxon rank-sum tests otherwise. Effect sizes were calculated, and comparisons were made between the five most feminine and masculine children in each grade.

Linear mixed models (LMMs) with the lme4 package [19] examined changes from first to third grade (research aim 2). We tested grade effects on CPPS, including interactions with gender. For all LMMs, Likelihood ratio tests and the anova function were used to finalize models. Post hoc Tukey tests were conducted with the emmeans package [20].

For evaluating how the most masculine and feminine children are perceived on the attributes hoarse – clear (research aim 3), we ran LMMs with attribute ratings as the dependent variable, testing effects of gender, priming, and grade level, and including speaker and listener as random intercepts.

To analyze the impact of voice quality on gender perception (research aim 4), we ran LMMs to test the effects of CPPS on gender perception, also examining interactions with gender and grade level, and including speaker, listener, and stimulus as random intercepts.

## 3 Results

### 3.1 Listening experiment 1 – Gender perception

Figure 2 illustrates the mean gender ratings for each child across listeners, stimulus type, and grade level. A higher value indicates that a child's voice is perceived as more feminine, while a lower value suggests a more masculine perception.

Among the children, 16 first-graders, 21 second-graders, and 16 third-graders exhibit phonetic patterns that result in systematic gender ratings, with ratings below 2 or above 6 (see [1]). In contrast, the remaining children produce acoustic patterns leading to more ambiguous and variable gender ratings.

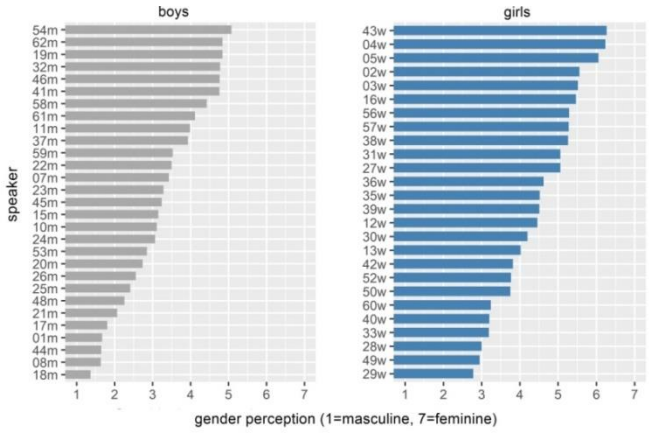



**Figure 2** - Mean values of the gender perception score in listening experiment 1.

The linear mixed model (LMM) that best explains variations in gender perception involves a three-way interaction between gender, grade level, and stimulus type ($\chi^2(10) = 402.66$, $p < .0001$). Post hoc Tukey tests reveal that the gender difference between boys and girls increases with grade level. Additionally, the difference is influenced by stimulus type: the gender distinction between boys and girls is more pronounced in the picture description task than in sentence 1 or sentence 2 (see figure 3).

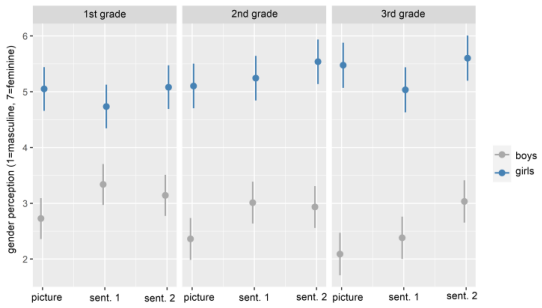



**Figure 3** - Three-way interaction between gender, grade level and stimulus type.

## 3.2 Listening experiment 2 – Attribute rating

A significant effect of gender in interaction with grade level on the perception of the attribute pair hoarse-clear can be observed ($\chi^2(1) = 23.12$, $p = .002$). The boys are perceived as significantly hoarser than the girls (see figure 4, left), with the difference being greater in second grade than in first grade. There is no effect of priming on the perception ($\chi^2$ (1) = 0.26, $p = .61$).

A significant influence of CPPS on the perception of hoarseness in interaction with gender and grade level can be found ($\chi^2(3) = 21.16$, $p = .01$, see figure 4 right). The post hoc Tukey tests show a significant positive trend in all cases: The higher the CPPS value, the clearer the children's voices sound. In first grade, the effect is stronger for boys, while in second grade, it is stronger for girls.

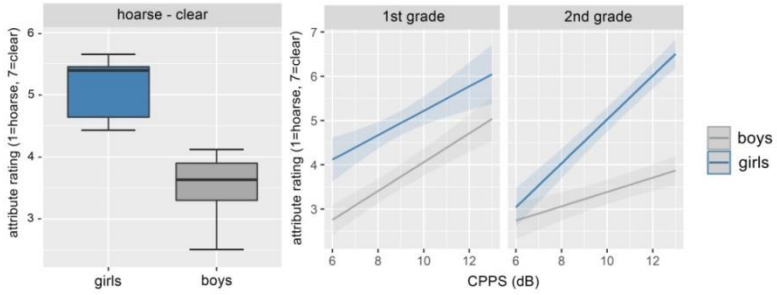



**Figure 4** - Attribute rating of listening experiment 2 (mean values of 1st and 2nd grade) (left) and relationship between perceived hoarseness and measured CPPS (right).

## 3.3 Gender differences and development in voice quality

Table 1 includes all means, standard deviations and results of the statistical tests of every grade level concerning gender differences in CPPS. Gender differences are also illustrated in figure 5. The two-sample independent t-test and Wilcoxon rank-sum tests comparing girls and boys do not reveal any significant gender differences at all. However, boys tend to show slightly higher values than girls, especially in the first and second grade (see table 1), indicating that the boys speak with a slightly more modal voice, which is in contrast to the perceived hoarse voice of the masculine-sounding boys.


**Figure 5** - Gender differences in CPPS between boys and girls and bewteen the most feminine- and masculine-sounding children of listening experiment 1.

**Table 1** - Means, standard deviations and results of the statistical tests comparing gender differences in CPPS between boys and girls (left) and unambiguous feminine- and masculine-sounding children (right) separated by grade and stimulus.

**First grade**

| | | **All children (N=62)** | | | | | | | **Unambiguous children (N=10)** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Boys** | | **Girls** | | | | | **Masculine** | | **Feminine** | | | | |
| | | M | SD | M | SD | t/W | p | d | M | SD | M | SD | t/W | p | d |
| **CPPS (dB)** | Targets | 11.6 | 1,6 | 10.9 | 1.4 | 1.7 | .09 | .43 | 10.9 | 1.6 | 11.3 | 0.8 | -0,5 | .65 | .30 |
| | Sent. 1 | 9.0 | 1.3 | 8.6 | 1.1 | 1.4 | .19 | .34 | 8.7 | 1.6 | 9.1 | 0.9 | -0.5 | .63 | .32 |
| | Sent. 2 | 8.2 | 1.5 | 7.9 | 1.3 | 0.7 | .46 | .19 | 8.4 | 1.8 | 8.5 | 0.8 | -0.1 | .91 | .08 |

**Second grade**

| | | **All children (N=60)** | | | | | | | **Unambiguous children (N=10)** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Boys** | | **Girls** | | | | | **Masculine** | | **Feminine** | | | | |
| | | M | SD | M | SD | t/W | p | d | M | SD | M | SD | t/W | p | d |
| **CPPS (dB)** | Targets | 11.6 | 1,8 | 11.3 | 1.3 | 0.8 | .42 | .21 | 11.6 | 1,0 | 11.4 | 1,0 | 0.1 | .89 | .09 |
| | Sent. 1 | 11.4 | 1.6 | 11.1 | 1.5 | 0.6 | .57 | .15 | 11.4 | 3.2 | 11.0 | 1.6 | 0.3 | .79 | .17 |
| | Sent. 2 | 11.3 | 1.6 | 11.1 | 1.9 | 0.6 | .56 | .15 | 11.5 | 2.9 | 12.1 | 1.5 | -0.4 | .68 | .27 |

**Third grade**

| | | **All children (N=56)** | | | | | | | **Unambiguous children (N=10)** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Boys** | | **Girls** | | | | | **Masculine** | | **Feminine** | | | | |
| | | M | SD | M | SD | t/W | p | d | M | SD | M | SD | t/W | p | d |
| **CPPS (dB)** | Targets | 11.6 | 1,3 | 11.5 | 1,3 | 0.2 | .85 | .05 | 11.2 | 1,8 | 11.4 | 1,0 | 0.2 | .85 | .05 |
| | Sent. 1 | 11.9 | 1.5 | 11.7 | 1.4 | 0,6 | .56 | .16 | 11.8 | 1.9 | 11.5 | 20.2 | 0.6 | .56 | .16 |
| | Sent. 2 | 11.4 | 1.7 | 11.7 | 1.6 | -0.6 | .54 | .17 | 10.8 | 2.3 | 11.9 | 2.3 | -0.6 | .54 | .17 |

If we only compare the unambiguous feminine- and masculine-sounding children, we find contrary results with higher values for CPPS in the feminine group. Nevertheless, none of these gender differences reaches statistical significance and the corresponding effect sizes remain negligible.

The LMM examining the development of voice quality shows a significant effect of grade level on CPPS ($\chi^2(2) = 244.70$, $p < .0001$). The post hoc Tukey test reveals a significant increase of CPPS over time, especially from first to second grade. In other words, boys and girls speak with a more modal voice over time.

### 3.4 Effect of voice quality on gender perception

When examining the impact of voice quality on gender perception, we observe a significant effect of CPPS on gender perception in interaction with both gender and grade level

($\chi^2(2) = 166.27$, $p < .0001$). The post hoc Tukey tests show that with an increase in CPPS, first-grade boys and third-grade girls are perceived as more feminine (see figure 6).

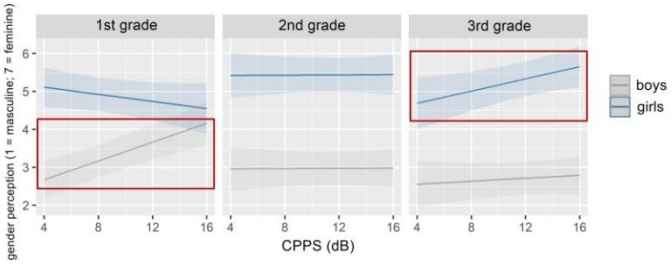



**Figure 6** - Effect of CPPS on gender perception in interaction with gender and grade level. Significant effects in red.

## 4 Discussion

The first research aim was to examine differences in voice quality between prepubertal girls and boys and between unambiguous female- and male-sounding children. The acoustic analysis of our data shows that boys and girls do not differ significantly in CPPS. However, similar to the findings of [5], there is a tendency for higher CPPS values in boys, suggesting a slightly more modal voice quality. In contrast, the feminine-sounding children tend to show higher CPPS values than the masculine-sounding children. Consequently, more modal voice qualities might appear more feminine to adult listeners. However, no gender-specific difference reaches statistical significance in this case either.

For the second research aim, it can be shown that CPPS values increase over time, similar to the findings of [21]. Therefore, children speak with a more modal voice as they grow older. This could be partly explained by the increase in vocal control with age. Alternatively, the children may have become more confident in front of the microphone as they got older.

For the third research aim, it can be clearly demonstrated that feminine voices are perceived as clearer and less hoarse compared to masculine voices, similar to the findings of [8]. Modal voices are therefore considered as an indicator of femininity in children's voices. Additionally, significant correlations between perceived hoarseness and measured CPPS can be identified. Consequently, listeners are able to correctly identify CPPS as an indicator of voice quality and may potentially use it for gender perception, even though boys and girls in general were not found to differ in CPPS.

This relationship between perceived femininity and voice quality can also be demonstrated for the fourth research aim. The higher the CPPS values of the children and the more modal the voice quality, the more feminine the voices are perceived to be. However, this is significant only for first-grade boys and third-grade girls. Other acoustic parameters (such as f0 or formant frequencies) may be overlaying the effect at least in some of the children.

In the further course of the research project, recordings of the children are to be continued throughout puberty and published in the LoKiS database (see [1] [2]). Additional acoustic analyses regarding the influence of various acoustic parameters on gender perception are planned. Furthermore, additional listening experiments are to be conducted, which will, for example, investigate the relationship between gender perception and age or the gender perception of children's voices by other children.

## Acknowledgments

The project is supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG grant SI 743/9-1, http://www.dfg.de/). We would like to thank our student

research assistant Johanna Katharina Baumbach for labeling the recordings, Tu Anh Nguyen Thi for drawing the beautiful pictures and Christoph Draxler (MLU Munich) for his help in constructing and creating the gender perception listening experiment. Most of all we are very grateful to the children and their parents agreeing to take part in this project.

## References

[1] FUNK, R., M. WEIRICH and A. P. SIMPSON: *The Effect of Fundamental Frequency on Gender Perception in Prepubertal Children: Insights from the LoKiS Database.* In *Journal of Voice*, in press, 2024.

[2] FUNK, R. and A. P. SIMPSON: *The acoustic and perceptual correlates of gender in children's voices.* In: *Speech, Language, and Hearing Research* 66(9), S. 3346 – 3363, 2023.

[3] GLAZE, L. E., D. M. BLESS, P. MILENKOVIC and R. D. SAUER: *Acoustic characteristics of children's voice. Journal of Voice.* In *Journal of Voice*, 2(4), S. 312 – 319, 1988.

[4] SHEENA, M., B. B. ASWIN and S. AMBETHKAR: *Variation of harmonics to noise ratio from the age range of 9–18 years old in both the genders.* In: *Journal of Otolaryngol Head Neck Surgery* 74(3), S. 5518 – 5523, 2022.

[5] MANJUNATHA, U.: *Cepstral analysis of voice in young children and adults.* In: *Global Journal of Otolaryngol* 11(1), S. 10 – 14, 2017.

[6] FERRAND, C. T.: *Harmonics-to-noise ratios in normally speaking prepubescent girls and boys.* In: *Journal of Voice* 14(1), S. 17 – 21, 2000.

[7] WONG, E. and B. MUNSON: *How do children mark gender phonetically? An analysis of voice quality.* In: *Proceedings of the 20th International Congress of Phonetic Sciences.* 14(1), S. 17 – 21, Guarant International Prague, 2023.

[8] SIMPSON, A. P., R. FUNK and F. PALMER: *How do children mark gender phonetically? An analysis of voice quality.* In: *Proceedings of the 20th International Congress of Phonetic Sciences.* 14(1), S. 17 – 21, Guarant International Prague, 2023.

[9] MARYN, Y., P. CORTHALS, P. VAN CAUWENBERGE, N. ROY and M. DE BODT: *Toward Improved Ecological Validity in the Acoustic Measurement of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels.* In: *Journal of Voice* 24(5), S. 540 – 555, 2010.

[10] AUDACITY: *Audacity(r): Free Audio Editor and Recorder. [Computer software]. URL https://audacityteam.org/,* 2021.

[11] KISLER, T., U. D. REICHEL and F. SCHIEL: *Multilingual processing of speech via web services.* 2021. In: *Computer Speech & Language* 45, S. 326 – 347, 2017.

[12] BOERSMA, P and D. WEENINK: *Praat, Doing phonetics by computer. [Computer software] URL http://www.praat.org, Version 6.0.40,* 2022.

[13] MAYER, J.: *Phonetische Analysen mit Praat. Ein Handbuch für Ein-und Umsteiger.* 2017.

[14] LEINER, D. J.: *Sosci survey (version 3.1.06) [computer software]. URL https://www.soscisurvey.de,* 2019.

[15] DRAXLER, C.: *Online experiments with the Percy software framework – experiences and some early results.* In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* S. 235 – 240, 2014.

[16] FUNK, R., M. WEIRICH, & A. P. SIMPSON.: *Phonetic correlates of gender in prepubertal voices.* In *R. Skarnitzl & J. Volìn (eds.), Proceedings of the 20th International Congress of Phonetic Sciences. Prague: Guarant International,* S. 22 – 26, 2023.

[17] RCORE: *R: A language and environment for statistical computing. [Computer software] URL https://www.R-project.org,* 2022.

[18] WICKHAM, H.: *ggplot2: Elegant Graphics for Data Analysis. [Computer software] URL https://ggplot2.tidyverse.org,* 2016.

[19] BATES, D., M. MÄCHELER, B. BOLKER and S. WALKER: *Fitting Linear Mixed-Effects Models Using lme4.* In: *Journal of Statistical Software* 67(1), S. 1 – 48, 2015.

[20] LENTH, R. V..: *emmeans: Estimated Marginal Means, aka Least-Squares Means. [Computer software] URL https://CRAN.R-project.org/package=emmeans,* 2016.

[21] INFUSINO, S. A., G. R. DIERCKS, J. ROGERS, D. J.AND GARCIA, S. OJHA, R. MAURER, G. BUNTING and C. J. HARTNICK: *Establishment of a Normative Cepstral Pediatric Acoustic Database.* In: *JAMA Otolaryngology–Head and Neck Surgery* 141(4), S. 358 – 363, 2015.

# THE INFLUENCE OF GRAMMATICAL CATEGORY ON SOUND CHANGE: AN EXPLORATORY STUDY OF METAPHONY IN VERBS IN THE LAUSBERG AREA (SOUTHERN ITALY)

*Pia Greca*[1]

[1]*Institute for Phonetics and Speech Processing, LMU Munich*
*greca@phonetik.uni-muenchen.de*

**Abstract:** It is still debated whether morpho-lexical components of a language can influence sound change in a top-down manner, or whether the path to phonologisation/morphologisation can only be uni-directional and bottom-up. In this study, we analyse the case of metaphony (i.e. raising or opening diphthongisation of stem vowels triggered by high suffix vowels) in verbs vs other grammatical categories in the Lausberg area (southern Italy). Previous descriptions suggest that metaphony systematically affects nouns and adjectives carrying a high suffix vowel, but it may be absent in verbs. To verify this, more than 17,000 /e, o/ stem vowels in metaphonic /i, u/ vs non-metaphonic /a, e/ suffix vowel contexts were analysed from nouns, adjectives, and verbs produced by speakers from three regions of the Lausberg area. In particular, the first frequencies of these vowels were parameterised by using Discrete Cosine Transformation (DCT), from which the first two scores $k_0$ (signal's mean) and $k_1$ (signal's slope) were analysed. The results confirmed, on the one hand, the presence of metaphony at different degrees in nouns and adjectives in the three regions. However, in all regions, differences between metaphonic and non-metaphonic contexts were virtually absent in verbs. These data therefore suggest that grammatical category can influence sound change. A possible functional explanation is that, if a trade-off of cues exists between stems and suffixes, then the higher semantic load of verb suffixes compared to the ones in nouns and adjectives may inhibit coarticulation in the stem.

## 1 Introduction

It has been recently debated whether sound changes progress from the lower phonetic level up to the morpho-lexical layer, with no possibilities of top-down influences [1, 2], or whether they may be sensitive to "higher" morpho-lexical components [3, 4]. Following lexical phonology [5, 1, 2], phonological change forms part of a modular, feed-forward phonological system, in which phonetic variation, which is initially not under cognitive control, stabilises and then becomes phonological as it enters the lexicon. In this model, there can be no direct influence of lexical or morphological levels on the phonetic output. On the other hand, however, other studies showed that morpheme boundaries [4] or grammatical or morpho-lexical category [6, 7, 8] can influence sound change in various languages. Some evidence that verbs can be more or less sensitive to sound change than other grammatical categories supports this view [7, 6, 8, 9].

In this study, metaphony in verbs vs nouns and adjectives is analysed in three regions of the Lausberg area [10]. Metaphony has its phonetic origins in trans-consonantal vowel-to-vowel coarticulation [11] and affects primarily stems containing phonetically mid vowels /e, o/, which are expected to rise or diphthongise in high suffix vowel contexts, i.e. word-final inflectional suffixes /i, u/. For instance, /b[ɛ]lla/, 'beautiful', fem. sg., becomes in the masculine /b[e]llu/

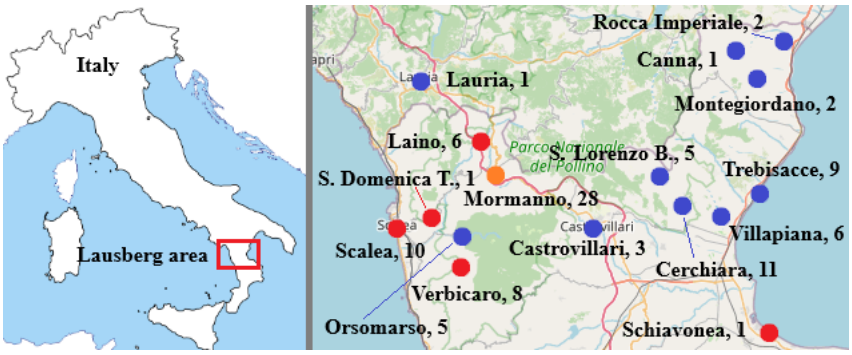



**Figure 1** – The Lausberg area and the villages in this study (map data ©OpenStreetMap 2024), including numbers of speakers for each village. Regions are colour-coded: orange for Mormanno, red for the Zwischenzone, blue for the Mittelzone.

in the mountain village of Mormanno (henceforth MM), /b[i]llu/ in the region that Lausberg [10] named *Mittelzone* (henceforth MZ), mainly extending from the Eastern coast of the area to the border between Calabria and Basilicata, and /b[jɛ]llu/ in the region that Lausberg called *Zwischenzone* (henceforth ZZ), concentrated in the western part of the area. However, this is generally not expected as systematically in verbs [10, 12]. In verbs of most Italo-Romance varieties, metaphony mainly occurs in second person singular forms of the indicative (ending in /i/) and only exceptionally in the first person singular (ending in /u/) [13]. Verbs in the Lausberg area, however, are reported not to generally show metaphonic alternations in the stem. In particular, in the dialects along the Calabrian northern east coast, only verbs derived from the Latin fourth conjugation (infinitive in -ĪRE) are expected to show metaphony, historically triggered by a /i/-suffix in the second and third person singular [10, 14].

Given this peculiar absence of metaphony in verbs, together with a lack of documentation of the phenomenon for these dialects, the main aim of the present study is to establish empirically whether metaphony in the Lausberg area affects verbs differently than other lexical categories, and, more generally, whether metaphony and sound change can be affected by morpho-syntax and grammatical category differences.

## 2 Method

### 2.1 Speakers and regions

99 participants took part to the study (53 females and 46 males, age range 13 to 92 years, mean age 48.2, *SD* = 19.2). Local varieties were considered from three clusters of villages in the Lausberg area that, according to previous impressionistic studies and acoustic analyses [15, 10, 12], show metaphony in the stem vowel to different degrees: least in MM (showing a raising from mid-low to mid-high vowels; 28 speakers), intermediate in ZZ (opening diphthongisation of mid vowels; 26 speakers), and most advanced in MZ (raising of mid vowels to high ones; 45 speakers). Figure 1 shows the different villages ($n$ = 16) included in this study and the number of participants for each village.

### 2.2 Materials and procedure

The lexical items containing the analysed stem vowels were elicited through a picture-naming task, which was carried out using the software SpeechRecorder, version 3.28.0 [16], installed on a laptop, and a headset with integrated microphone. The order of appearance of the pictures was randomised differently for each speaker. Each picture corresponded to one inflected form of a lexical item. The nouns and adjectives were first produced in isolation, then within the

carrier sentence “I say ... two times” ([(jɛ) ˈdiku ... dui ˈvotə]) in the dialect. Verbs could not be elicited in isolation but within a sentence described by a picture. Similarly, inflected adjectives had to be elicited in combination with a noun (some examples of eliciting stimuli are available in Appendix B in [15]).

Each speaker produced 123 different words (out of a total of 57 different lexical stem types, e.g. the words /bella, bellu/ share the lexical stem /bell/) phonologically containing primary stressed mid vowels /e, o/ in the stem and suffix vowels /a, e, i, u/, from which /i, u/ are those expected to trigger metaphony. 58 words (27 lexical stem types) included a stem-/e/ vowel and 65 words (30 lexical stem types) a stem-/o/ vowel. These words included 104 different nouns and adjectives (out of 50 different lexical stem types) and 19 verbs (out of 7 different lexical stem types). Most verbs were organised in triplets (first, second and third person singular of the indicative, e.g. /ˈdormu, ˈdormisi, ˈdorme/, ‘to sleep’); while nouns and adjectives were mainly organised into pairs (mainly singular vs plural, or feminine vs masculine). Most word types ($n$ = 92) were disyllabic, while there were some words ($n$ = 31) that were trisyllabic, in which the stem and suffix vowels were either contiguous (e.g. /niˈpote/, ‘grandchild’) or separated by one syllable (e.g. /ˈtenisi/, ‘(you) have’).

The total number of potentially available stem vowels for analysis was: 123 words × 2 repetitions × 99 speakers = 24,354 tokens. However, some productions had to be removed since misarticulated or produced in Standard Italian: this left 17,259 vowels for the analysis (5161 tokens were analysed produced by speakers from MM, 3983 tokens by speakers from ZZ, and 8115 tokens by speakers from MZ). Most tokens ($n$ = 15,195) were nouns and adjectives, while 2064 tokens were verbs. In addition to the words carrying mid stem vowels, other words were elicited with high and low /i, a, u/ stem vowels: these were used to extract the F1 and F2 values of their stem vowels that were necessary for the speaker-normalisation of vowel formants.

### 2.3 Data pre-processing

The speech signals were segmented and labelled semi-automatically by using MAUS (*Munich Automatic Segmentation System*, [17]), a forced alignment system integrated in the `emuR` package (version 1.1.2) [18] available in the R programming environment (version 4.4.0, [19]). The first two formant frequencies (F1, F2) of stem vowels were extracted using the `PraatR` package (version 2.4) [20] in R with a 25 ms window and a 5 ms frame shift. Around 40% of the data had to be manually corrected because of misplaced segment boundaries or mistracked formants. Formant trajectories were time-normalised to 11 equidistant time points between the acoustic onset and offset of the stem vowel. Formants were then speaker-normalised following the method described by [21].

The time-varying shapes of speaker-normalised F1 of /e/ and /o/ vowel stems ($n$ = 17,259) were parameterised using the discrete cosine transformation (DCT). The DCT decomposes any time-varying signal into a set of ½ cycle cosine waves such that the amplitudes or coefficients of the first two, $k_0$, $k_1$, are proportional respectively to the signal’s mean and linear slope [22, 23]. It follows that $k_0$ varies with changes in F1/vowel height, while $k_1$ is expected to deviate from zero if the vowel is diphthongised, since diphthongisation causes a formant trajectory’s slope to be rising or falling over the vowel interval.

In order to test for significance of differences in metaphonic influence between regions and grammatical categories, linear mixed model statistical tests were run by using the `lmerTest` package [24] in R, while post-hoc tests were computed whenever the fixed factors interacted using the `emmeans` package (version 1.10.2, [25]). The model was of the form (R notation):

$$k \sim \text{Metaphony} * \text{Grammatical category} * \text{Region} + (1|\text{Stem}) + (1|\text{Speaker}) \qquad (1)$$

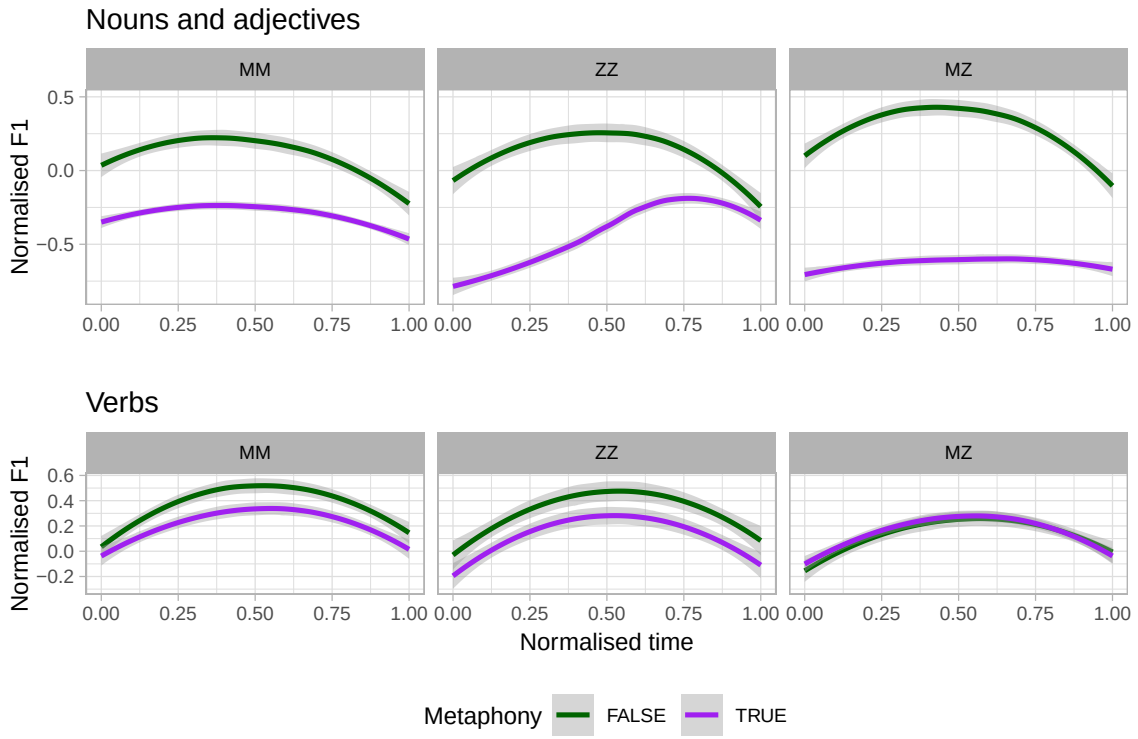



**Figure 2** – Mean F1 trajectories of /e, o/ stem vowels aggregated by region, grammatical category ("NAdj" = nouns and adjectives; "V" = verbs), and metaphonic context (TRUE = suffix vowels /i, u/; FALSE = suffix vowels /a, e/).

in which the response $k$ was either $k_0$ or $k_1$, while the fixed factors were *Metaphony*, having two levels of high /i, u/ vs non-high /e, a/ suffix vowels; *Grammatical category* (also two levels: nouns and adjectives vs verbs), and *Region* (three levels: MM, ZZ, MZ), and also included their interaction. The intercept *Stem* included lexical stems of words independently of their suffix (e.g. the stem representation for the different inflected forms of the adjective 'good' /bona, bone, boni, bonu/ was /bon/).

## 3 Results

For nouns and adjectives, the F1 trajectories of Figure 2 (upper plots) clearly show lowering in all three regions (purple curves), thus describing stem vowel raising. However, this lowering is mostly marked in MZ, least in MM, while in ZZ it is mostly marked at the onset and less at the offset, thus suggesting a diphthongal quality of the metaphonic stem. Vice-versa, for verbs (Figure 2, lower plots) these differences are either absent (MZ), or far reduced (MM, ZZ).

The resulting variation of coefficients from the DCT analysis is consistent with these observed differences. Thus, in Figure 3 showing $k_0$ (which is proportional to the F1 mean), there are very clear differences between nouns/adjectives and verbs in the degree of metaphonic influence (dark green vs purple plots). By contrast, these differences are almost absent in verbs.

Consistently with these data, the statistical analysis with $k_0$ as the dependent variable showed a significant interaction between region, grammatical category, and metaphony ($F_{2,\ 17167.0} = 65.1$, $p < 0.001$). Post-hoc tests showed a significant influence of metaphony on the stem vowel only for nouns and adjectives (MM: $z = 18.7$, $p < 0.001$; ZZ: $z = 25.8$, $p < 0.001$; MZ: $z = 57.9$, $p < 0.001$) but not for verbs. The results also showed a greater metaphonic effect in nouns and adjectives compared with verbs in all three regions (MM: $z = –3.4$, $p = 0.01$; ZZ: $z = –3.5$, $p = 0.01$; MZ: $z = –5.3$, $p < 0.001$).

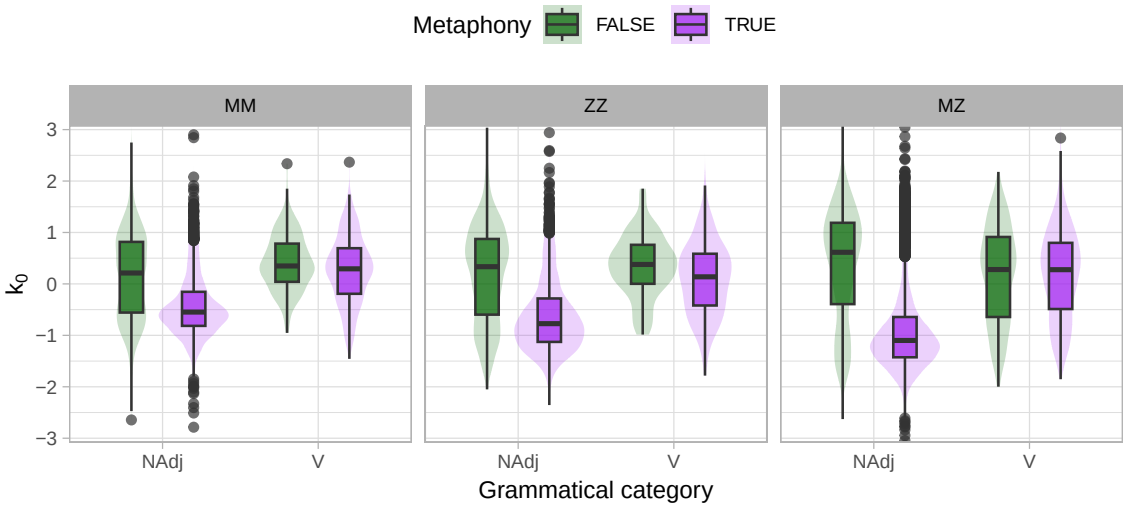



**Figure 3** – Distribution of $k_0$ DCT-scores for stem vowels /e, o/ shown separately by region and grammatical category. Presence vs absence of metaphonic context is colour-coded (TRUE = suffix vowels /i, u/; FALSE = suffix vowels /a, e/).

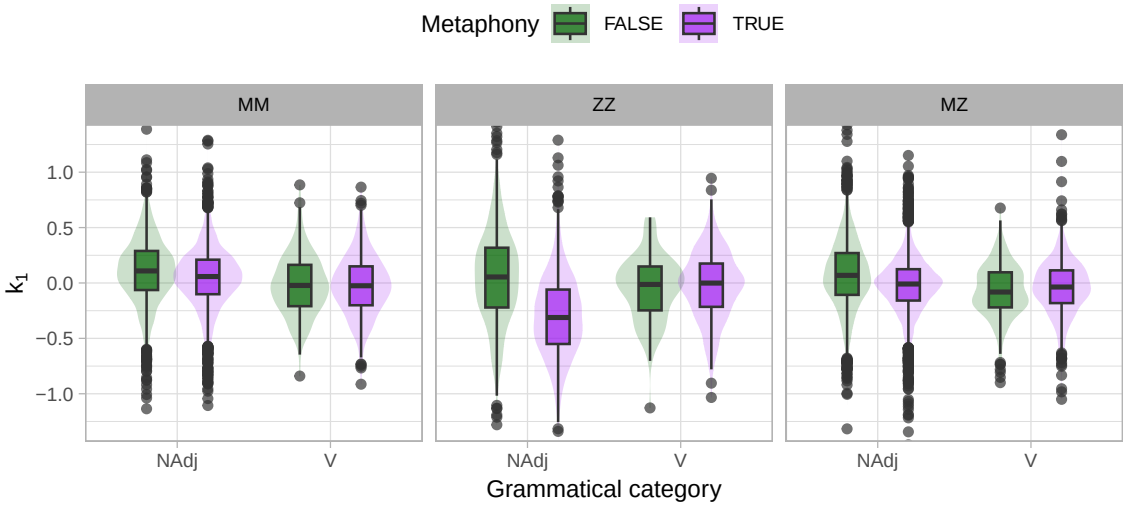



**Figure 4** – Distribution of $k_1$ DCT-scores for stem vowels /e, o/ shown separately by region and grammatical category. Presence vs absence of metaphonic context is colour-coded (TRUE = suffix vowels /i, u/; FALSE = suffix vowels /a, e/).

Figure 4 shows that the differences in $k_1$, which is proportional to the linear slope ($k_1$ values below 0 signal a more rising slope), are clearly evident in ZZ for nouns and adjectives but not for verbs. The statistical analysis showed also for $k_1$ a significant interaction between region, grammatical category, and metaphony ($F_{2,\ 17105.6} = 36.1$, $p < 0.001$). Post-hoc tests showed a significantly lower $k_1$ in metaphonic context in ZZ than both MM ($z = -17.3$, $p < 0.001$) and MZ ($z = -15.3$, $p < 0.001$). Consistently with Figure 4, the extent of $k_1$ lowering in metaphonic context was significantly greater in nouns and adjectives than verbs in ZZ only ($z = -4.1$, $p < 0.001$). Also, $k_1$ was significantly lower in metaphonic context for nouns and adjectives only in ZZ ($z = 30.5$, $p < 0.001$) and to a minor extent also in MZ ($z = 5.0$, $p < 0.001$).

## 4 Discussion

In line with previous analyses [15], these data confirmed that metaphony is present to different degrees across the three regions. The degree of raising of stem vowels (parameterised by $k_0$)

was mostly marked in MZ, least marked in MM, and intermediate in ZZ. In addition, the $k_1$ analysis revealed that metaphony in ZZ was mainly characterised by diphthongisation, in which F1 values are lower at the onset but rise towards the offset. The second main finding was that verbs showed overall less metaphony than both adjectives and nouns together. This was the case for all regions: even in MM, in which metaphony is more reduced than the other two regions, effects were still greater in nouns and adjectives than verbs.

A discrepancy between verbs and other grammatical categories in some sound changes has also been observed in other languages, although the direction and extent of the sound change can vary [7]. A possible explanation may have to do with the morpho-synctactic function of verbs: on the one hand, their central role in syntax and semantics makes them highly relevant and informative for both speaker and hearer; on the other hand, however, not all morphemes within a verb may carry the same functional load in all languages. For example, inflectional information may be redundant and easily retrievable from the context: this happens for instance in northern Italian varieties [9], in which lack of metaphony in the verb is functionally interpreted as a way to avoid linguistic redundancy, since other affixes already express inflectional information.

Along redundancy, another aspect that may explain why metaphony in the verbs of the Lausberg area does not take place is the hypothesis that coarticulation-based sound changes like metaphony originate from a trade-off of phonetic cues between stem and suffix vowels. In [15], it was observed that, between the three regions in the Lausberg area, a higher degree of suffix erosion corresponded to a greater degree of coarticulation in the stem. In general, verb suffixes in pro-drop languages like the Lausberg dialects carry information about person and number that has to be perceptually salient to the listener, whereas in nouns and adjectives inflectional information can be also retrieved word-externally through determiners and inflected modifiers. Additionally, in these varieties, the second-person suffix is always bisyllabic (/asi/ for the first conjugation and /isi/ for other conjugations), meaning that it cannot be completely eroded, as phonetic erosion primarily affects word-final vowels. It is possible, therefore, that the higher functional load on suffixes in verbs may inhibit their reduction and therefore cue-transfer to the stem. To confirm this hypothesis, however, a comparison of suffixes (word-final vowels) between nouns/adjectives and verbs would be needed.

Since a uni-directional mapping from phonetic variation to morphology cannot explain the different effects on sound change of different grammatical categories, then it is plausible that the lexicon can directly influence phonology [7, 3, 4]. In particular, the case of metaphony in the Lausberg area seems to be based on complex trading relationships that involve multiple layers of the grammar: the lack of external cues that may compensate for an information loss due to phonetic suffix erosion (or neutralisation) may be a great inhibitor for the change in words like verbs [26]. The ways in which the morpho-lexical component can interact with phonology can vary across sound changes and languages. Some changes may progress from phonetics to a higher degree of cognitive control through various forms of grammaticalisation [1, 2], as appears to be the case for metaphony in the Mittelzone [15]. Nevertheless, this does not preclude the possibility that higher modules in the grammar can bias sound change in a certain way, suggesting that sound change is not entirely independent of higher grammatical levels, such as morphology (e.g. [4]) and the lexicon [7, 3, 27].

Overall, these results suggest that the morpho-lexical level can influence sound change in a "top-down" manner. Future research should give increasing attention to lexically-specific phonetic effects and, more broadly, to evidence from diverse languages that suggests interactions between phonetics/phonology and higher linguistic levels.

## Acknowledgments

This study was co-funded by the European Union (ERC, *SoundAct*, project N° 101053194).

## References

[1] BERMÚDEZ-OTERO, R. and G. TROUSDALE: *Cycles and continua: on unidirectionality and gradualness in language change*. In *The Oxford handbook on the history of English*, pp. 691–720. Oxford University Press, Oxford, 2012.

[2] RAMSAMMY, M.: *The life cycle of phonological processes: Accounting for dialectal microtypologies*. *Language and Linguistics Compass*, 9(1), pp. 33–54, 2015. doi:10.1111/lnc3.12102.

[3] WEDEL, A.: *Lexical contrast maintenance and the organization of sublexical contrast systems*. *Language and Cognition*, 4(4), pp. 319–355, 2012. doi:10.1515/langcog-2012-0018.

[4] STRYCHARCZUK, P. and J. M. SCOBBIE: *Gradual or abrupt? the phonetic path to morphologisation*. *Journal of Phonetics*, 59, pp. 76–91, 2016.

[5] KIPARSKY, P.: *Phonologization*. In J. HONEYBONE and J. SALMONS (eds.), *The Oxford handbook of historical phonology*, vol. 1, pp. 563–582. Oxford University Press, Oxford, 2015.

[6] BLEVINS, J. and J. LYNCH: *Morphological conditions on regular sound change? A reanalysis of *l-loss in Paamese and Southeast Ambrym*. *Oceanic Linguistics*, pp. 111–129, 2009.

[7] SMITH, J. L.: *Category-Specific Effects*, chap. 102, pp. 1–25. John Wiley & Sons, Ltd, 2011. doi:https://doi.org/10.1002/9781444335262.wbctp0102. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444335262.wbctp0102`. `https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444335262.wbctp0102`.

[8] SERENO, J. A. and A. JONGMAN: *Acoustic correlates of grammatical class*. *Language and Speech*, 38(1), pp. 57–76, 1995.

[9] MAIDEN, M.: *Interactive morphonology: Metaphony in Italy*, vol. 16. Routledge, London/New York, 1991.

[10] LAUSBERG, H.: *Die Mundarten Südlukaniens*. Niemeyer, Halle an der Saale, 1939.

[11] RECASENS, D.: *Coarticulation and sound change in Romance*. John Benjamins, Amsterdam, 2014.

[12] RENSCH, K.-H.: *Beiträge zur Kenntnis nordkalabrischer Mundarten.* Aschendorffsche Verlagsbuchhandlung, Münster, 1964.

[13] MAIDEN, M. and L. M. SAVOIA: *Metaphony*. In P. M. MAIDEN M. (ed.), *The dialects of Italy*, pp. 15–25. Routledge, London/New York, 1997.

[14] ROHLFS, G.: *Grammatica storica della lingua italiana e dei suoi dialetti. Morfologia*, vol. 2. Einaudi, Torino, 1968.

[15] GRECA, P., M. GUBIAN, and J. HARRINGTON: *The relationship between the coarticulatory source and effect in sound change: evidence from italo-romance metaphony in the lausberg area*. *Laboratory Phonology*, 15, 2024. doi:10.16995/labphon.9228.

[16] DRAXLER, C.: *Percy – an HTML5 framework for media rich web experiments on mobile devices*. In *Proceedings of the 12th annual conference of the International Speech Communication Association (Interspeech 2011)*, pp. 3339–3340. Florence, Italy, 2011. URL `https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2011/i11_3339.pdf`.

[17] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 45, pp. 326–347, 2017. doi:10.1016/j.csl.2017.01.005.

[18] WINKELMANN, R., J. HARRINGTON, and K. JÄNSCH: *EMU-SDMS: Advanced speech database management and analysis in R*. *Computer Speech & Language*, 45, pp. 392–410, 2017. doi:10.1016/j.csl.2017.01.002.

[19] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

[20] ALBIN, A. L.: *Praat*R*: An architecture for controlling the phonetics software "*P*raat" with the* R *programming language*. *The Journal of the Acoustical Society of America*, 135(4), pp. 2198–2199, 2014. doi:10.1121/1.4877175.

[21] LOBANOV, B. M.: *Classification of* R*ussian vowels spoken by different speakers*. *The Journal of the Acoustical Society of America*, 49, pp. 606–608, 1971. doi:10.1121/1.1912396.

[22] WATSON, C. I. and J. HARRINGTON: *Acoustic evidence for dynamic formant trajectories in Australian English vowels*. *The Journal of the acoustical society of America*, 106(1), pp. 458–468, 1999.

[23] HARRINGTON, J.: *Acoustic phonetics*. In W. J. HARDCASTLE, J. LAVER, and F. E. GIBBON (eds.), *The handbook of phonetic sciences*, pp. 107–124. Wiley-Blackwell, Chichester, U.K. & Malden, MA, 2010.

[24] KUZNETSOVA, A., P. B. BROCKHOFF, and R. H. B. CHRISTENSEN: *lmerTest package: Tests in linear mixed effects models*. *Journal of Statistical Software*, 82(13), pp. 1–26, 2017. doi:10.18637/jss.v082.i13.

[25] LENTH, R. V.: *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024. URL `https://CRAN.R-project.org/package=emmeans`. R package version 1.10.2.

[26] BLEVINS, J. and A. WEDEL: *Inhibited sound change: An evolutionary approach to lexical competition*. *Diachronica*, 26(2), pp. 143–183, 2009.

[27] HAY, J. and P. FOULKES: *The evolution of medial /t/ over real and remembered time*. *Language*, 92(2), pp. 298–330, 2016. doi:10.1353/lan.2016.0036.

# DER EINFLUSS DER SPRECHGESCHWINDIGKEIT AUF DIE REALISATION DER BETONTEN VOKALE IM DEUTSCHEN

*Reinhold Greisbach[1], Kornélia Juhász[3,4,2], Zsuzsa Szánthó[2,3,4], Szabina Zsoldo[2], Andrea Deme[2,3,4]*

*[1]IfL - Phonetik, Universität zu Köln, [2]Eötvös Loránd University Budapest, [3]Hungarian Research Centre for Linguistics, [4]Lendület "Momentum" Neurophonetics Research Group*

**Kurzfassung:** Mithilfe der Variation der Sprechgeschwindigkeit wird untersucht, welcher der beiden Parameter Lautquantität oder Lautqualität eine größere Bedeutung für die phonologische Beschreibung der deutschen Vokale hat. Es zeigt sich, dass sich die Größe der Quantitätsopposition bei schnellerem Sprechen verringert, während die Qualitätsopposition stabil bleibt. Andererseits zeigt sich eine Expansion der Quantitätsopposition auf alle Laute der getesteten Minimalpaare bei normaler Sprechgeschwindigkeit, ein Phänomen, das auch beim schnelleren Sprechen erhalten bleibt. Eine genauere Analyse der zeitlichen und lautlichen Zusammensetzung der Minimalpaarwörter in Form einer relativen Positionskarte zeigt weiterhin, dass diese statistisch gemittelten Ergebnisse nicht die tatsächlichen Strategien der einzelnen Sprecher beim Beschleunigen ihrer Aussprache widerspiegeln.

## 1 Motivation

Phonetische und phonologische Beschreibungen des Lautinventars und der Lautstruktur einer Sprache beruhen auf der präzisen Aussprache von Sprechern dieser Sprache bzw. mechanisch gesehen auf einer Artikulation bei langsamer Sprechgeschwindigkeit. Um schneller zu sprechen, stehen den Sprechern unterschiedliche Möglichkeiten zur Verfügung. Sie können die Artikulation beschleunigen indem sie ihre Artikulatoren (Lippen, Zungenspitze, Zungenkörper, Velum) schneller bewegen. Weiterhin können sie die Bewegungen der Artikulatoren in ihrer Extension (Auslenkung) reduzieren, was z.B. zu einer Spirantisierung von Plosiven oder einer Zentralisierung bei Vokalen führt. Und zuletzt können sie einzelne Sprechgesten vollständig auslassen, d.h. Laute elidieren. In dieser Untersuchung wollen wir feststellen, welche der obigen Beschleunigungsstrategien bei der Aussprache der akzentuierten deutschen Vokale zu Anwendung kommen und wie sich dies auf die Gewichtung der Beschreibungsparameter der phonologischen Quantitäts- bzw. Qualitätsopposition des deutschen Vokalsystems auswirkt.

## 2 Vokalsystem des Deutschen

Im Deutschen können fast alle Vokale, die in akzentuierten Silben auftreten, zu Paaren zusammengefasst werden. Diese Paare unterscheiden sich in der Quantität (Lautdauer) und zumeist zusätzlich auch noch in einem weiteren phonetischen Merkmal. Was dieses weitere Merkmal ist und ob nicht sogar die unterschiedliche Lautdauer ein Nebeneffekt dieses weiteren Merkmals ist, wird in der deutschen Phonetik und Phonologie kontrovers diskutiert. In Frage kommen dafür die Merkmale: Qualität, Gespanntheit und Silbenschnitt (vgl. z.B. Becker [10]). In jedem Fall beobachten wir auf der phonetischen Ebene außer bei den tiefen Vokalen /a - aː/ zwei Unterscheidungsmerkmale gleichzeitig: Quantität und Qualität, die so zu einer Paarbildung bei /ɪ - iː/, /ʏ - yː/, /ʊ - uː/, /ɔ - oː/, /œ - øː/ sowie /ɛ - eː/ führen. In unserem Experiment untersuchen wir die Fragestellung, wie robust sich Quantität und Qualität bei den Eckvokalpaaren gegenüber einer Erhöhung der Sprechgeschwindigkeit erweisen. Für unsere akustische Analyse operationalisieren wir die Vokalquantität durch die Lautdauer (in ms) und die Vokalqualität durch die ersten beiden Formanten, F1 und F2 (in Hz).

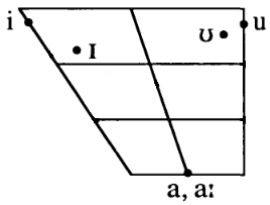



**Abbildung 1** – Deutsche Eckvokalpaare nach Kohler [5]. Alle anderen Vokale sind getilgt.

Abb.1 zeigt die Eckvokalpaare des Deutschen gemäß der auf phonetischer Transkription beruhenden Beschreibung des deutschen Lautsystems nach Kohler [5]. Die Vokalpaare unterscheiden sich in der Lautquantität (in Abb. 1 nur für [a:] notiert) und der Lautqualität. Dabei findet sich laut Kohler der stärkste Qualitätsunterschied bei den i-Vokalen, während sich die a-Vokale qualitativ nicht unterscheiden.

# 3 Experiment

In unserem Experiment wurden die untersuchten Vokale in Minimalpaare tatsächlicher gesprochener deutscher Wörter eingebettet und diese wiederum in einen Rahmensatz. Gewählt wurden die Paare *„Schiefe“-„Schiffe“, „biete“-„bitte“* für die i-Laute, *„Maße“-„Masse“, „wate“-„Watte“* für die a-Laute und *„Muße“-„Musse“, „Buße“-„Busse“* für die u-Laute mit dem eher ungebräuchlichen Plural des Wortes *„Muss“*, wie es in der Phrase *„Es ist ein Muss, dass …“* vorkommt.

Eingebettet wurden die Wörter in den Rahmensatz *„Mehr ZIELWORT bitte!“*. Die 12 Wörter kamen je 12 mal in randomisierter Folge vor. Im ersten Durchlauf wurden die Versuchspersonen gebeten mit normaler Sprechgeschwindigkeit zu sprechen, im zweiten Durchlauf sollten sie so schnell wie möglich sprechen. In einer Trainingsphase vor dem zweiten Durchlauf feuerte der Versuchsleiter die Sprecher mehrmals an schneller zu sprechen.

Insgesamt sprachen 19 Sprecher (10 weiblich, 9 männlich) des Standarddeutschen aus Nordrhein-Westfalen ohne regiolektale Auffälligkeiten im Altersbereich von 22 bis 63 Jahren die Sätze. Für die Analyse wurden alle 12 Realisationen mit normaler Geschwindigkeit und die 6 schnellsten Realisationen mit schneller Geschwindigkeit ausgewählt, da zu beobachten war, dass die Sprecher ihre maximale Sprechgeschwindigkeit nicht konstant über die Versuchsdauer beibehalten konnten. Die Dauer- und Formantanalyse erfolgte mit dem Programm PRAAT.

# 4 Ergebnisse

## 4.1 Sprechgeschwindigkeit

Der Versuchsaufbau erlaubt tatsächlich eine Steigerung der Artikulationsgeschwindigkeit. Die durchschnittliche Silbenrate stieg beim schnellen Sprechen von 5,69 Silben/Sekunde auf 7,67 Silben/Sekunde, was einer Beschleunigung auf 73,8% der benötigten Zeit relativ zur normalen Artikulationsgeschwindigkeit entspricht.

Es waren nahezu keine Elisionen und Gestenreduktionen zu beobachten. Die Beschleunigung scheint deshalb nur von der Fähigkeit der Sprecher ihre Artikulatoren schneller zu bewegen abhängig zu sein. Die Befähigung der Sprecher für diese Geschwindigkeitssteigerung unterschied sich jedoch sehr stark. Die langsamste Sprecherin (Denise) erreichte 4,44 Silben/Sekunde, der schnellste Sprecher (Peter) erreichte 6,58 Silben/Sekunde bei normalem Tempo. Bei schneller Sprache war der langsamste Sprecher 6,18 Silben/Sekunde schnell, der schnellste erreichte 9,94 Silben/Sekunde. Die Beschleunigung variierte zwischen 95,6 % (nahezu keine Beschleunigung) und 58,0% (nahezu Verdopplung der Artikulationsgeschwindigkeit). Es gab

keine Korrelation zwischen der Ausgangsgeschwindigkeit und der Beschleunigung, d.h. die Befähigung der Sprecher schneller zu sprechen, hängt nicht von ihrer normalen Sprechgeschwindigkeit ab.

## 4.2 Vokaldauer

Die Messungen der Vokaldauer der Eckvokale lassen eine Systematik in der Bildungsdauer der einzelnen Vokale erkennen. Es sind stets die offenen Vokale (sowohl kurz wie lang, sowohl normal wie schnell gesprochen), die am längsten dauern, gefolgt von den hinteren geschlossenen und den vorderen geschlossenen Vokalen, was aufgrund der physiologischen Produktionsbedingungen (unterschiedliche beteiligte Zungenmuskeln) zu erwarten und auch so für das Deutsche (z.B. Ramers [3]) und das Englische (z.B. Crystal & House [4]) dokumentiert ist.

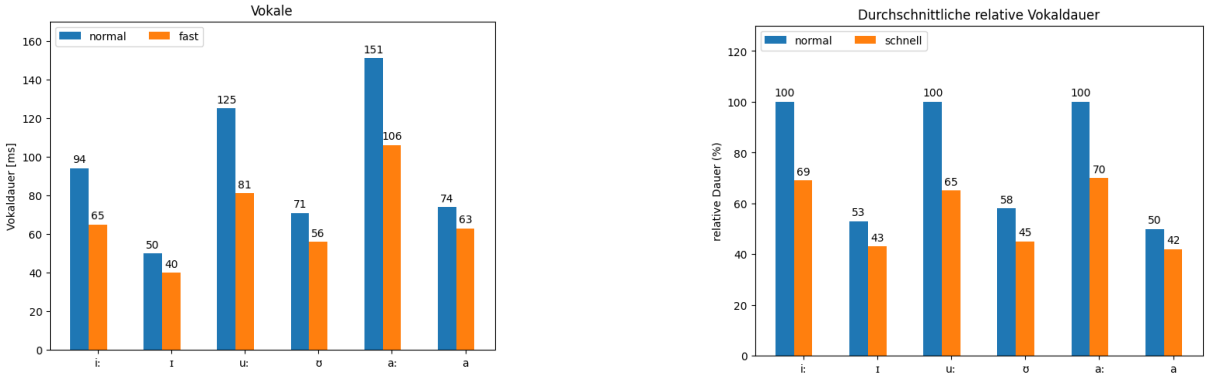



**Abbildung 2** – Durchschnittliche absolute (links) und relative (rechts) Lautdauer

Die relativen Dauerunterschiede zwischen den Vokalgruppen bleiben dagegen nahezu konstant (Abb 2, rechts). Für die Berechnung wurden die absoluten Dauern der normal gesprochenen langen Eckvokale jeweils auf 100% gesetzt. Es zeigt sich, dass die u-Vokale am stärksten von der Beschleunigung betroffen sind, die a-Vokale am geringsten. Im Durchschnitt über alle Vokalpaare werden die Langvokale auf 68,0% beschleunigt, die Kurzvokale nur auf 81,3% der normal gesprochenen Vokale.

Phonologisch interessiert jedoch in erster Linie der Dauerkontrast zwischen Lang- und Kurzvokalen, d.h. der Quotient Langvokal/Kurzvokal. Dieser beträgt im Schnitt bei normaler Geschwindigkeit 1,89. Bei schneller Geschwindigkeit dagegen nur noch 1,58. Der relative Dauerkontrast zwischen Lang- und Kurzvokalen reduziert sich also beim schnelleren Sprechen.

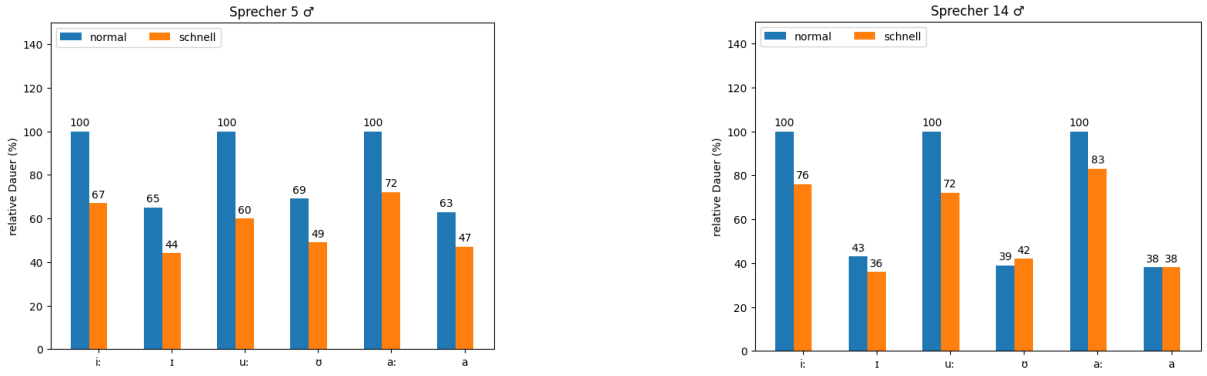



**Abbildung 3** – Relative Lautdauer von Sprecher 5 (links) und Sprecher 14 (rechts)

Die einzelnen Sprecher zeigen jedoch ein unterschiedliches Verhalten. So gibt es Sprecher, die bei normaler Artikulationsgeschwindigkeit einen starken Kontrast von 2,44 zwischen Lang- und Kurzvokal aufweisen (Abb. 3, rechts) und andere, bei denen ein eher schwachen Kontrast von nur 1,52 auftritt (Abb. 3, links). Die Sprecher mit starkem Kontrast reduzieren

typischerweise die Dauern der Kurzvokale beim schnelleren Sprechen nicht mehr, während die Sprecher mit schwächerem Kontrast die Dauern der Langvokale bis hin zur Dauer der normal gesprochenen Kurzvokale hin reduzieren.

## 4.3 Dauerkontrast in den Minimalpaaren

### 4.3.1 Expansion des Vokalkontrastes

In unserem Korpus liegen die Beschleunigungswerte von Lang- und Kurzvokalen relativ symmetrisch um den Durchschnittswert der Beschleunigung, der sich aus dem Vergleich der Silbengeschwindigkeiten von 73,8% ergibt. Dieser Durchschnittswert entspricht in etwa auch der durchschnittlichen Beschleunigung der restlichen Laute in den Minimalpaarwörtern [$C_1V(:)C_2ə$] (75,2% bei $C_1$ und 74,8% bei $C_2$ sowie 75,1% bei [ə]).

Setzen wir auch hier die Dauer der einzelnen Laute in den Minimalpaaren mit langem Vokal [$C_1V:C_2ə$] auf jeweils 100%, so zeigt sich ein Effekt, der über den erwarteten Dauerkontrast hinausgeht. Denn es ändern sich nicht nur die Lautdauern der betonten Vokale, sondern auch die der sie umgebenden Laute der Minimalpaare systematisch.

Beim Übergang von Lang- zu Kurzvokal verkürzt sich bei normaler Sprechgeschwindigkeit der Vokal auf 52,7% (Kontrast Langvokal/Kurzvokal q=1,90) aber auch der Onsetkonsonant der Silbe auf 92,4% (q=1,08). Gleichzeitig verlängern sich die Dauern der Laute der 2. Silbe verlängern, die des Konsonanten auf 104,2% (q=0,96) und die des Schwas sogar auf 117,8% (q=0,85). Bei erhöhter Sprechgeschwindigkeit bleibt dieser Effekt leicht abgeschwächt erhalten (Abb. 4 rechts).

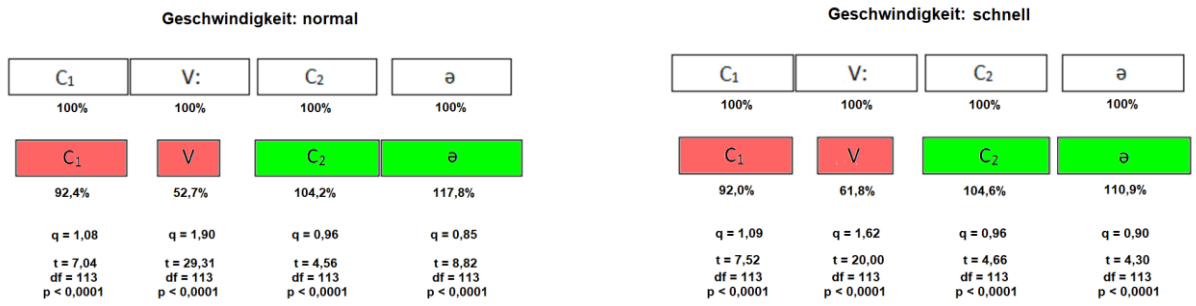



**Abbildung 4** – Gemittelte relative Dauerunterschiede in den Lauten der Minimalpaare [$C_1V:C_2ə$] (oben) vs. [$C_1VC_2ə$] (unten) für normale Geschwindigkeit (links) und schnelle Geschwindigkeit (rechts) sowie Längenkontrast q und Werte der gepaarten Tests über 6 Wortpaare × 19 Sprecher

Bei den Wörtern mit Langvokalen verlängert sich nicht nur der Vokal, sondern auch in geringerem Maße der Onsetkonsonant der Silbe, während sich die Laute der 2. Silbe gegenüber den Wörtern mit Kurzvokalen verkürzt. Diese Unterschiede sind im Rahmen gepaarter t-Tests jeweils signifikant (Abb. 4). Die Ergebnisse zeigen, dass sich die phonologische Opposition der deutschen Vokale zusätzlich in den relativen Dauern der umgebenden Laute innerhalb der Minimalpaare widerspiegelt und dies unabhängig von der Sprechgeschwindigkeit. Der Quantitätskontrast expandiert also über den eigentlichen Ort der phonologischen Opposition hinaus und führt zu einer unterschiedlichen zeitlichen Organisation des gesamten Minimalpaares.

### 4.3.2 Relative Positionskarte

Diese Organisation hängt sehr stark von der lautlichen Zusammensetzung des Wortes und von den spezifischen Artikulationsstrategien des Sprechers ab, wie eine andere Darstellung dieser Daten zeigt. Bei dieser Darstellung, hier „relative Positionskarte“ genannt, wird jedes Wort durch einen Punkt in einer zweidimensionalen Karte repräsentiert. Zur Lokalisierung des Wortes [masə] beispielsweise mit den Lautdauern 50 ms, 150 ms, 100 ms und 100 ms werden

zunächst die relativen Dauern gebildet, also 12,5% für [m], 37,5% für [a] sowie jeweils 25 % für [s] und [ə]. Anschließend werden den 4 Lauten des Wortes die Koordinaten der Eckpunkte eines Quadrates zugewiesen, hier also z.B. [m] = [-100|100], [a] = [100|100] , [s] = [-100|-100] und [ə] = [100|-100]. Zuletzt werden die relativen Dauern der Laute mit den jeweiligen Eckpunkten des Quadrates multipliziert. Dies ergibt dann in diesem Beispiel die Koordinaten des Wortes [masə] als [x|y] = 0,125*[-100|100] + 0,375*[100|100] + 0,25*[-100|-100] + 0,25*[100|-100] = [-25|0] (siehe Abb. 5, links).

Die Position eines Wortes in dieser Karte ist bei 4 Lauten nicht eindeutig (anders als bei 3 Lauten und einem gleichseitigem Dreieck der Eckpunkte), vermittelt aber einen Eindruck von der relativen Zeitstruktur der Wörter: Liegt das Wort weiter rechts, dann ist der vokalische Anteil im Wort größer, liegt es weiter links, der konsonantische Anteil. Liegt das Wort weiter oben, dann ist der Anteil der 1. Silbe stärker, liegt es weiter unten der Anteil der 2. Silbe.

Abb. 5 rechts zeigt die Lage der einzelnen Wörter des Korpus in der relativen Positionskarte für normale Sprechgeschwindigkeit gemittelt über alle Sprecher. Die Positionen der Lang- und Kurzvokalwörter sind klar getrennt: Die Kurzvokalwörter haben wie schon gesehen einen längeren Anteil der 2. Silbe und der konsonantischen Bestandteile. Deutlich sichtbar wird auch, dass die Wörter mit a-Vokalen am weitesten rechts liegen, also allesamt einen größeren vokalischen Anteil haben. Deutlich wird auch der unterschiedliche inhärente Daueranteil der einzelnen Konsonanten, wobei sich das wortinitiale [ʃ] als besonders prominent erweist und die Wortpunkte von *„Schiefe“* und *„Schiffe“* nach links oben Richtung $C_1$ zieht. Beim schnelleren Sprechen verändern sich die Lagen der Wörter nicht wesentlich.

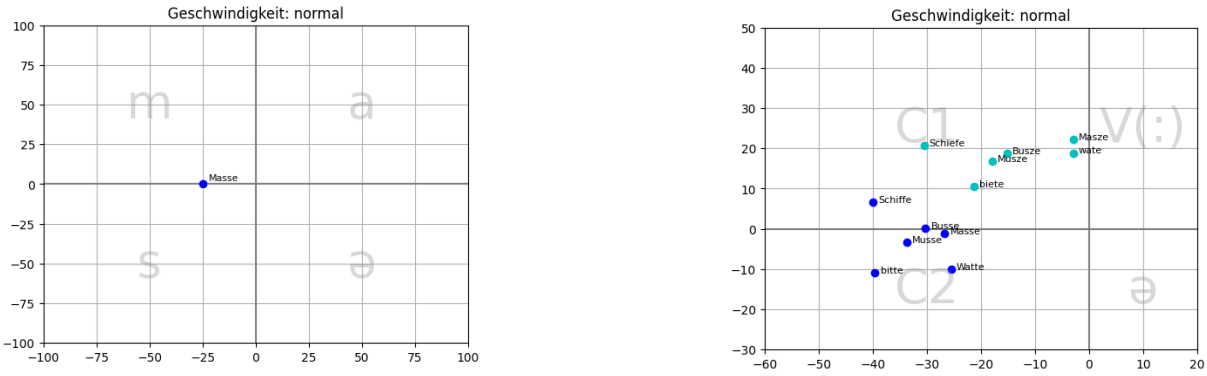



**Abbildung 5** – links: Relative Posititonskarte des Erklärungsbeispiels; rechts: Relative Positionskarte (gezoomt) der Minimalpaarwörter bei normaler Artikulationsgeschwindigkeit

Die relative Positionskarte erlaubt auch einen Blick auf die unterschiedliche zeitliche Organisation der Wörter bei den einzelnen Sprechern.

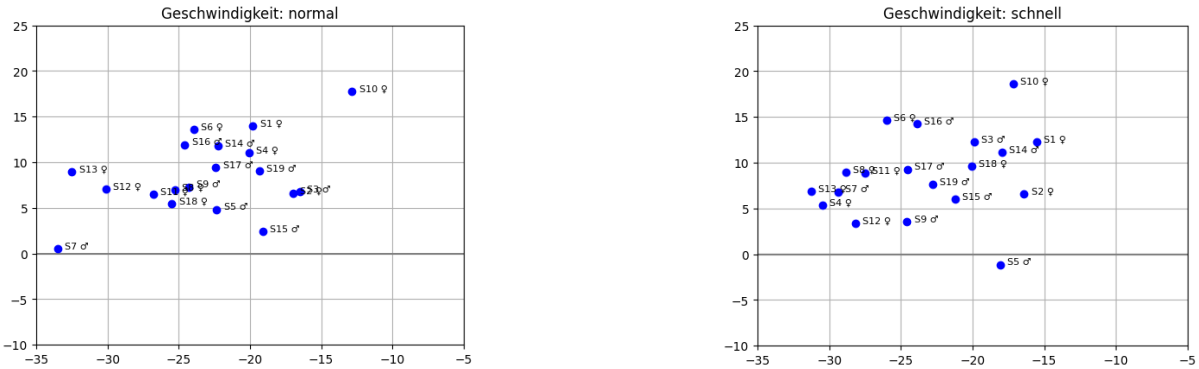



**Abbildung 6** – Relative Positionskarte (gezoomt) der Sprecherartikulationen bei normaler und schneller Artikulationsgeschwindigkeit

Dazu wird über alle Äußerungen eines Sprechers bei normaler Sprechgeschwindigkeit gemittelt. In Abb. 6, links stechen einige Sprecher besonders hervor. Sprecherin S10 weist im Mittel über alle Wörter sowohl mit Kurz- als auch mit Langvokalen den größten vokalischen Anteil in der akzentuierten Silbe im Vergleich zu allen anderen Sprechern auf. Dagegen zeigt sich bei Sprecher S7 ein nahezu gleicher Anteil der Dauer von betonter und unbetonter Silbe in den Wörtern bei gleichzeitig stark dominantem konsonantischem Anteil. Bei höherer Sprechgeschwindigkeit zeigen sich keine wesentlichen Unterschiede (Abb. 6, rechts).

Ein möglicher perzeptiver Effekt dieser sprecherspezifischen Eigenheiten könnte die Verständlichkeit der Sprecher sein, wobei Sprecher mit einem größeren vokalischen Anteil in den Wörtern verständlicher erscheinen als stärker konsonantisch orientierte Sprecher.

## 4.4 Vokalqualität

### 4.4.1 Formantnormierung

Die Vokalqualitäten wurden in dieser Untersuchung akustisch durch die ersten beiden Formanten F1 und F2 erfasst. Anders als bei den Lautdauern besteht bei den Formantfrequenzen jedoch eine starke Sprecherabhängigkeit. Um diese auszugleichen gibt es eine Reihe von Vorschlägen. So vergleichen beispielsweise Adank et al. [1] und Flynn & Foulkes [2] mehrere verschiedene Normalisierungsverfahren auf deren Effektivität.

Wir entscheiden uns hier für ein weiteres, speziell für unsere Datenlage angepasstes Normierungsverfahren, wo wir grob gesagt die Formantfrequenzen der normal gesprochenen langen Eckvokale für jeden Sprecher in die Ecken eines gleichseitigen Dreiecks platzieren und die anderen Vokale (normal gesprochene Kurzvokale und schnell gesprochene Lang- und Kurzvokale) als Linearkombinationen dieser Eckvokale berechnen. Der euklidische Abstand der Vokale vom Zentrum des ipsativen Dreiecks gibt dann den Grad der Zentralisierung bzw. die relative Expansion in % wieder. Eine genaue Beschreibung des Algorithmus findet man in Greisbach et al [9].

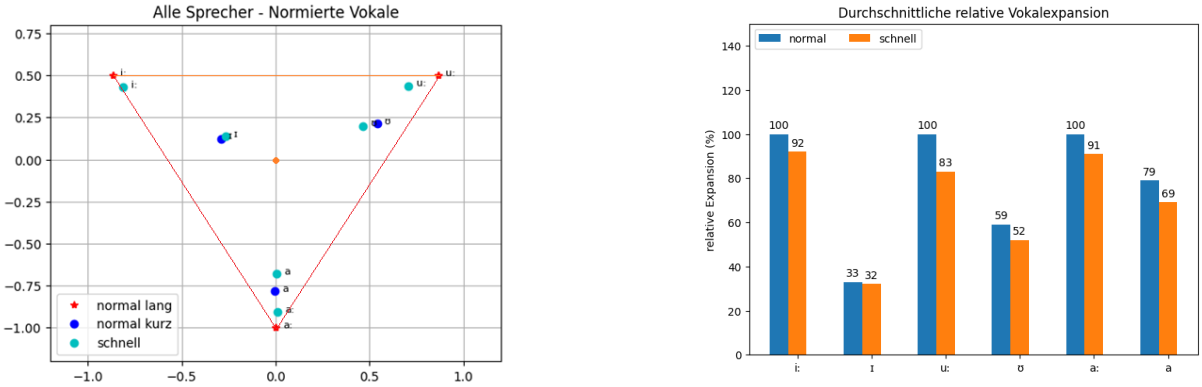



**Abbildung 7** – Links: Durchschnittliche ipsative Formantkarte: Der Abstand der Eckpunkte (rote Sterne) zum Zentrum (orangener Punkt) ist gleich. Rechts: Relative Expansion der Vokale gemittelt über alle Sprecher

### 4.2.2 Vokalkontrast

Bei normaler Sprechgeschwindigkeit zeigt sich der größte Qualitätsunterschied zwischen den i-Vokalen, ein Resultat, dass die Beschreibung von Kohler (Abb. 1) stützt. Aber anders als dies in seiner transkriptionsbasierten Vokallokalisierung im Vokalviereck zum Ausdruck kommt, zeigt sich hinsichtlich unserer Messgrößen einen Lageunterschied zwischen /a:/ und /a/. Dieser ist für den deutschen Muttersprachler nicht perzipierbar, findet sich aber genauso in anderen Formantkarten des Standarddeutschen (vgl. z.B. Sendlmeier & Seebode [8]).

Unter den Langvokalen wird beim schnelleren Sprechen das /u:/ am stärksten in Richtung Zentrum verschoben. Bei den Kurzvokalen ist es das /a/, gefolgt vom /ʊ/. Der Kontrast (Quotient von Lang- und Kurzvokal) ist sehr stark von der Vokalqualität abhängig. Am größten ist er zwischen dem i-Vokalpaar am geringsten zwischen dem a-Vokalpaar. Der Kontrast bleibt bei schnellerem Sprechen für jedes Paar jeweils quantitativ gleich.

Auch hier ergeben sich wieder Muster, die auf unterschiedliche sprecherspezifische Strategien für das schnellere Sprechen hinweisen. So verschieben sich bei Sprecher 5 fast alle Vokalqualitäten Richtung Zentrum, was darauf hindeutet, dass er seine Zunge bei schnellerem Sprechen nicht mehr soweit auslenkt (Sprechgeschwindigkeit steigt von 5,34 auf 7,07 Silben/Sekunde). Dagegen verschieben sich die Formanten des Sprechers 14 abgesehen von den u-Vokalen kaum, obwohl sich seine Sprechgeschwindigkeit von 6,34 auf 8,03 Silben/Sekunde erhöht.

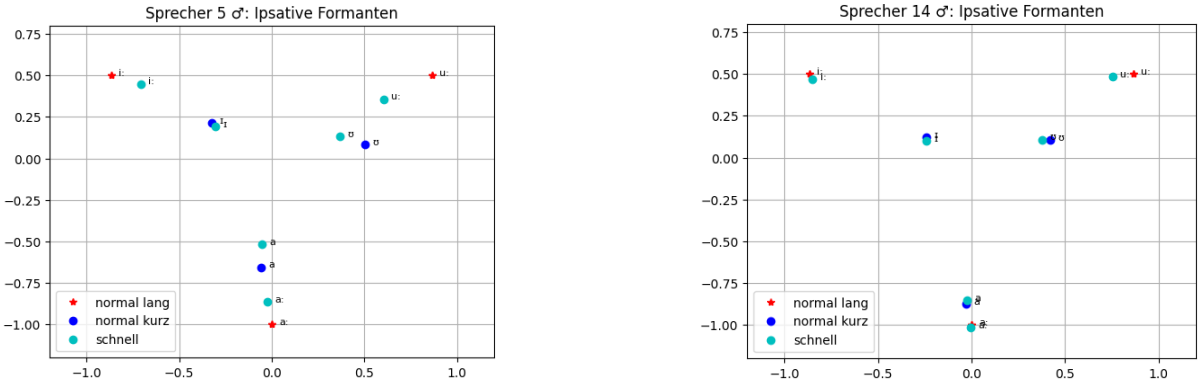



**Abbildung** 8 – Relative Positionierung der Vokale in den ipsativen Formantkarten für Sprecher 5 und 14

Bei diesem Sprecher zeigt sich sogar ein Overshoot: Der Langvokal /a:/ ist bei höherer Sprechgeschwindigkeit etwas weiter vom Zentrum entfernt (102%) als das /a:/ bei normaler Geschwindigkeit (Abb. 9, rechts). Einen vergleichbaren Effekt beobachten auch Siebenhaar & Hahn [6] bei einigen ihrer Sprecher und erklären ihn mit der kleinen Ausgangsgröße des jeweiligen Sprecher-Vokalraumes bei langsamem Sprechtempo.

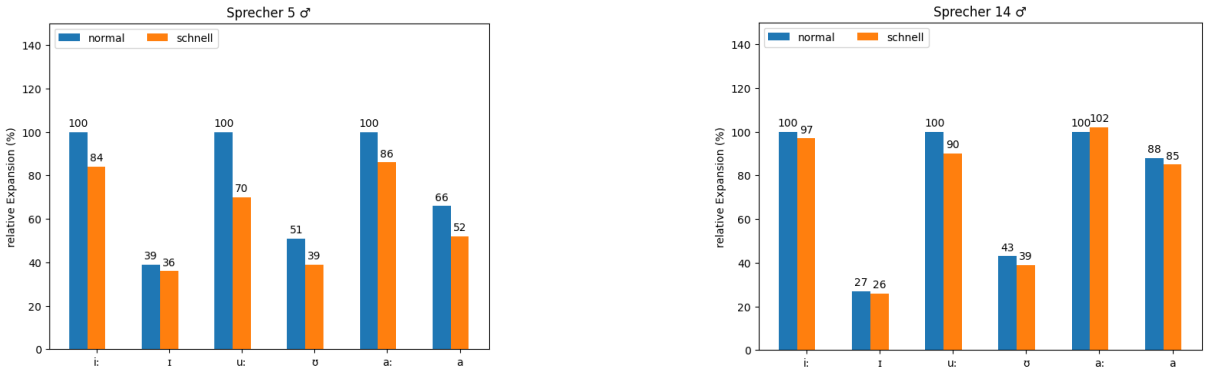



**Abbildung 9** – Relative Expansion der Vokale für Sprecher 5 und 14

In Abb. 3, 8 und 9 sind nur zwei Sprecher herausgegriffen, um deren Abweichungen vom statistischen Mittel exemplarisch zu dokumentieren. Tatsächlich lässt sich keine Typologie der Sprecher finden: Es gibt keinen Zusammenhang der Artikulationsstrategien etwa der Art: Wenn größerer Dauerkontrast, dann geringerer Qualitätskontrast, o.ä. Stattdessen verhalten sich die Sprecher je nach Vokalpaar und sogar je nach Minimalpaar individuell verschieden!

## 5 Diskussion

Mit der im Experiment gewählten Methode lässt sich die Sprechgeschwindigkeit tatsächlich erhöhen. Der Grad der Beschleunigung ist aber sehr stark abhängig von der Befähigung der Sprecher, ihre Artikulatoren schneller zu bewegen. Die Quantitätsopposition (Lautdauern) zwischen Lang- und Kurzvokalen ist unabhängig von den untersuchten Vokalpaaren etwa gleich und bleibt bei Erhöhung der Sprechgeschwindigkeit bestehen, reduziert sich aber in quantitativer Hinsicht. Tatsächlich lässt sich der Längenkontrast auch in den umgebenden Lauten der Minimalpaare beobachten, ein Phänomen, das auch bei schnellerem Sprechen bestehen bleibt. Bei der Qualitätsopposition (Vokalformanten) bestehen zwischen den Vokalpaaren unterschiedlich große Kontraste: am größten bei den i-Vokalpaaren, am geringsten bei den a-Vokalpaaren. Bei schnellerem Sprechen reduzieren sich die quantitativen Werte der jeweiligen Kontraste anders als bei der Längenopposition nicht.

Dies würde bedeuten, dass sich die Qualitätsopposition als robuster gegenüber einer Erhöhung der Sprechgeschwindigkeit erweist als die Quantitätsopposition. Andererseits dokumentiert die Ausdehnung des Längenkontrastes auf die zeitliche Struktur des gesamten Minimalpaares die Bedeutung der Quantitätsopposition.

Diese Ergebnisse bilden nicht unbedingt die Strategien der Sprecher ab, die diese verwenden, um schneller zu sprechen. Zu unterschiedlich sind die Ausgangsgeschwindigkeit und die erreichte Beschleunigung beim schnelleren Sprechen, zu unterschiedlich ist die Gewichtung von Quantitäts- und Qualitätskontrast zwischen den Sprechern bei Ausgangsgeschwindigkeit und schnellerer Artikulation bis hin zur sprecherspezifischen zeitlichen Komposition der Minimalpaarwörter, die auf eher vokalische vs. eher konsonantische Sprechertypen hindeuten.

## Dank

Diese Untersuchung entstand im Rahmen des durch TKA und DAAD geförderten Projektes: „Phonetische Aspekte schneller Sprache im Ungarischen und Deutschen“

## Literatur

[1] ADANK, P.M., R. SMITS, R. VAN HOUT: A comparison of vowel normalization procedures for language variation research. JASA 116 3099-3107 (2004).
[2] FLYNN, N., FOULKES, P.: Comparing Vowel Formant Nomalization Methods. In: Proc. of the 17th Int. Congr. of Phon. Sciences, Hongkong, 683-686 (2011).
[3] RAMERS, K.H.: *Vokalquantität und -qualität im Deutschen*. Niemeyer, Tübingen 1988
[4] CRYSTAL, T.H., A.S. HOUSE: *The duration of American-English vowels: an overview*. Journal of Phonetics 16, 263-284 (1988).
[5] KOHLER, K.J.: *"German"*. Journal of the International Phonetic Association, 20: 48–50 (1990).
[6] SIEBENHAAR, B., M. HAHN: *Vowel Space, Speech Rate and Language Space*. In: S. Calhoun ea. (Hrsg.): Proc. of the 19th Int. Congr. of Phon. Sciences, Melbourne, 879-883 (2019).
[7] HERMES, A., S. TILSEN, R. RIDOUANE: *Cross-linguistic timing contrast in geminates: A rate-independent perspective.* 12th International Seminar on Speech Production, Dec 2020, virtual, United States. ⟨halshs-03083573⟩
[8] SENDLMEIER, W.F., J. SEEBODE: *Formantkarten des deutschen Vokalsystems*. https://www.static.tu.berlin/fileadmin/www/10002019/Forschung/Formantkarten_des_deutschen_Vokalsystems_01.pdf (2006).
[9] GREISBACH, R., K. JUHÁSZ, Z. SZÁNTHÓS, S. ZSOLDOS, A. DEME: *Durational Aspects of fast speech in German.* Proceedings of the 5th ISAPh, Tartu (2024)
[10] BECKER, T.: Das Vokalsystem der deutschen Standardsprache. Lang, Frankfurt a.M. 1998

# WICKELN, FÜTTERN, SPIELEN – AUFNAHME UND TRANSKRIPTION VON ELTERN-BABY-GESPRÄCHEN IM ALLTAG

*Lisa Hartley[1], Eva Thoma[1], Christoph Draxler[1]*

*[1]Ludwig-Maximilians-Universität München*
*Lisa.hartley@campus.lmu.de*

**Kurzfassung:** Ziel dieser Analyse ist die Leistung KI-gestützter Spracherkennung für die Transkription natürlicher, geräuschvoller Eltern-Kind-Interaktionen (EKI) zu bewerten. Dazu wurden durch das Modell WhisperX generierte und manuell validierte Transkripte mit vollständig manuell erstellten Transkripten verglichen. Die zu transkribierenden Sprachdaten stammen aus dem Promotionsprojekt der ersten Autorin, für das 307 EKI in Alltagssituationen aufgenommen wurden.

Die Vorergebnisse dieser laufenden Analyse zeigen, dass die Nachbearbeitung der KI-generierten Transkripte, insbesondere die Korrektur der Segmentgrenzen, durchschnittlich zeitaufwendiger ist als eine vollständige manuelle Transkription. Dies deutet darauf hin, dass KI-Systeme derzeit Schwierigkeiten haben, lange und geräuschintensive Aufnahmen korrekt zu segmentieren und zu diarisieren. Dieser Beitrag macht allerdings auch deutlich, dass weitere Analysen erforderlich sind, um die Effizienz der KI für diese speziellen Sprechsituationen umfassender bewerten zu können.

## 1 Einleitung

Künstliche Intelligenz (KI) wird in vielen Bereiche eingesetzt, um bestimmte Arbeitsschritte zu vereinfachen, zu beschleunigen oder überhaupt zu ermöglichen. Die Transkription von Sprachdaten ist eine der Tätigkeiten, die enorm von den Fortschritten der KI profitieren kann. So ist es beispielsweise möglich, nahezu sofortige Untertitel oder Übersetzungen gesprochener Texte zu erstellen. Zwar ist die Ausgabe nie fehlerfrei, doch in vielen Situationen ist die Qualität so hoch, dass es die menschliche Be- und Nacharbeitung erheblich erleichtern kann. Am besten funktioniert dies allerdings mit Sprachproben, auf denen ein KI-Modell intensiv trainiert wurde.

Dieser Beitrag geht der Frage nach, wie hilfreich KI für die Transkription weniger optimaler Sprechsituationen ist, wie etwa bei spontanen Eltern-Kind-Interaktionen (EKI) in natürlicher und teilweise geräuschvoller Umgebung. Zu diesem Zweck wird die Leistung von WhisperX [1] bei der Transkription von alltäglichen EKI, die im Rahmen einer Doktorarbeit aufgenommen wurden, evaluiert und mit der Ausgabe einer erfahrenen Transkribiererin verglichen.

## 2 Daten: Ausschnitt aus dem „Gendered Baby Talk“ Projekt

Der Forschungsgegenstand des Dissertationsprojekts, aus dem die Daten für diesen Beitrag stammen, ist alltägliche an Kleinkindern gerichtete Sprache. Untersucht wird, inwiefern sich elterliche Ansprachen an Jungen und Mädchen unterscheiden, in Bezug auf Tonhöhe, Tonhöhevariation, Sprechgeschwindigkeit und Größe des Vokalraums. Ziel war es, ein Sprachkorpus aufzubauen, das den linguistischen Input eines vorsprachlichen Kindes möglichst realistisch abbildet. Aus diesem Grund wurde großer Wert auf die Alltäglichkeit, Spontaneität und Authentizität der Äußerungen während den Aufnahmen gelegt.

Fünfundzwanzig Mütter und sieben Väter von 7- bis 10-monatigen Kindern nahmen sich zu Hause mit ihrem eigenen Smartphone während alltäglicher Aktivitäten wie wickeln, füttern und spielen auf. In einer Vorbesprechung wurden die Teilnehmer*innen darum gebeten, ihren Alltag und ihre Routine nicht zu verändern und die Aufnahmen in den entsprechenden Situationen zu machen, wenn sie allein zu Hause sind, keinen Stress oder Zeitdruck haben und das Kind

zufrieden und kooperativ ist. Die Forscherin wies des Weiteren darauf hin, Geräte wie Radio, Fernseher oder Waschmaschine während den Aufnahmen auszuschalten. Dies wurde jedoch nicht systematisch eingehalten. Den Teilnehmer*innen wurden selbst überlassen, zu welcher Tageszeit die Aufnahmen gemacht wurden, in welcher Reihenfolge und in welchem Abstand. Auch der Inhalt sowie die Länge der Aufnahmen sind den Teilnehmer*innen freigelassen worden. Die Anweisungen lauteten beispielsweise: „Solange gefüttert wird“ oder „Womit ihr normalerweise spielt“. Um die Spielsituationen anzuregen, wurde jeder teilnehmenden Familie eine Auswahl an Spielsachen für die Dauer der Studie zur Verfügung gestellt. Um deren Verwendung während den Aufnahmen zu ermutigen, aber nicht zu erzwingen, wurde dies von der Studienleiterin nur mit „Damit ihr auch etwas Schönes von der Teilnahme an der Studie habt“ kommentiert.

Insgesamt wurden 307 Eltern-Kind-Interaktionen (EKI) mit einer Gesamtdauer von ca. 60 Stunden gesammelt. Da die Aufnahmebedingungen weitgehend den Sprecher*innen selbst überlassen worden sind, zeigen sich große Variationen in der Länge der Aufnahmen (von 2 bis 50 Minuten), in der Audioqualität (mit unterschiedlich vielen Hintergrundgeräuschen) und im sprachlichen Inhalt. Zusätzlich nahm jede*r Sprecher*in an einem etwa 30-minütigen Leitfadeninterview teil, wodurch insgesamt rund 17 Stunden an „Adult-Directed Speech“ Proben gesammelt wurden. Der Umfang der gesammelten Daten ist in Tabelle 1 ausgeführt.

**Tabelle 1 – Umfang der Datenerhebung für das „Gendered Baby Talk“ projekt**

| **Interaktionssituation** | **Anzahl** | **Gesamtdauer (ss:mm)** | **durchnittliche Dauer (min)** |
|---|---|---|---|
| Wickeln | 102 | 09:05 | 5 |
| Füttern | 100 | 25:17 | 15 |
| Spielen | 105 | 26:02 | 13 |
| **Gesamt EKI** | **307** | **60:25** | **11,5** |
| Interview | 32 | 17:07 | 32 |

Der auf Authentizität und Spontaneität ausgerichtete Datenerhebungsprozess führte dazu, dass die Sprachdaten viele Nebengeräusche enthalten. In dieser Hinsicht stellt sich die Frage, ob die aktuellen automatischen Spracherkennungssysteme bereits in der Lage sind, besser mit der Transkription solcher Daten umzugehen als menschliche Transkribierer*innen. Die Bewertung der automatisch erstellten Transkripte kann sowohl quantitativ ausgewertet werden, indem die Zeit berechnet wird, die den menschlichen Transkribierern eingespart wurde, als auch qualitativ, beispielsweise mit der Berechnung der Levenshtein-Distanz. In diesem Beitrag wird nur auf den quantitativen Aspekt eingegangen.

# 3 Transkription

Ein Ausschnitt des gesamten Korpus wurde ausgewählt, welcher fünf Interviews sowie 25 Wickelsituationen und 22 Füttersituationen umfasst. Die ausgewählten Aufnahmen wurden in drei Szenarien aufgeteilt.

## 3.1 Szenario 1: manuelle Vortranskription

Bis auf die 22 Füttersituationen, die für Szenarien 2 und 3 verwendet wurden, wurde das gesamte Korpus mit ELAN [2] bearbeitet. Die Redebeiträge der Eltern, Kindern und der Interviewerin wurden getrennt und in Äußerungen segmentiert. Die Sprechbeiträge der Eltern wurden orthographisch transkribiert, wobei die Transkriptionsrichtlinien sich an der *literarischen*

*Umschrift ohne Glättung* von Fuß und Karbach [3] mit eigenen Anpassungen orientieren. Die manuell erstellten Transkripte der ausgewählten 5 Interviews und 25 Wickelsituationen wurden anschließend in den Online-Editor Octra [4] importiert, in welchem sie dann jeweils manuell von einer erfahrenen Transkribiererin validiert wurden.

### 3.2 Szenario 2: automatische Vortranskription

Für 12 der verbleibenden Füttersituationen wurden mit WhisperX [1] orthografische Transkripte automatisch angefertigt. Die Ausgaben wurden ebenfalls in OCTRA importiert und von derselben Transkribiererin korrigiert. Das Ergebnis sind diarisierte und zeitlich ausgerichtete Transkripte der Sprechbeiträge der Erwachsenen.

### 3.3 Szenario 3: ohne Vorlage

Die Audiodateien der restlichen 10 Füttersituationen wurden ohne Vorlage in OCTRA importiert und vollständig von der Transkribiererin bearbeitet. Sie segmentierte die Beiträge der Eltern in Äußerungen und erstellte eine orthographische Transkription. Dabei folgte sie denselben Transkriptionsrichtlinien wie die in ELAN erstellten Vortranskripte aus Szenario 1 (§3.1).

## 4 Vorläufige Ergebnisse

Die aktuelle Auswertung stützt sich auf die automatisch erstellten Fortschrittsprotokolle in OCTRA. Da es angenommen wird, dass lange Pausen auf Unterbrechungen der Transkriptions- bzw. Validierungsarbeit (Beispielweise wegen einer Mittagspause) zurückzuführen sind, wurden die Fortschrittsprotokolle um Pausen von mehr als 10 Minuten bereinigt [5]. Die Transkriptionsfaktoren, die für Validierungs- und Transkriptionsaktivität berechnet wurden, sind in Tabelle 2 wiedergegeben. Die Transkriptionsfaktore für die Wickel- und für die Füttersituationen sind für eine manuelle Validierung bzw. sogar Neuerstellung eines Transkripts sehr niedrig. Im Durchschnitt hat die Transkribieren 2,7 Minuten gebraucht, um eine Minute aus den Füttersituationen zu transkribieren (Szenario 3, Siehe §3.3) und nur 1,34 Minuten für die Validierung einer Minute der manuellen vortranskribierten Wickelsituationen (Szenario 1, Siehe §3.1). Dagegen sind die Transkriptionsfaktoren für die Validierung der Interviews und der automatisch bearbeiteten Füttersituationen höher.

**Tabelle 2 – T-Faktor (in Octra) und Anteil an elterliche Beiträge**

| Situation | Anz. | Szenario – Tool | Tool für die Valiedierung | T-Faktor | Erwachsene Sprache (%) |
|---|---|---|---|---|---|
| Interview | 5 | Szenario 1 – ELAN | Octra | 6,7 | 63,0 |
| Wickeln | 25 | Szenario 1 – ELAN | Octra | 1,34 | 8,4 |
| Füttern | 12 | Szenario 2 – WhisperX | Octra | 5,12 | 25,2 |
| | 10 | Szenario 3 – Octra | - | 2,70 | 13,4 |

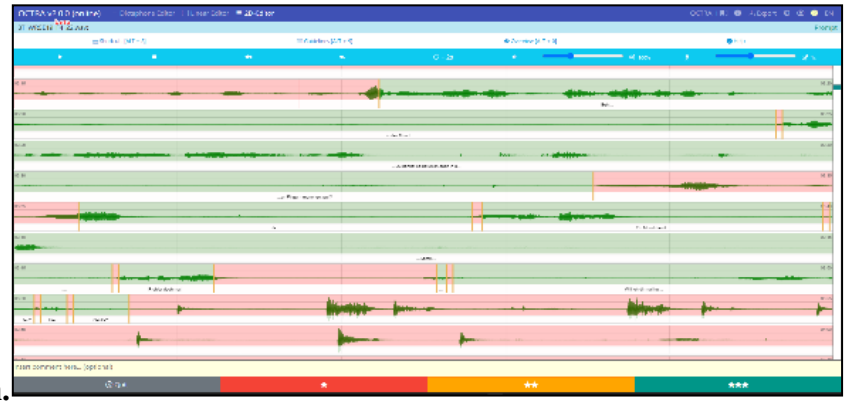
Der im Vergleich zu den Wickelsituationen hohe Arbeitsaufwand bei der Validierung der Interviewtranskripte (Transkriptionsfaktor = 6,7) innerhalb desselben Szenarios (Siehe §3.1) lässt sich durch die Situation erklären: Die Sprache in den Interviews zeichnet sich nämlich im Vergleich zu typisch an Kinder gerichteter Sprache durch längere Sätze, mehr Unsicherheiten, komplexere Wörter und Syntax sowie ein schnelleres Sprechtempo aus. Zudem wurde insgesamt deutlich mehr gesprochen, da bei den Interviews 63 % der Aufnahmedauer auf die Sprechbeiträge der Teilnehmer*innen entfallen, aber nur 8,4 % bei den Wickelsituationen. Dies führt zwangsläufig zu einem höheren Arbeitsaufwand pro Zeiteinheit bei den Interviews.

Besonders interessant sind die Ergebnisse der Füttersituationen. Der Transkriptionsfaktor bei der Validierung der automatisch erstellten Transkripte (Szenario 2, Siehe §3.2) ist mit 5,12 fast doppelt so hoch wie der für die neuerstellten Transkripte (Szenario 3). Die Transkribiererin hat also länger für die Korrektur der Ausgabe der KI gebraucht, als sie für die Erstellung eines vollständigen neuen Transkripts brauchte. Dafür gibt es zwei mögliche Erklärungen.

Erstens, obwohl die Füttersituationen zufällig aufgeteilt wurden, ist der Sprechanteil der Erwachsenen nicht gleichmäßig verteilt. In der Gruppe von Aufnahmen, die mit WhisperX vortranskribiert wurden (Szenario 2), sprechen Eltern durchschnittlich 25,2 % der Zeit, während sie in den Aufnahmen, die ohne Vorlage transkribiert wurden (Szenario 3), nur 13,4% der Aufnahmezeit sprechen.

Zweitens macht ein Blick auf die Ausgabe von WhisperX (Abbildung 1a) allein schon deutlich, welche Bearbeitungsschritte der Transkribiererin besonders zeitintensiv waren: Die die meisten Segmentgrenzen mussten korrigieren werden, um zum finalen Ergebnis zu kommen (Abbildung 1b.)

**1a.**



**1b.**

**Abbildung 1 –** Oben 1a: Ausgabe des Vortranskripts mit WhisperX. Unten 1b: Nach Validierung in Octra desselben Ausschnitts.

## 5 Interpretation und weitere Schritte

Diese vorläufigen Ergebnisse deuten darauf hin, dass die an Kinder gerichtete Sprache, im Vergleich zu Gesprächen zwischen erwachsenen Sprecher*innen, von Natur aus einfacher zu transkribieren ist. Außerdem deuten diese Ergebnisse darauf hin, dass automatische Spracherkennungssysteme lange und geräuschvolle Aufnahmen (noch) nicht gut segmentieren können. In diesem Fall scheint die manuelle Transkription zeiteffizienter zu sein.

Um jedoch aussagekräftige Ergebnisse zu erzielen, sind einige zusätzliche Analyseschritte erforderlich. Die Autor*innen planen, die Transkriptionsfaktoren anhand des Sprechanteiles zu

normalisieren. Zudem soll auch die Qualität der maschinellen Vortranskripte (z.B. Berechnung der Levenshtein-Distanz) ausgewertet werden. Schließlich wäre eine genauere Analyse der Validierungsschritte (Korrektur der Segmentgrenzen vs. Korrektur des Textes) aufschlussreich.

## 6 Fazit

Die vorliegende Untersuchung deutet daraufhin, dass die automatische Spracherkennung, insbesondere im Hinblick auf spontane Eltern-Kind-Interaktionen und geräuschvolle Aufnahmen auf Herausforderungen stößt, die ihre Effektivität beeinträchtigen können. Dies betrifft insbesondere die Sprecherdiarisierung und die Segmentierung von Redebeiträgen. In diesem Bereich ist womöglich eine manuelle Arbeit (derzeit noch) zeiteffizienter. Um aussagekräftige Ergebnisse zu erhalten, sind jedoch weitere Auswertungen erforderlich.

## 7 Literatur


[1] BAIN, M., J. HUH, T. HAN, und A. ZISSERMAN: *WhisperX: Time-Accurate Speech Transcription of Long-Form Audio*, Proc. Interspeech, Dublin, 2023.

[2] SLOETJES, H., A. RUSSEL, und A. KLASSMANN: *ELAN: a free and open-source multimedia annotation tool*. Proc. Interspeech 2007, Antwerpen, 2007.

[3] FUß, S., und U. KARBACH: *Grundlagen der Transkription*. Opladen, Toronto: Verlag Barbara Budrich, 2019.

[4] PÖMP, J., und C. DRAXLER: *Octra – a configurable browser-based editor for orthographic transcription*. P&P 2017, Berlin, 2017

[5] DRAXLER, C.: *Analysis of Transcriptions Using Octra – A Pilot Study*. Studientexte zur Sprachkommunikation, Elektronische Sprachsignalverarbeitung 2023, S. 17-23, 2023. https://www.essv.de/pdf/pdf/2023_17_23.pdf

# THE INFLUENCE OF VISUAL CONTEXT ON THE PERCEPTION OF VOICE ASSISTANT GENDER

*Jiaman He*[1] *and Iona Gessinger*[2]
[1]*RMIT University, Australia,* [2]*ADAPT Centre, University College Dublin, Ireland*
*s4125570@student.rmit.edu.au, iona.gessinger@ucd.ie*

**Abstract:** The use of voice assistants (VAs) with default female voices has raised concerns about reinforcing the image of women tending towards submissive societal roles. To address this bias, some companies are working on gender-neutral voices, such as Apple Siri's American English voice "Quinn." This study explores whether users perceive "Quinn" as neutral and if visual context influences their perception. A total of 73 native speakers of Chinese watched a video of "Quinn" uttering the same content under three conditions: with neutral, female, or male visual context. Participants perceived "Quinn" as more female across all conditions, yet especially when a female visual context was present. A correlation between perceived gender and likability of the voice was not found. Achieving gender neutrality is challenging, as perception depends on the interpretation of the context in addition to the voice itself. When addressing gender in VAs, we recommend considering the influence of visual factors on gender perception.

## 1 Introduction

Voice assistants (VAs) like Apple's Siri or Amazon Alexa are becoming an integral part of many people's everyday lives [1]. They often come with female names and voices as their default setting. Researchers and policy makers have raised concerns about the societal implications of this gendering phenomenon. As VAs predominantly engage in domestic tasks and comply with user commands, subconscious associations may arise, perpetuating the perception of women as submissive in various social scenarios [2, 3, 4]. Some have even called for legislative regulation around VA gender, since it may infringe on the right to equality [5]. Companies have come under scrutiny for perpetuating gender bias through their VA voices. In 2019, GenderlessVoice released "Q", a gender-neutral voice "created to end gender bias in AI assistants" [6], challenging the big tech companies to offer a non-binary voice option for their VAs. In 2022, Apple introduced their own gender-neutral voice named "Quinn" for the American English version of Siri (iOS 15.4, voice 5) [7]. However, is "Quinn" truly perceived as gender-neutral by users?

Listeners may in fact perceive identical speech stimuli differently based on the provided context and their corresponding experiences and beliefs. This has been demonstrated for the perception of *what* is said (i.e., the phonetic detail of specific speech sounds), which is, for example, influenced by the supposed regional origin of a speaker [8, 9, 10] and the supposed speaker gender [11]. Further, the influence of context has also been demonstrated for the perception of *who* is speaking. For example, the fact that a topic is associated with a particular gender can lead to the voice talking about it being more likely associated with that gender [12].

Given the important role that VAs play in many people's lives and the concerns about societal implications of perceived VA gender, it is important to gain a better understanding of how users perceive VA voices and what influences their perceptions. The present work hence explores how visual context influences the perception of gender for a presumed gender-neutral VA voice and how likability ratings are impacted by this.

## 2 Background

### 2.1 Gender-neutral voices

Listeners routinely extract socially significant non-linguistic cues like age and gender from voices—even when speakers use an unfamiliar language or lack embodiment (e.g., a foreign radio broadcast) [13]. Yet, what qualifies as a gender-neutral voice? Gender perception from voice is mainly associated with a shift in the voice's fundamental frequency jointly with the formant frequencies (i.e., resonance frequencies of the vocal tract) [14]. The fundamental frequency (f0) refers to the perceived pitch of the voice, with mean f0 being lower for men (100-120 Hz) and higher for women (200-220 Hz) [14, 15]. Male voices generally have an f0 range of 85 to 180 Hz, while female voices typically span a range of 140 to 255 Hz [16]. A mean f0 in the range of 145 to 165 Hz is considered as potentially gender-neutral [17, 14]. While various other factors, such as speech style, social context, and age can influence the gender attribution of a voice, pitch stands out as the most significant determinant [18, 12, 19]. The gender-neutral voice "Q", for example, was "recorded by people who neither identify as male nor female" [6] and has a mean f0 between 154 and 157 Hz. Apple only stated that their gender-neutral voice "Quinn" was recorded by an individual from the LGBTQ+ community [20]. Measuring f0 in the stimulus material of the present study showed that "Quinn" has a mean f0 of about 152 Hz with a range of about 93 to 206 Hz. It hence seems well positioned at the intersection between typical male and female f0 ranges.

### 2.2 Perception of gender in VAs

VAs operate similarly to disembodied human voices (e.g., on the radio), since they are lacking a tangible physical presence or real-world identity. Individuals may use the VA's voice as a source of information to imagine an identity for it. However, will users perceive the voice of "Quinn" as embodying the same degree of gender neutrality or will some categorize it as distinctly male or female? Prior work by Mooshammer and Etzrodt [12] tested the perception of German male, female, and gender-neutral synthetic voices that were generated using Google WaveNet [21] and Praat [22]. They found that German listeners perceived the gender-neutral voice as neutral overall, but showing a very wide range of ratings compared to the gendered voices. Only 20 % of the participant chose the neutral midpoint of the rating scale, while 46 % tended towards the male side and 34 % towards the female side of the scale. As previously indicated, the perception of a voice can vary based on personal experiences and beliefs. Gender-neutral voices may offer more room for such individual influence and hence show more variability in gender ratings.

According to Nass and Gong [23], users tend to apply gender stereotypes to machines. They found that evaluations delivered by male voices were deemed more credible than those by female voices. Female voices are often perceived as assisting, whereas male voices are seen as authoritative providing direct solutions [23].

### 2.3 Influence of visual context on gender perception

As mentioned above, contextual information can influence perception of *what* is said and *who* is speaking [24, 12]. Crucially, such contextual information may be provided in subtle ways that are unrelated to the concrete interaction, such as by placing culturally significant objects in the listener's field of vision [25, 26, 27]. Work by Sarigul et al. [28] showed that participants rate a voice accompanied by a visual stimulus resembling a robot as having more robotic quality than the same voice accompanied by a visual stimulus resembling a human. Such findings indicate a strong tendency among listeners to use contextual cues to anticipate the quality of voices.

This is likely also the case for the perception of gender. Our first research question (RQ1) is therefore: ***Is the perception of VA gender influenced by visual context unrelated to the VA?***

### 2.4 Influence of gender perception on likability

Mitchell et al. [29] investigated preferences for gender in synthesized voices. Their study indicated that synthesized female voices were perceived as conveying more warmth than male voices. However, Eyssel et al. [30] found that people perceived a robot more positively when the voice aligned with their own gender. These findings suggest that the perceived VA gender may influence the VA's perceived likability. Our second research question (RQ2) is therefore: ***Is the perceived VA likability influenced by the perceived VA gender?***

## 3 Material and Methods

The present study received ethical approval from the *Research Ethics Committee* at the *School of Information and Communication Studies* at *University College Dublin*. Participants were provided with an information sheet and invited to contact the researchers, if they had any questions, before completing a consent form, if they wished to participate. The study was conducted on *SurveyMonkey* (`https://www.surveymonkey.com/`).

### 3.1 Participants

The study included 73 participants (35 male, 38 female) with a mean age of 29.3 years (range: 23 to 44 years) who were recruited by word of mouth. Participants were native speakers of Chinese and proficient in English (A1/A2: 4 %, B1/B2: 36 %, C1/C2: 60 %). More than half of the participants (51 %) indicated to use VAs at least several times a month, while 41 % use them rarely, and 8 % never. Apple's Siri was the most used VA among the participants (64 %). When asked whether VAs should have a gender, only 19 % of participants answered no, while 43 % replied maybe, and 38 % said yes. The latter include 18 % with no gender preference, 19 % who would prefer VAs to be female, and only 1 % who would prefer them to be male.

### 3.2 Stimuli

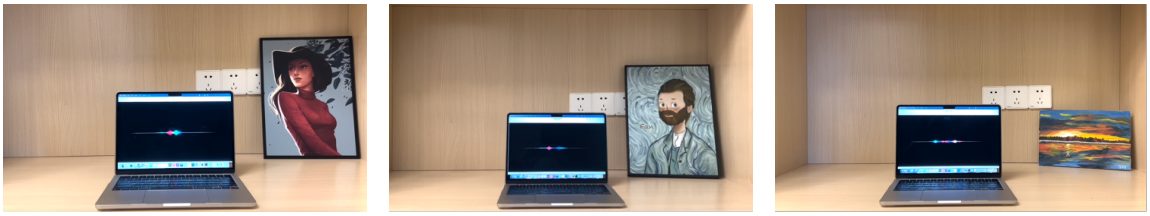
#### 3.2.1 Visual context conditions

The stimulus material in the present study consisted of short videos showing a laptop computer on a desk in three visual context conditions: neutral, female, and male. For the neutral condition, a landscape painting was put at the right back corner of the desk (see Figure 1c). For the female and male conditions, a female or male portrait painting was put at the same place as the landscape painting in the neutral condition, respectively (see Figures 1a and 1b).[1] Participants were randomly assigned to one of the three conditions.

#### 3.2.2 VA voice

The voice used in the present study is the gender-neutral Apple Siri voice 5, also called "Quinn," of macOS version 13.4. As shown in Figure 1, the classic Siri voice visualization was shown on the laptop screen to suggest to the participants that the VA voice came from the computer itself. The speech content was designed by the first author of the present study. It encompasses

[1]The artwork in the neutral condition is by the first author. The female portrait is by Janice Sung (`https://www.janicesung.com/`) and the male portrait is by Fanfan (`https://weibo.com/u/7088277897`).



**(a)** Female condition **(b)** Male condition **(c)** Neutral condition

**Figure 1** – Scenes depicting different conditions with corresponding paintings in the background.

the persona of the VA, which has been active as an assistant for a span of three years. It outlines the spectrum of tasks this VA is proficient in, ranging from resolving technical complexities like programming to handling household chores, as well as engaging in conversations to uplift people's spirits. Given that participants may associate technical issue-solving with male-oriented roles, while communication and domestic tasks are often attributed to female-oriented roles [31], this amalgamation of content seeks to mitigate the potential gender bias triggered by the content itself [12]:

*Hello, I'm a virtual assistant. I've been working as a virtual assistant for over 3 years now, and I love it! I work with people who are looking to get their business off the ground, or who just need some extra help around the house. My clients are always happy with my work. In addition to being a virtual assistant, I'm also a reader and writer. I enjoy helping others in any way that I can, whether it's by offering advice or helping them write their copy for landing pages or emails. I've worked on projects ranging from simple text posts to complex programming tasks, so whatever your project requires, you'll find me very capable of handling it. Please don't hesitate to let me know if you need any help from me, or you just need a conversation about your day or your feelings. I'm very excited to meet you and know more about you!*

#### 3.2.3 Video stimuli

The videos for each condition are about 50 seconds long. They are accessible through the following link: `https://osf.io/9mteh/`.

### 3.3 Perception task

Participants completed the study remotely from their homes and were asked to use laptop or desktop computers with large screens for clear image display. They were required to watch the video in a quiet environment with a clean desk and background wall, using a device capable of playing sound. As a sanity check, participants had to confirm compliance with these conditions by ticking a checkbox to progress with the study.

After watching one of the videos described above (depending on the assigned condition), participants were asked to first rate the perceived gender of the VA on a scale from 0 (male or female) to 9 (female or male), and then they were asked to rate the perceived likability of the VA on a scale from 0 (very unlikable) to 9 (very likable). Participants could not revisit the video during the rating and the two rating scales were presented one at a time to prevent bias of the likability rating on the gender rating. To avoid positional bias in the gender rating, half of the participants were presented with a version of the scale showing the "female" label on the left and the "male" label on the right, while the other half was presented with the reversed order. Demographic questions were asked after the perception task.

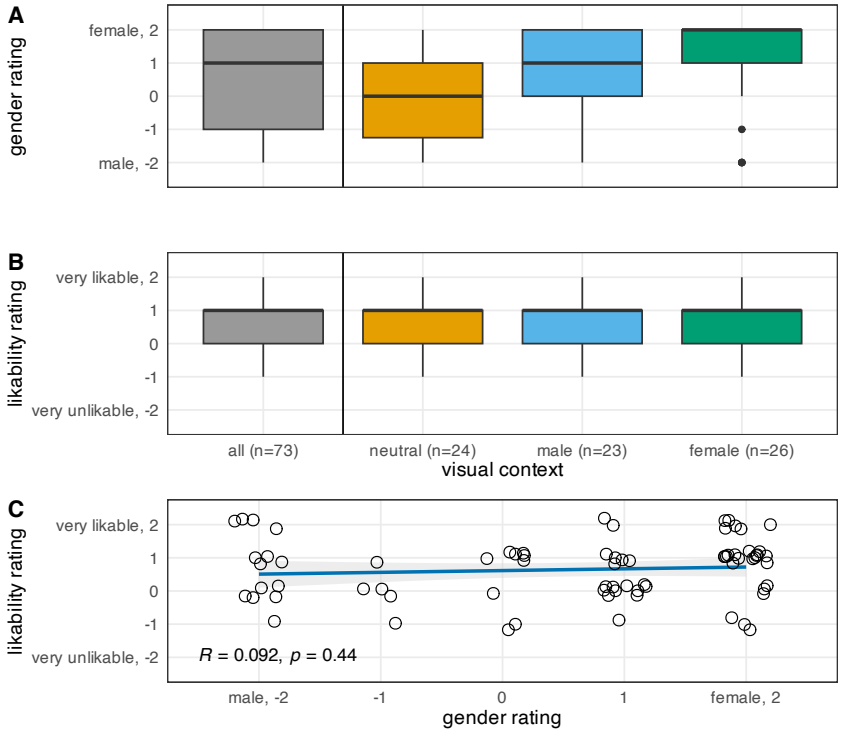



**Figure 2** – **(A)** Gender rating from *male* to *female* and **(B)** likability rating from *very unlikable* to *very likable* of VA voice in different visual contexts. **(C)** Correlation of gender and likability ratings. Gray area indicates 95% confidence interval; vertical and horizontal jitter of 0.2 were applied to data for better visibility in the plot; Pearson correlation coefficient R is indicated with corresponding p-value.

# 4 Analysis and Results

The gender ratings were reverse-scored, where necessary, and then re-scaled to a coarser 1 to 5 scale by grouping adjacent values together (0/1, 2/3, 4/5, 6/7, 8/9). In this way, we created a neutral midpoint out of ratings that indicated only a slight tendency towards male or female gender. Ratings were then centered around zero, so that a rating of $-2$ indicates perceiving "Quinn" as male and a rating of 2 indicates perceiving it as female, while a rating of 0 is considered a gender-neutral perception point (see Figure 2a).

The overall mean rating of 0.53 ($SD = 1.52$) indicates a tendency for the voice to be perceived as rather female. In the neutral condition, "Quinn" received a neutral mean rating of 0.00, with the high standard deviation of 1.47 revealing a wide range of perceptions. Notably, only 17 % of the participants in this condition perceived "Quinn" as neutral, while the majority tended to attribute a gender (male: 38 %, female: 46 %). In the female condition, the same voice was rated as more female with a mean rating of 1.04 and a standard deviation of 1.51. However, in the male condition, "Quinn" was also perceived as tendentially female with a lower mean rating of 0.52 and a lower standard deviation of 1.44 compared to the female condition.

Given that the gender ratings can be treated as continuous and observations are independent, we employed independent t-tests for two samples to assess the female and male conditions in comparison to the neutral condition. This revealed a significant difference in the gender perception scores between the female and neutral condition ($t = 2.46$, df $= 47.8, p = 0.018$). However, the difference between the male and neutral condition was not statistically significant ($t = 1.23$, df $= 45.0, p = 0.226$).

The likability ratings were re-scaled and centered around zero as well, so that a rating of $-2$ indicates perceiving "Quinn" as very unlikable and a rating of 2 indicates perceiving it as very likable, while a rating of 0 is considered a gender-neutral perception point (see Figure 2b).

The overall mean rating of 0.64 ($SD = 0.89$) indicates that the voice was perceived as rather likable. In the neutral condition, likability ratings ($M = 0.58, SD = 0.97$) were lower than those in the female condition ($M = 0.65, SD = 0.85$) and the male condition ($M = 0.70, SD = 0.88$). However, independent t-tests for two samples indicated that likability ratings in the neutral condition did not differ significantly from those in the female ($t = 0.27$, df $= 45.8, p = 0.787$) and the male condition ($t = 0.42$, df $= 44.8, p = 0.679$).

Figure 2c illustrates the relationship between gender and likability ratings. A Pearson correlation coefficient of 0.092 (df $= 71, p = 0.441$) indicates that there is no evidence of a relationship between likability and gender perceptions of this VA voice, which means that "Quinn" was not rated more likable when it was perceived more female.

## 5 Discussion

The present study suggests that Siri's American English voice "Quinn" is not universally perceived as gender-neutral. This was demonstrated with a sample of 73 native Chinese speakers who are proficient in English. They perceived "Quinn" as female overall. While we have verified that our stimulus material of "Quinn" exhibits an f0 mean and range that corresponds to the intersection between typical male and female voices, there may be other acoustic aspects of the voice that skewed the overall perception towards the female pole. However, another explanation is that the voice was presented as belonging to a VA which in itself provided context for the perception. Listeners have been mainly exposed to VAs with default female voices in the past which may skew their perception towards this expectation for gender-neutral voices as well. Similar to Mooshammer and Etzrodt [12], gender ratings had a wide range in the neutral condition of the present study, with some listeners rating "Quinn" as more male, others as more female, and only a minority giving a neutral rating. This illustrates the individual differences in gender perception that may be rooted in each listener's unique experiences and beliefs.

We investigated whether a gendered visual context influences the perception of VA gender (RQ1). Even though the information was presented in the background, unrelated to the VA itself, we found that a female visual context led to more female ratings. However, a male visual context did not lead to more male ratings. This may be due to the specific paintings we selected, which can be investigated in future work. A positive aspect is that this difference indicates that participants did not consciously choose whichever gender was indicated by the visual context. We can only speculate that the female interpretation of the voice might be more plausible (i.e., congruent with user expectation), given that VAs had default female voices for a long time. It has been shown that audio-visual integration is stronger when information provided by both modalities is congruent [32, 33, 34]. Hence, the female contextual information might be integrated into perception more easily than the male contextual information.

Since the perceived topic gender has been shown to inform gender perception [12], we designed the content "Quinn" uttered during the experiment including topics that relate to stereotypically male- and female-oriented roles. However, it is possible that participants perceived this content differently, again shaped by their individual experiences and beliefs, which may have influenced their ratings. We believe that it is difficult to have study participants also evaluate the content in terms of perceived gender, as they may be influenced by the voice they heard. Future work could pretest the content in text form with a separate group of participants. However, this will not provide insight into the individual interpretation of each listener.

We investigated how native Chinese speakers perceived an American English VA. However, gender perception from voice is likely to differ across languages and cultures [35, 36]. Hence, different listener groups need to be tested to acquire a more comprehensive view with regard to the perceived gender of "Quinn"—and other gender-neutral VA voices, as they be-

come available. The cross-cultural aspect is also relevant for the selection of visual context. Future work could improve the study design by selecting male and female portraits that are specifically tailored to the listener group. Illustrations that match the listeners' culture may lead to a increased effects of visual context [37, 38].

We also explored a potential association between the perception of gender and likability ratings (RQ2). Our results show that visual context did not influence the perceived likability of "Quinn," and the voice was not rated more likable when it was perceived more female.

Achieving gender neutrality in VA voices is challenging. This may in part be due to bias in design, since design is influenced by societal power structures and the designers' views on gender [39, 40, 41]. Further, a gender-neutral voice may be impractical if other design elements remain gendered [42]. Crucially, what constitutes a gender-neutral voice needs to be based on user perceptions.

## 6 Conclusion

This study investigated perceptions of Apple Siri's gender-neutral American English voice "Quinn." The native Chinese speakers participating in the study perceived the voice as rather female overall. Visual context in the form of a female portrait influenced gender ratings in the female direction compared to a neutral condition showing a landscape painting. The presence of a male portrait had no equivalent effect. The perceived gender was not correlated with likability ratings. The results show that it is challenging to achieve gender neutrality for voice assistants (VAs), as perception depends not only on the voice itself, but also on the interpretation of the context—even if seemingly unrelated to the VA. We recommend taking the influence of visual factors on gender perception into account when dealing with the topic of gender in VAs.

## Acknowledgments

This work was funded by Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. We thank the artists Janice Sung (`https://www.janicesung.com/`) and Fanfan (`https://weibo.com/u/7088277897/`), who provided the female and male portraits respectively.

## References

[1] PORCHERON, M., J. FISCHER, S. REEVES ET AL.: *Voice interfaces in everyday life*. In *Conference on Human Factors in Computing Systems*, p. 1–12. 2018.

[2] BERGEN, H.: *'I'd blush if I could': Digital assistants, disembodied cyborgs and the problem of gender*. *Word and Text, a Journal of Literary Studies and Linguistics*, 6(1), pp. 95–113, 2016.

[3] NI LOIDEAIN, N. and R. ADAMS: *Female servitude by default and social harm: AI virtual personal assistants, the FTC, and unfair commercial practices*. 2019.

[4] WEST, M., R. KRAUT, and H. CHEW: *I'd blush if I could: closing gender divides in digital skills through education*. Programme and Meeting Document, 2019.

[5] NI LOIDEAIN, N. and R. ADAMS: *From Alexa to Siri and the GDPR: The gendering of virtual personal assistants and the role of data protection impact assessments*. *Computer Law & Security Review*, 36, p. 105366, 2020.

[6] GENDERLESS VOICE: 2019. `https://genderlessvoice.com/` [Accessed: Mar 25, 2024].

[7] PEREZ, S.: *Siri gains a new gender-neutral voice option in latest ios update*. 2022. https://techcrunch.com/2022/02/24/siri-gains-a-new-gender-neutral-voice-option-in-latest-ios-update [Accessed: Feb 25, 2024].

[8] NIEDZIELSKI, N.: *The effect of social information on the perception of sociolinguistic variables*. *Journal of Language and Social Psychology*, 18(1), pp. 62–85, 1999.

[9] HAY, J., A. NOLAN, and K. DRAGER: *From fush to feesh: Exemplar priming in speech perception*. *The Linguistic Review*, 23(3), pp. 351–379, 2006.

[10] JANNEDY, S. and M. WEIRICH: *Sound change in an urban setting: Category instability of the palatal fricative in Berlin*. *Laboratory Phonology*, 5(1), pp. 91–122, 2014.

[11] STRAND, E. and K. JOHNSON: *Gradient and visual speaker normalization in the perception of fricatives*. In D. GIBBON (ed.), *Natural Language Processing and Speech Technology*, pp. 14–26. De Gruyter Mouton, Berlin, Boston, 1996.

[12] MOOSHAMMER, S. and K. ETZRODT: *Gender ambiguity in voice-based assistants: Gender perception and influences of context*. *Human-Machine Communication*, 5, pp. 49–74, 2022.

[13] LATINUS, M. and P. BELIN: *Human voice perception*. *Current Biology*, 21(4), pp. R143–R145, 2011.

[14] GELFER, M. and Q. BENNETT: *Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender*. *Journal of Voice*, 27(5), pp. 556–566, 2013.

[15] SIMPSON, A.: *Phonetic differences between male and female speech*. *Language and Linguistics Compass*, 3(2), pp. 621–640, 2009.

[16] DANIELESCU, A.: *Eschewing gender stereotypes in voice assistants to promote inclusion*. In *Conference on Conversational User Interfaces*, pp. 1–3. 2020.

[17] GALLENA, S., B. STICKELS, and E. STICKELS: *Gender perception after raising vowel fundamental and formant frequencies: Considerations for oral resonance research*. *Journal of Voice*, 32(5), pp. 592–601, 2018.

[18] JØRGENSEN, A.: *Speaking (of) gender: A sociolinguistic exploration of voice and transgender identity*. Københavns Universitet, 2016.

[19] JUNGER, J., K. PAULY, S. BRÖHR ET AL.: *Sex matters: Neural correlates of voice gender perception*. *Neuroimage*, 79, pp. 275–287, 2013.

[20] FRIED, I.: *Apple gives Siri a less gendered voice*. 2022. https://www.axios.com/2022/02/23/apple-gives-siri-less-gendered-voice [Accessed: Feb 13, 2024].

[21] VAN DEN OORD, A., S. DIELEMAN, H. ZEN ET AL.: *WaveNet: A generative model for raw audio*. 2016. ArXiv:1609.03499.

[22] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer*, 2020. http://www.praat.org/.

[23] NASS, C. and L. GONG: *Speech interfaces from an evolutionary perspective*. *Communications of the ACM*, 43(9), pp. 36–43, 2000.

[24] STRAND, E.: *Uncovering the role of gender stereotypes in speech perception*. *Journal of Language and Social Psychology*, 18(1), pp. 86–100, 1999.

[25] HAY, J. and K. DRAGER: *Stuffed toys and speech perception*. *Linguistics*, 48(4), pp. 865–892, 2010.

[26] WALKER, M., A. SZAKAY, and F. COX: *Can kiwis and koalas as cultural primes induce*

*perceptual bias in Australian English speaking listeners? Laboratory Phonology*, 10(1), pp. 1–29, 2019.

[27] HURRING, G., J. HAY, K. DRAGER ET AL.: *Social priming in speech perception: Revisiting kangaroo/kiwi priming in New Zealand English. Brain Sciences*, 12(6), p. 684, 2022.

[28] SARIGUL, B., I. SALTIK, B. HOKELEK ET AL.: *Does the appearance of an agent affect how we perceive his/her voice? Audio-visual predictive processes in human-robot interaction.* In *International Conference on Human-Robot Interaction*, pp. 430–432. 2020.

[29] MITCHELL, W., C.-C. HO, H. PATEL ET AL.: *Does social desirability bias favor humans? Explicit-implicit evaluations of synthesized speech support a new HCI model of impression management. Computers in Human Behavior*, 27(1), pp. 402–412, 2011.

[30] EYSSEL, F., D. KUCHENBRANDT, S. BOBINGER ET AL.: *'If you sound like me, you must be more human' on the interplay of robot and user features on human-robot acceptance and anthropomorphism.* In *International Conference on Human-Robot Interaction*, pp. 125–126. 2012.

[31] WHITE, M. and G. WHITE: *Implicit and explicit occupational gender stereotypes. Sex Roles*, 55, pp. 259–266, 2006.

[32] CALVERT, G., P. HANSEN, S. IVERSEN ET AL.: *Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the bold effect. Neuroimage*, 14(2), pp. 427–438, 2001.

[33] BEAUCHAMP, M., K. LEE, B. ARGALL ET AL.: *Integration of auditory and visual information about objects in superior temporal sulcus. Neuron*, 41(5), pp. 809–823, 2004.

[34] SONG, J.-J., H.-J. LEE, H. KANG ET AL.: *Effects of congruent and incongruent visual cues on speech perception and brain activity in cochlear implant users. Brain Structure and Function*, 220, pp. 1109–1125, 2015.

[35] VALENTINE, C. and B. SAINT DAMIAN: *Gender and culture as determinants of the 'ideal voice'. Semiotica*, 71(3-4), pp. 285–304, 1988.

[36] SULPIZIO, S., F. FASOLI, A. MAASS ET AL.: *The sound of voice: Voice-based categorization of speakers' sexual orientation within and across languages. PloS one*, 10(7), p. e0128882, 2015.

[37] SEGALL, M., D. CAMPBELL, and M. HERSKOVITS: *The influence of culture on visual perception*, vol. xvii, 268. Bobbs-Merrill, Oxford, England, 1966.

[38] PHILLIPS, W.: *Cross-cultural differences in visual perception of color, illusions, depth, and pictures.* In *Cross-cultural psychology: Contemporary themes and perspectives*, chap. 13, pp. 287–308. Wiley Blackwell, Hoboken, NJ, US, 2019.

[39] AKRICH, M.: *The de-scription of technical objects.* In *Shaping Technology/Building Society: Studies in Sociotechnical Change*, pp. 205–224. MIT Press, 1992.

[40] OUDSHOORN, N., E. ROMMES, and M. STIENSTRA: *Configuring the user as everybody: Gender and design cultures in information and communication technologies. Science, Technology, & Human Values*, 29(1), pp. 30–63, 2004.

[41] SØRAA, R.: *Mechanical genders: How do humans gender robots? Gender, Technology and Development*, 21(1-2), pp. 99–115, 2017.

[42] SUTTON, S.: *Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity.* In *Conference on Conversational User Interfaces*, pp. 1–8. 2020.

# WERBESTIMMEN: ZUSAMMENHANG ZWISCHEN INHALTLICHEN ASPEKTEN VON RADIOWERBUNG UND PARAMETERN DER GRUNDFREQUENZ

*Inke Henneberger[1], Melanie Weirich[1], Adrian P. Simpson[1]*

*[1]Friedrich-Schiller-Universität Jena*
*inke.henneberger@uni-jena.de*

**Kurzfassung:** Die Studie untersucht den Einfluss inhaltlicher Aspekte von Werbung auf Parameter der Grundfrequenz anhand aktueller Radiowerbung. Analysiert wurden 292 Werbestimmen (170 männlich kategorisiert, 122 weiblich kategorisiert) aus 220 Werbespots, die während der Morningshow-Programme von vier deutschen Radiosendern ausgestrahlt wurden. Als inhaltliche Aspekte wurden der Preis des beworbenen Produkts (bis 10€/ 10€ bis 100 €/ 100€ bis 1000 €/ über 1000€) und die Form der Anrede, die für die Hörer*innen verwendet wurde (formelle Anrede „Sie"/informelle Anrede „Du"), einbezogen. Um Unterschiede der Grundfrequenz ($f_0$) und Grundfrequenzvariation ($f_0$-Spannweite) der Werbestimmen in verschiedenen inhaltlichen Kontexten zu ermitteln, wurden unter Einbezug des kategorisierten Geschlechts der Stimmen lineare gemischte Modelle (LMMs) durchgeführt.

In den Rohdaten lässt sich beobachten, dass die weiblich kategorisierten Stimmen stärker in der Grundfrequenz zwischen den unterschiedlichen Preiskategorien variieren als die männlich kategorisierten Stimmen. Die LMMs bestätigen den Einfluss von Preiskategorie und Form der Anrede auf die Grundfrequenzparameter. Interaktionen zwischen der Preiskategorie und dem kategorisierten Geschlecht sowie zwischen der Form der Anrede und dem kategorisierten Geschlecht zeigten jedoch keine Signifikanz.

## 1 Einleitung

Der Einsatz von Stimmen ist ein wesentliches Gestaltungsmittel in der Werbung. Besonders in der Radiowerbung, bei der der gesamte Wirkungsmechanismus auf die auditive Ebene entfällt, ist die Auswahl der Werbestimmen eng mit dem Erfolg des Produkts verknüpft, den sich Werbetreibende versprechen. Busse et al. [1] stellten in deutscher Fernsehwerbung fest, dass eindeutig männliche und weibliche Produkte meist von Stimmen des gleichen Geschlechts präsentiert wurden. Bei genderneutralen Produkten überwog jedoch der Anteil männlich kategorisierter Stimmen. Eine ähnliche Tendenz zeigte sich bei Henneberger [2]. Eine Untersuchung US-amerikanischer Sender aus dem Jahr 2020 [3] zeigte eine nahezu ausgeglichene Repräsentation von weiblich und männlich kategorisierten Voiceovern, stellte jedoch fest, dass weiblich kategorisierte Stimmen in Haushaltskontexten überrepräsentiert und in Nicht-Haushaltskontexten unterrepräsentiert waren.

Werbestimmen unterstützen nicht nur die Werbebotschaft, sondern beeinflussen auch die Darstellung von Personengruppen. In empirischen Studien wurde vielfach festgestellt, dass die Ausprägungen der Grundfrequenz ($f_0$) und der Grundfrequenzvariation geschlechtsspezifisch zu unterschiedlichen Persönlichkeitswahrnehmungen führen. Bei weiblich kategorisierten Stimmen geht eine höhere $f_0$ beispielsweise mit höheren Einschätzungen von Attraktivität [4] und Sympathie [5] einher. Weiblich kategorisierte Stimmen mit tieferer $f_0$ führen zu höheren Einschätzungen von Kompetenz, Stärke und Glaubwürdigkeit [6]. Eine höhere Variation der $f_0$ wird vor allem weiblich kategorisierten Stimmen zugeschrieben, die als emotional und empathisch bewertet werden [7], während niedrigere Variation mit einer geringeren Einschätzung von Attraktivität verbunden ist [8;9].

Bezüglich männlich kategorisierter Stimmen wird eine niedrigere $f_0$ als attraktiver eingeschätzt als eine höhere [8;9]. Außerdem werden tiefere männlich kategorisierte Stimmen

ebenfalls als glaubwürdiger und kompetenter wahrgenommen [6]. Überdurchschnittlich hohe und niedrige $f_0$-Variationen führten in Untersuchungen von Brown et al. [10] zu geringeren Einschätzungen von Kompetenz und Wohlwollen.

Diese Studie untersucht den Einfluss von inhaltlichen Merkmalen der Radiowerbung auf diese Grundfrequenzmerkmale. Unsere Forschungsfragen lauten wie folgt:

1. Beeinflussen Inhaltsmerkmale (Preis des Produkts, Form der Anrede („Du“ oder „Sie“)) die Parameter der Grundfrequenz (mittlere $f_0$, $f_0$-Spannweite) von Werbestimmen?
2. Unterscheiden sich diese Effekte zwischen weiblich und männlich kategorisierten Stimmen?

# 2 Methode

## 2.1 Korpus

Der Untersuchung liegt ein Korpus zugrunde, das Radiowerbespots enthält, die bei vier verschiedenen Radiostationen ausgespielt wurden: Berliner Rundfunk, Kiss FM, Antenne Thüringen und Radio Top 40. Zwei dieser Radiostationen haben ihr Sendegebiet im urbanen Berliner Raum, während die anderen beiden in Thüringen und damit in einer ländlich geprägten Region angesiedelt sind. Auch das Publikum unterscheidet sich altersbedingt: Berliner Rundfunk und Antenne Thüringen richten sich an ein älteres Publikum (Durchschnittsalter: Berliner Rundfunk - 52,7 Jahre, Antenne Thüringen - 53,1 Jahre) [11;12], während Kiss FM und Radio Top 40 ein jüngeres Publikum ansprechen (Durchschnittsalter: Kiss FM - 38,8 Jahre, Radio Top 40 - 35,6 Jahre) [11;13].

Die Werbespots wurden zu drei Messzeitpunkten mit ca. einem Monat Abstand im November 2023, Dezember 2023 und Januar 2024 während der Morning-Show der Radiosender von 5:00 bis 10:00 Uhr aufgezeichnet, da die Hörer*innenschaft zu dieser Sendezeit am größten ist. Das Radioprogramm wurde mithilfe einer Webradio-Anwendung [14] auf einem mobilen Endgerät abgespielt, während das interne Audiosignal zeitgleich durch eine Soundrecorder-Anwendung [15] auf dem Gerät aufgezeichnet wurde. Die Werbeblöcke wurden anschließend mit *Audacity* [16] mit einer Amplitudenauflösung von 32 Bit und einer Abtastrate von 32 kHz extrahiert.

Während jeder Morning-Show-Ausstrahlung gab es sieben bis neun Werbeunterbrechungen mit drei bis zehn Werbespots. Einige Spots wurden während der gesamten Erhebungszeit nur einmal bei einem Sender ausgespielt, andere wurden mehrfach über die Sender hinweg gesendet. Für die vorliegende Untersuchung wurden die Vorkommenshäufigkeit der einzelnen Stimmen im Korpus nicht berücksichtigt. Ohne Einbezug der Wiederholungen ergab sich ein Korpus mit 220 verschiedenen Werbespots und 292 verschiedenen Werbestimmen. Die Stimmen wurden auditiv als männlich oder weiblich kategorisiert, da keine Informationen über das tatsächliche Geschlecht der Sprecher*innen vorliegen. 170 Stimmen wurden als männlich und 122 Stimmen als weiblich kategorisiert. Das Verhältnis von 58 % männlich kategorisierten zu 42 % weiblich kategorisierten Stimmen zeigt, dass männlich kategorisierte Stimmen im Korpus weiterhin stärker vertreten sind, auch wenn das Ungleichgewicht im Vergleich zu früheren Studien [17;18] geringer ist.

## 2.2 Inhaltliche Aspekte

Zur Untersuchung der Unterschiede der Grundfrequenzparameter anhand inhaltlicher Aspekte der Werbung wurde zum einen der Preis des beworbenen Produkts erfasst. Die Produkte wurden jeweils in Preisspannen von „bis 10€“, „10€ bis 100€“, „100€ bis 1000€“ und „über 1000€“ eingeordnet.

Des Weiteren wurde die Form der Anrede untersucht, die die Werbesprecher*innen zur Ansprache der Hörer*innen nutzten. Hierbei wurde zwischen der formellen Anrede „Sie“ und der informellen Anrede „Du“ unterschieden.

Ein Blick auf die Fallzahlen von weiblich und männlich kategorisierten Stimmen innerhalb der inhaltlichen Merkmalsausprägungen zeigt, dass männlich kategorisierte Stimmen in jedem Kontext vorherrschend sind (siehe Tabelle 1). Besonders groß ist die Differenz zwischen weiblich und männlich kategorisierten Stimmen in der teuersten Preiskategorie sowie bei der Verwendung der formellen Anrede „Sie“.

**Tabelle 1** – Vorkommenshäufigkeit in absoluten Zahlen von weiblich und männlich kategorisierten Stimmen unter inhaltlichen Aspekten im Gesamtkorpus der analysierten Radiowerbung

| | | **Weiblich kategorisiert** | **Männlich kategorisiert** |
|---|---|---|---|
| **Preis des beworbenen Produkts** | Bis 10 € | 15 | 24 |
| | 10€ - 100 € | 35 | 44 |
| | 100€ - 1000€ | 17 | 27 |
| | Über 1000€ | 13 | 31 |
| | Kein Preis definiert | 42 | 44 |
| | | | |
| **Form der Anrede** | Du | 44 | 50 |
| | Sie | 15 | 43 |
| | Keine Anrede verwendet | 63 | 77 |

## 2.3 Akustische Analysen

Da dieser Untersuchung reale Werbeinhalte zugrunde liegen, sind die meisten untersuchten Werbestimmen durch atmosphärische Geräusche oder ein Musikbett unterlegt, das Aufmerksamkeit erregen oder eine Anpassung an den medialen Kontext bewirken soll [19]. Diese Hintergrunduntermalung kann zu Verzerrungen der akustischen stimmbezogenen Daten führen. Daher wurden die Werbeaufnahmen mit der KI *LALAL.AI* [20] bearbeitet, die üblicherweise in der Musikindustrie zur Trennung von Gesangs- und Instrumentalspuren verwendet wird. Pretests zeigten, dass die Verzerrung der Grundfrequenzdaten durch die Segregation von Stimme und Geräusch- bzw. Musikkulisse mit *LALAL.AI* verringert werden konnte.

Für die akustische Analyse wurden die Audioaufnahmen segmentiert und mit *WebMAUS* [21] annotiert. Die resultierenden Praat-Textgrid-Dateien wurden manuell überprüft und korrigiert. Mit *Praat* [22;23] wurde die durchschnittliche Grundfrequenz ($f_0$) über eine Äußerung hinweg erfasst.

Zusätzlich wurde die Grundfrequenzvariation betrachtet, gemessen durch die $f_0$-Spannweite in Form eines 90%-Intervalls. Dafür wurde die Spannweite zwischen dem 5%- und dem 95%-Perzentil aller $f_0$-Werte einer Äußerung erfasst.

## 2.4 Datenanalyse

Für die Betrachtung des Einflusses der Preiskategorie wurde ein Subset erstellt, in dem alle Stimmen ausgeschlossen wurden, welche Produkte bewarben, bei denen kein Preis definiert werden konnte. Ein weiteres Subset wurde für die Betrachtung des Einflusses der Anrede gebildet, in dem nur die Stimmen ausgewählt wurden, die eine direkte Ansprache der Hörer*innen mit „Du“ oder „Sie“ nutzen.

Die statistischen Analysen wurden in *R* [24] durchgeführt, wobei die Grafiken mit dem Paket *ggplot2* [25] erstellt wurden. Mithilfe linear gemischter Modelle (LMMs), erstellt mit dem Paket *lme4* [26], testeten wir die Haupteffekte der Anredeform und des Produktpreises auf die Grundfrequenzparameter sowie mögliche Interaktionen mit dem kategorisierten Geschlecht der Werbestimmen. Die Sprecher*innen wurden als zufälliger Effekt einbezogen. Für alle LMMs wurden Likelihood-Ratio-Tests sowie die Anova-Funktion verwendet, um die Modelle zu finalisieren. Post-hoc-Tukey-Tests wurden mit dem Paket *emmeans* [27] durchgeführt.

# 3 Ergebnisse

## 3.1 Mittlere Grundfrequenz ($f_0$)

Ein Blick auf die Rohdaten zeigt den Einfluss des Preises des beworbenen Produkts auf die mittlere $f_0$ der Werbestimmen (siehe Abbildung 1). Besonders bei weiblichen Stimmen lassen sich niedrigere $f_0$-Werte beobachten, wenn das beworbene Produkt teurer ist.

Durch die linearen gemischten Modelle können Haupteffekte vom kategorisierten Geschlecht ($\chi^2(1) = 134.19$, $p < .001$) und der Preiskategorie ($\chi^2(3) = 10.89$, $p < .05$) auf die mittlere $f_0$ nachgewiesen werden. Die Post-hoc-Tukey-Tests zeigen jedoch nur einen signifikanten Unterschied zwischen den Preiskategorien „bis 10€“ und „über 1000€“ (mittlere $f_0$ „bis 10 €“ = 199 Hz, mittlere $f_0$ „über 1000€“ = 172 Hz, $p < .01$) (siehe Abbildung 2). Eine getestete Interaktion zwischen dem kategorisierten Geschlecht und der Preiskategorie ist nicht signifikant ($\chi^2(3) = 2.43$, $p = 0.49$).

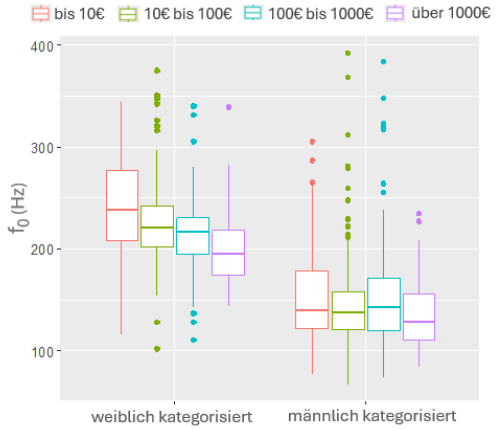



**Abbildung 1** – mittlere $f_0$ gruppiert nach kategorisiertem Geschlecht und Preis des beworbenen Produkts

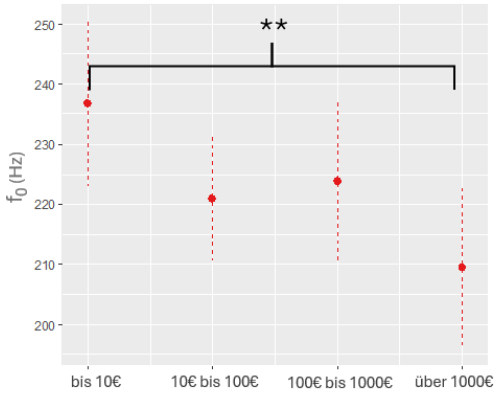



**Abbildung 2** – Haupteffekt der Preiskategorie auf $f_0$, signifikanter Unterschied gekennzeichnet (* = $p < .05$, ** = $p < .01$, *** = $p < .001$)

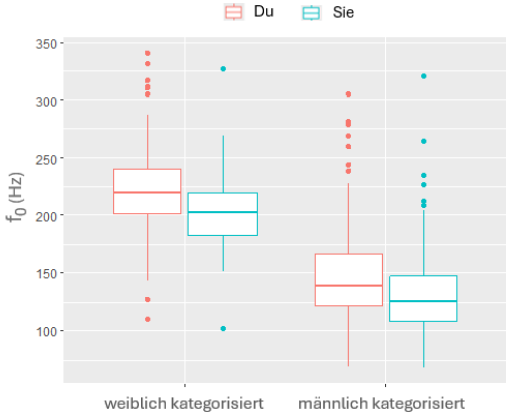



**Abbildung 3** – mittlere $f_0$ gruppiert nach kategorisiertem Geschlecht und Anrede.

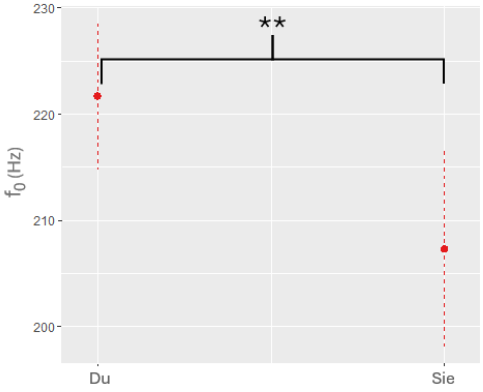



**Abbildung 4** – Haupteffekt der Anrede auf $f_0$, signifikanter Unterschied gekennzeichnet (* = $p < .05$, ** = $p < .01$, *** = $p < .001$)

Bezüglich der Anrede lässt sich beobachten, dass sowohl weiblich als auch männlich kategorisierte Stimmen eine niedrigere $f_0$ aufweisen (siehe Abbildung 3), wenn eine förmliche Anrede mit „Sie“ verwendet wird. Die LMMs bestätigen signifikante Haupteffekte des kategorisierten Geschlechts ($\chi^2$ (1) = 177.54, p < 0.001) und der Form der Anrede ($\chi^2$ (1) = 10.43, p < .01) (siehe Abbildung 4). Auch hier ist die getestete Interaktion zwischen der Anrede und dem kategorisierten Geschlecht der Stimmen nicht signifikant ($\chi^2$ (1) = 0.31, p = 0.58).

### 3.2 Grundfrequenzvariation ($f_0$-Spannweite, 90%-Intervall)

Betrachtet man die $f_0$-Spannweite in Form des 90%-Intervalls, zeichnet sich ein ähnliches Bild wie bei den Daten der mittleren $f_0$ ab: Die $f_0$-Spannweite ist geringer, wenn teurere Produkte durch weiblich kategorisierte Sprecher*innen beworben werden (siehe Abbildung 5). Bei den männlich kategorisierten Stimmen wirken die Unterschiede weniger stark ausgeprägt. Trotzdem lassen sich auch hier durch die LMMs lediglich die Haupteffekte des kategorisierten Geschlechts ($\chi^2(1) = 44.81$, p < .001) und der Preiskategorie ($\chi^2(3) = 9.41$, p < .05) auf die $f_0$-Spannweite nachweisen. Die Post-Hoc-Tukey-Tests bestätigen nur zwischen den Preiskategorien „bis 10€“ und „über 1000€“ einen signifikanten Unterschied ($f_0$-Spannweite „bis 10€“ = 178 Hz, $f_0$-Spannweite „über 1000€“ = 150 Hz, p < .05) (siehe Abbildung 6). Eine Interaktion zwischen der Preiskategorie und dem kategorisierten Geschlecht der Werbestimmen verbessert das Modell nicht signifikant ($\chi^2(3) = 4.08$, p = 0.25).

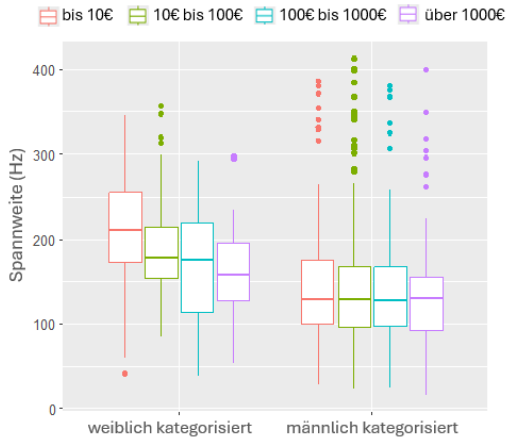



**Abbildung 5** – $f_0$-Spannweite gruppiert nach kategorisiertem Geschlecht und Preiskategorie

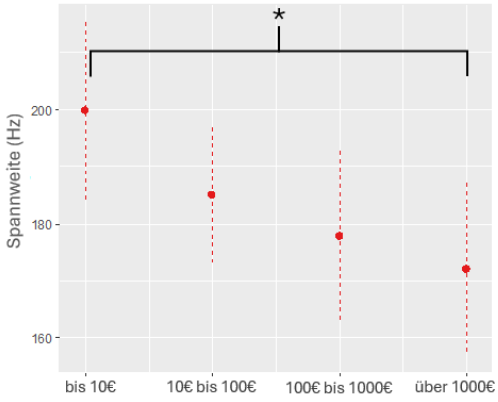



**Abbildung 6** – Haupteffekt von Preiskategorie auf $f_0$-Spannweite, signifikanter Unterschied gekennzeichnet (* = p < .05, ** = p < .01, *** = p < .001)

Bezüglich der Anrede lassen sich nur geringfügige Unterschiede in der $f_0$-Spannweite der Werbestimmen betrachten (siehe Abbildung 7). Die LMMs bestätigen einen signifikanten Haupteffekt des kategorisierten Geschlechts ($\chi^2(1) = 33.97$, p < 0.001). Der zusätzlich getestete Haupteffekt der Anrede auf die $f_0$-Spannweite hingegen erreicht keine Signifikanz ($\chi^2(1) = 3.63$, p = 0.057).

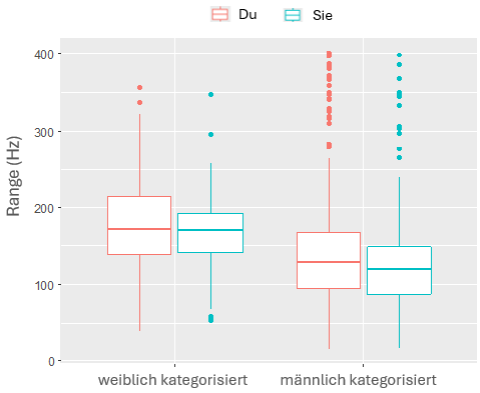



**Abbildung 7** – $f_0$-Spannweite gruppiert nach kategorisiertem Geschlecht und Anrede

# 4 Diskussion

Im Rahmen dieser Studie sollte untersucht werden, ob inhaltliche Merkmale in Radiowerbespots einen Einfluss auf die Grundfrequenzparameter von Werbesprecher*innen haben und ob sich dieser Einfluss zwischen weiblich kategorisierten und männlich kategorisierten Stimmen unterscheidet.

Bezüglich beider Parameter konnten Unterschiede zwischen den inhaltlichen Merkmalsausprägungen festgestellt werden, wie z. B. eine tiefere durchschnittliche Grundfrequenz, wenn Sprecher*innen die formelle Form der Anrede mit „Sie“ nutzten (siehe Abbildung 3). Folglich könnte der Einsatz tieferer Stimmen für die formelle und seriöse Form der Anrede als angemessener eingeschätzt werden.

Nachweislich unterscheidet sich die mittlere Grundfrequenz der Werbestimmen, je nachdem, wie teuer das beworbene Produkt ist, als auch dahingehend, ob die Sprecher*innen eine informelle Anrede mit „Du“ oder eine formelle Anrede mit „Sie“ nutzen. In Werbespots mit formellen Anredeformen und höheren Preiskategorien werden Stimmen mit tieferer $f_0$ genutzt (siehe Abbildung 2 & 4). Auch bezüglich der $f_0$-Spannweite konnte ein signifikanter Einfluss der Preiskategorie nachgewiesen werden, mit einer geringeren $f_0$-Spannweite in höheren Preiskategorien (siehe Abbildung 6).

Von besonderem Interesse sind die Unterschiede zwischen männlich und weiblich kategorisierten Stimmen. Es zeigt sich ein zahlenmäßiges Ungleichgewicht im gesamten Korpus zwischen männlich und weiblich kategorisierten Stimmen, das sich auch in früheren Studien belegen lässt [17;18]. Dieses Ungleichgewicht ist vor allem bei den höherpreisigen Werbeinhalten zu sehen (siehe Tabelle 1).

Unterschiede in den Ausprägungen der Einflüsse von Preiskategorie und Anrede auf die Grundfrequenzparameter zwischen den männlich und weiblich kategorisierten Stimmen lassen sich in den Rohdaten beobachten. Die Beobachtungen deuten darauf hin, dass weiblich kategorisierte Stimmen in der Tendenz größere kontextabhängige Unterschiede aufweisen (siehe Abbildung 1 & 5). Die getesteten Interaktionen zwischen Preiskategorie und kategorisiertem Geschlecht, sowie Form der Anrede und kategorisiertem Geschlecht wurden jedoch nicht signifikant.

Aus Wahrnehmungsstudien zeichnet sich ab, dass hohe, weiblich kategorisierte Stimmen eher als attraktiv und tiefe, weiblich kategorisierte Stimmen eher als kompetent eingeschätzt werden. Tiefe, männlich kategorisierte Stimmen werden hingegen sowohl als attraktiv als auch als kompetent wahrgenommen [4;5;6]. Da der höhere Preis für ein Produkt eine größere Hemmschwelle für den Kauf darstellt, ist davon auszugehen, dass hier Stimmen eingesetzt werden, von denen man sich eine große Überzeugungskraft verspricht. Dies könnte ein Erklärungsansatz sein, warum männlich kategorisierte Stimmen in teuren Preiskategorien stärker vertreten sind und warum sich nominelle Unterschiede zwischen den Preiskategorien stärker bei weiblich kategorisierten Stimmen beobachten lassen, auch wenn diese nicht statistisch signifikant sind.

Martín-Santana et al. [28] wiesen in einem Experiment mit spanischer Radiowerbung jedoch auch nach, dass tiefere, weiblich kategorisierte Stimmen generell eine positivere Einstellung zum Produkt hervorriefen. Dies gilt es auch auf Basis des erstellten Korpus für deutsche Hörer*innenpräferenzen zu überprüfen.

Außerdem könnte die Einschätzung höherer Attraktivität, die mit einer höheren $f_0$ und $f_0$-Variation weiblich kategorisierter Stimmen einhergeht [4;9;10], einen Erklärungsansatz bieten, warum diese beiden stimmlichen Merkmale sich in dieser Untersuchung nominell stärker in günstigeren Produktkontexten auftreten. Es könnte darauf hindeuten, dass die wahrgenommene Attraktivität weiblich kategorisierter Stimmen in dieser Preiskategorie als Verkaufsargument genutzt wird.

Im weiteren Verlauf dieses Forschungsprojekts sollen noch weitere Faktoren der Werbestimmen (Formanten, Vokalraum, Sprechgeschwindigkeit, Stimmqualität) sowie weitere inhaltliche Kontexte (Produktkategorie, Werbestrategie etc.) betrachtet werden, um den Einfluss der Werbeinhalte auf die Auswahl bzw. Beschaffenheit von Werbestimmen umfassender darzustellen. Des Weiteren soll unter Einbezug weiterer phonetischer Merkmale eine Clusteranalyse durchgeführt werden, um etwaige Stimmtypen im Korpus zu ermitteln und zu analysieren, ob diese mit bestimmten inhaltlichen Faktoren der Werbung korrelieren. Ein weiterer Blick gilt außerdem den Unterschieden in der Werbung zwischen den Sendegebieten und den altersspezifischen Zielgruppen der ausgewählten Sender.

## Dank

Das Projekt wird von der Deutschen Forschungsgemeinschaft gefördert (WE 5757/6-1, 529274597, http://www.dfg.de/). Wir möchten unseren Studentischen Hilfskräften Julius Dorn und Lilia Kurnosova für die Überarbeitungen der Annotationen danken.

## Literatur

[1] BUSSE, C., BÖHM, E., SUHM, P., BAUMANN, I., und KIPP, K.: Typisch Werbung! Bedient die Auswahl von Voice-Over-Stimmen geschlechtsspezifische und stereotype Vorstellungen? In: Kranich, Wieland (Hrsg.): Sprechwissenschaft heute. Sprache und Sprechen 52, S. 91-100, Schneider Hohengehren, Baltmannsweiler, 2020.

[2] HENNEBERGER, I. (2021): Stereotype Werbestimmen – Untersuchung von geschlechterspezifischen Stimm- und Sprech-typen in der deutschen Primetime-Werbung anhand phonetischer Parameter. Masterarbeit, Martin-Luther-Universität Halle-Wittenberg, 2021.

[3] LEONARDO, A.; POLLARD, A.; CLARK, R.: The gender of product representatives and voice-overs in television commercials: An update. In: Sociology Between the Gaps: Forgotten and Neglected Topics 5(1), S. 6, 2020.

[4] FEINBERG, D., DE BRUINE, L., JONES, B., und PERRETT, D.: The Role of Femninity and Averageness of Voice Pitch in aesthetic Judgments of Women's Voices. In: Perception 37(4),615–623, 2008.

[5] SCHÜTTE, V.: 17 Hz machen einen Unterschied: Die Stimme der als managementkompetentgeltenden Frau ist in ihrer Wahrnehmung männlich. Eine Kurzdarstellung. In: sprechen 57, 77–85, 2014.

[6] KLOFSTAD, C., ANDERSON, R., und PETERS, S.: Sounds like a Winner: Voice Pitch influences Perception of Leadership Capacity in both Men and Women. In: Proceedings of the Royal Society of London B: Biological Sciences, 2698-2704, 2012.

[7] KIESE-HIMMEL, C.: *Körperinstrument Stimme*, Springer Berlin Heidelberg, 2016.

[8] RIDING, D., LONSDALE, D., und BROWN, B.: The Effects of Average Fundamental Frequency and Variance of Fundamental Frequency on male Vocal Attractiveness to Women. In: Journal of nonverbal Behavior 30(2), 55–61, 2006.

[9] VÖLKERT, S.: Das Phänomen "Sexy Stimme". Was macht Männer- und Frauenstimmen sexy? In: sprechen 54, 74–87, 2012.

[10] BROWN, B., STRONG, W., RENCHER, A.: Fifty-four Voices from Two: The Effects of Simultaneous Manipulations of Rate, Mean Fundamental Frequency, and Variance of Fundamental Frequency on Ratings of Personality from Speech. In: Journal of the Acoustical Society of America 55(2), 313–318, 1974."

[11] AUDIO MEDIA PLUS: *Mediadaten* [Online]. URL https://www.audiomediaplus.de/portfolio/mediadaten.php, Zugriff am 10.10.2024.

[12] TOPRADIO.DE: *Berliner Rundfunk 91.4.* [Online] Verfügbar unter: https://www.topradio.de/sender/berliner-rundfunk-914/ Zugriff am 10.10.2024.

[13] TOPRADIO.DE: *98.8 KISS FM – Der Beat von Berlin.* [Online] Verfügbar unter: https://www.topradio.de/sender/988-kiss-fm/ Zugriff am: 10.10.2024.

[14] APPMIND-RADIO FM, RADIO ONLINE: *InternetRadio Deutschland (Version 4.0.7)*, [Mobile App] Google Play Store, URL https://play.google.com/store/apps/details?/id=com.appmind.radios.de, 2015. Zugriff am 23.01.24
[15] PTD STUDIO: *Internal Audio Record – Sound (Version 1.3),* [Mobile App] Google Play Store, URL https://play.google.com/store/apps/details?/id=com.ptdstudio.internalaudiorecorder, 2021. Zugriff am 23.01.24
[16] AUDACITY: Audacity(r): *Free Audio Editor and Recorder.* [Computer software]. URL https://audacityteam.org/, 2021.
[17] PAEK, H-J., NELSON, M., und VILELA, A.: Examination of Gender-Role Portrayals in Television Advertising across Seven Countries. In: Sex Roles 64, 192–207, 2011.
[18] MATTHES, J., PRIELER, M., und ADAM, K.: Gender Role Portrayals in Television Advertising Across the Globe. In: Sex Roles 75(7), 314–327, 2016.
[19] Tauchnitz, Jürgen (1990): Werbung mit Musik. Theoretische Grundlagen und experimentelle Studien zur Wirkung von Hintergrundmusik in der Rundfunk- und Fernsehwerbung. Physica-Verlag, Heidelberg, 1990.
[20] LALAL.AI: *Vocal Remover & Instrumental AI Splitter.* [Online]. URL https://www.lalal.ai/, Zugriff am 10.10.24.
[21] KISLER, T., U. D. REICHEL and F. SCHIEL: Multilingual processing of speech via web services. 2021. In: Computer Speech & Language 45, S. 326 –347, 2017.
[22] BOERSMA, P and D. WEENINK*: Praat, Doing phonetics by computer.* [Computer software] URL http://www.praat.org, Version 6.0.40, 2022.
[23] MAYER, J.: Phonetische Analysen mit Praat. Ein Handbuch für Ein-und Umsteiger. 2017.
[24] RCORE: R: *A language and environment for statistical computing.* [Computer software] URL https://www.R-project.org, 2022.
[25] WICKHAM, H.: ggplot2: *Elegant Graphics for Data Analysis.* [Computer software] URL https://ggplot2.tidyverse.org, 2016.
[26] BATES, D., M. MÄCHELER, B. BOLKER and S. WALKER: *Fitting Linear Mixed-Effects Models Using lme4.* In: *Journal of Statistical Software* 67(1), S. 1 –48, 2015.
[27] LENTH, R. V.: emmeans: *Estimated Marginal Means, aka Least-Squares Means.* [Computer software] URL https://CRAN.R-project.org/package=emmeans, 2016.
[28] MARTÍN-SANTANA, J. D.; REINARES-LARA, E.; REINARES-LARA, P.: Influence of radio spokesperson gender and vocal pitch on advertising effectiveness: The role of listener gender. In: Spanish Journal of Marketing-ESIC 21(1), S. 63–71, 2017.

# POTENZIELLE VERÄNDERUNGEN IN VOT BEI PLOSIVEN WÄHREND DES MENSTRUATIONSZYKLUS

*Lotta Hilliger, Sophie Ziemer, Adrian P. Simpson, Melanie Weirich*

*Friedrich-Schiller-Universität Jena*

**Kurzfassung:** Several studies have analysed potential gender-specific differences in voice onset time (VOT) in onset plosives in different varieties of English. Results have been variable, finding differences in some cases [10,11,16], but not in others [5,13]. Where significant differences have been found, it is female speakers who have longer VOT. It has been suggested that at least one reason for the observed variability is the failure to consider differences in laryngeal behaviour brought about by hormonal changes during the menstrual cycle [5,12]. The present study examines possible menstrual cycle-related differences in the VOT of onset /p/ and /k/ in a sample of 79 (18-45 year-olds) German-speaking women, who are part of a larger project (DFG, WE 5757/4). Female subjects were recruited who reported having a regular cycle and not taking any hormonal contraception. They were recorded during the fertile and luteal phases of the menstrual cycle. Subjects recorded short samples of read and spontaneous speech. Saliva samples were collected for hormonal analysis (progesterone, estradiol, testosterone) during each recording session. VOT was measured in fortis labial and dorsal plosives in the verb forms gepachtet ("leased") and gekauft ("bought") from two read sentences from the luteal and fertile phases. Average utterance durations were used to normalise individual VOT durations to exclude any tempo differences. VOT varies significantly as a function of place of articulation (Estimate= -21.533, SD = 2.711 $p < .001$) but values do not vary as a function of recording time during the cycle. The results of this study, to our knowledge the first of its kind for German-speaking subjects, do not replicate those found for English-speaking subjects [5,12]. We did find, on the other hand, differences between the age groups. However, in contrast to the English studies, we have yet to analyze the realisation of the fortis lenis contrast in terms of VOT, as well as making a comparison with male controls, as was done in the English studies. Although two cycle phases were considered, it will be interesting to see whether the hormone values (not yet available) are able to reveal a more differentiated picture.

## 1 Einleitung

Die Stimme verändert sich im Laufe des Lebens, von der Kindheit zum Erwachsenenalter bis hin zum fortgeschrittenen Alter. Dass Hormone bei dieser Veränderung eine wichtige Rolle spielen, ist weitestgehend bekannt. Die weibliche Stimme unterliegt dabei von der Kindheit bis zur Menopause dem abwechselnden Einfluss der Hormone Östrogen, Progesteron und Testosteron [1].

In mehreren Studien wurden bereits stimmliche Veränderungen im Zusammenhang mit dem Menstruationszyklus nachgewiesen, typischerweise in Verbindung mit der Einnahme von Antibabypillen oder bedingt durch das Auftreten von Heiserkeit durch die Menstruation [1,2, 3,4,5]. Aber auch ohne den Einfluss hormoneller Verhütungsmittel konnte in mehreren Studien bereits ein Einfluss der weiblichen Hormone in den verschiedenen Phasen des Menstruationszyklus festgestellt werden. Beispielsweise vermerkten einige Studien einen Anstieg der mittleren Grundfrequenz (f0) vor dem Eisprung [6,7]. Weitere Studien zeigten, dass die mittlere f0 nach der Menopause abnimmt in Folge einer Abnahme der Östrogene und einer Zunahme der Androgene [1]. Bryant und Haselton [6] fanden in ihrer Studie heraus, dass Änderungen in der Tonhöhe die zyklische Fruchtbarkeit zu verfolgen scheinen. Somit galt diese

Studie als erster und vorläufig einziger Nachweis eines spezifischen zyklischen Fruchtbarkeitsmerkmals in der menschlichen Stimme. Die Studienlage dazu ist jedoch nicht eindeutig, da es auch Studien gibt, die zu einem anderen Ergebnis kommen und keinen Unterschied zwischen Grundfrequenzen vor und nach der Menopause finden, wie beispielsweise Mendes-Laureano et al. [8] Ob Fruchtbarkeitsmerkmale auch in einzelnen Lauten nachzuweisen sind, ist bisher im Englischen [5,12], aber noch nicht im Deutschen untersucht worden. In Anbetracht dessen soll in der vorliegenden Arbeit betrachtet werden, ob es einen Zusammenhang zwischen der Phase des Menstruationszyklusses (luteal vs. fertil) und der Voice Onset Time gibt. Untersuchungsgegenstand sind hierfür die stimmlosen Plosive /p/ und /k/.

## 2 Begriffsklärung

Pompino-Marschall [9] definiert die VOT als die Zeit zwischen der oralen Verschlusslösung und dem Einsetzen der Stimmlippenschwingung als Merkmal der Stimmhaft-stimmlos-Opposition. Diese unterscheidet als Parameter die Kategorien der stimmhaften (mit negativer VOT, d.h. Stimmlippen schwingen vor der Verschlusslösung), der stimmlos nicht aspirierten (VOT = 0 bis 20-30ms) und der stimmlos aspirierten Plosive (VOT > 20-30 ms) [9].

Karlsson, Zetterholm und Sullivan [10] belegten in ihrer Studie, dass eine einzelne akustische Messung der Zeitspanne zwischen dem Lösen eines Konsonanten und dem Beginn der Stimmbandschwingung ausreichend Informationen enthielt, um den Unterschied zwischen stimmhaften und stimmlosen Plosiven zu bestimmen. Bei der Messung sind besonders die Dauer zwischen dem Verschluss und dem Einsetzen der Stimmhaftigkeit und die Gesamtamplitude ausschlaggebend [10]. Um die Dauer zwischen Verschluss und Einsetzen der Stimmhaftigkeit des Vokals zu messen, liegt für die hier untersuchten Laute eine CVC-Silbenstruktur vor. Der Plosiv befindet sich dabei im Onset.

## 3 Forschungsstand zu VOT

Eine der ersten Studien, die sich mit der Untersuchung von genderspezifischen Unterschieden bei der VOT-Produktion beschäftigt, wurde von Swartz 1992 durchgeführt [11]. Darin untersuchte er mögliche Unterschiede der männlichen und weiblichen Stimmeinsatzzeit der Plosive /d/ und /t/, bei denen er einen signifikanten geschlechtsspezifischen Unterschied feststellte. Demnach produzieren Männer eine kürzere VOT als Frauen. Zudem brachte Swartz [11] die VOT in Verbindung mit der Sprechgeschwindigkeit (speech rate), hier konnte jedoch kein signifikanter Zusammenhang festgestellt werden.

Weitere Studien zu geschlechtsspezifischen Unterschieden der Produktion der VOT wurden im Laufe der Jahre durchgeführt [10,12,13]. Karlsson et al. [10] erläutern, dass die durchschnittliche Länge der Stimmlippenmembran bei erwachsenen Frauen 6mm kürzer ist als bei erwachsenen Männern. Die kürzere Länge der Membran ermöglicht eine schnellere Schließbewegung, was durch einen höheren durchschnittlichen f0-Wert der Sprache von Frauen im Vergleich zu der von Männern belegt wird [10]. Das bedeutet, wenn die VOT abhängig ist von der Abduktionsgeschwindigkeit der Stimmlippen, würde dies die von Frauen und von Männern produzierten Plosive ungleich beeinflussen. Dadurch entstünde ein *gender bias* [10].

Swartz [11] untersuchte diesen vermeintlichen Einfluss des Geschlechts auf die VOT. Er stellte einen signifikanten Unterschied der VOT in Abhängigkeit des Geschlechts unter Einsatz von Messungen der VOT fest, die anhand der Wellenform von Sprachproduktionen von männlichen und weiblichen Erwachsenen mit der Muttersprache Englisch erkannt wurden.

Karlsson et al. [10] stellten in ihrer Studie zum Schwedischen darüber hinaus fest, dass dieser Unterschied nicht in Zusammenhang mit der höheren Sprechgeschwindigkeit der Männer im Vergleich zu den Frauen steht. Ihre Studie war eine der ersten, die alle auch im Deutschen vorkommenden Plosive mit Blick auf die VOT untersucht hat und dabei den Fokus

auf die Unterschiede zwischen den binären Geschlechtern legte. Ein Unterschied zwischen den Geschlechtern trat nur in der VOT der stimmhaften Plosive auf: der Wert der VOT von den Frauen war dort kleiner als bei den männlichen Probanden.

In der Untersuchung von Morris und McCrea [13] zu VOT der Plosive /b/, /p/, /d/, /t/, /g/ und /k/ kam heraus, dass die VOT zusätzlich zum Artikulationsort auch davon abhängig ist, welcher Vokal nach dem Plosiv gebildet wird. So erzeugt der Vokal /a/ eine kürzere VOT des vorhergehenden Plosivs als die Vokale /i/ und /u/.

Whiteside, Hanson und Cowell [5] erforschten in ihrer Studie erstmals zusätzlich zum Einfluss des Geschlechts den Einfluss des weiblichen Zyklus auf die Produktion der VOT. Anlass für diese Studie war die Vermutung, dass bisherige Ergebnisse der geschlechtsspezifischen Unterschiede der Stimmeinsatzzeit aufgrund der fehlenden Betrachtung dieses Faktors inkonsistent waren [5]. In der englischsprachigen Studie wurden insgesamt die VOT der Plosive /b/, /p/, /d/, /t/, /g/ und /k/ von sieben weiblichen Sprecherinnen mit denen von sechs männlichen Sprechern verglichen. Eine Hälfte der Frauen wurde zuerst an einem Zeitpunkt aufgenommen, an dem die Östrogen- und die Progesteronwerte niedrig waren, während die andere Hälfte zuerst Aufnahmen während einer Phase machte, in der die Hormonwerte hoch waren. Ein Ergebnis war, dass besonders niedrige Progesteron- und Östrogenwerte die VOT der Frauen bei den stimmlosen Plosiven beeinflusst hat. Ein weiteres Ergebnis war, dass die Männer eine schnellere Sprechgeschwindigkeit aufwiesen, die eine Verkürzung der VOT zur Folge hatte.

Diese Studie wurde 2006 von Wadnerkar et al. [12] repliziert und erweitert: die 15 weiblichen Probandinnen wurden jeweils zweimal zu unterschiedlichen Zeitpunkten ihres Zyklus aufgenommen. Der Fokus dieser Studie lag ausschließlich auf den Plosiven /p/, /b/, /k/ und /g/. Die VOT-Werte der weiblichen Probandinnen aus den zwei Zeitpunkten wurde ebenfalls wieder mit Werten von männlichen Probanden verglichen. Die Ergebnisse von Whiteside et al. [5] wurden ebenfalls repliziert: der Kontrast der VOT zwischen stimmhaften und stimmlosen Plosiven bei den Frauen war zum Zeitpunkt, an dem die Hormonwerte hoch waren, größer als bei den Männern.

Die Forschungsfrage lautet, ob die VOT der stimmlosen Plosive /p/ und /k/ vom weiblichen Menstruationszyklus beeinflusst werden. Dazu wurden die Daten aus der lutealen und aus der fertilen Phase miteinander verglichen. Das Design der Untersuchung orientiert sich an dem Ansatz von Wadnerkar, Cowell, Whiteside [12] und wendet deren Vorgehen auf deutschsprachige Daten an. Eine ähnliche Untersuchung der VOT in Verbindung mit dem weiblichen Hormonzyklus liegt für das Deutsche nach unserem Stand noch nicht vor. Daraus leitet sich die Hypothese her, dass die Bildung der stimmlosen Plosive abhängig ist von den Hormonen in den unterschiedlichen Phasen des Menstruationszyklus und dass dies in der VOT nachzuweisen ist.

## 4 Methode und Daten

Als Datengrundlage für diese Untersuchung wurden die Daten aus dem DFG-Projekt (DFG, WE5757/4) verwendet. Im Zuge des Projekts wurden insgesamt 79 Frauen im Alter zwischen 18 und 45 Jahren (mean: 29 Jahre) zu zwei Zeitpunkten in ihrem Zyklus aufgenommen. Die Probandinnen wurden nach ihrem Alter in drei Gruppen unterteilt: F1, 18-25 Jahre (mean: 22,5 J), F2, 26-35 Jahre (mean: 29,7 J), und F3, 36-45 Jahre (mean: 40,8 J). In der Altersgruppe F1 befanden sich 35 Probandinnen, in F2 25 Probandinnen und 19 Probandinnen in F3. Die Teilnehmerinnen nahmen jeweils eine Sprachaufnahme in ihrer fertilen und eine in ihrer lutealen Phase auf, wobei 39 Probandinnen ihren ersten Termin zur Versuchsteilnahme in der fertilen und 38 ihren ersten Termin in der lutealen Phase hatten. Bei der Betrachtung der Daten wurden zwei Probandinnen, die jeweils nur an einem Termin teilgenommen hatten, ausgeschlossen. Die jeweiligen Phasen errechneten sich, indem von dem Anfang des nächsten

vorhergesagten Zyklus jeweils 15-17 (fertil) und 5-7 (luteal) Tage zurückgerechnet wurde. Alle Probandinnen gaben an, dass sie keine hormonellen Verhütungsmittel einnahmen, nicht schwanger waren oder stillten und einen geregelten Zyklus haben. Die Sprachaufnahmen zu den zwei Zeitpunkten im Zyklus umfassten das Lesen von fünfzehn Sätzen, das isolierte Sprechen der Vokale /a, e, i, o, u/, Zählen von 21 bis 30, eine kurze Bildbeschreibung und das Lesen eines kurzen Textes. Auch wurden über einen Fragebogen persönliche Daten der Probandinnen (Alter, Herkunft, Bildungsstand, Familienstand etc.), zu ihrem Zyklus und zur Wahrnehmung von Femininität/Maskulinität sowie zu Geschlechterrollenbildern (GEPAQ) erhoben. Außerdem gaben die Probandinnen an jedem Termin Speichelproben ab, indem sie innerhalb einer Stunde dreimal in demselben Röhrchen Speichel sammelten. Für die Teilnahme an der Studie erhielten die Probandinnen eine Aufwandsentschädigung von 25 Euro.

Für die Analyse der VOT in den stimmlosen Plosiven /k/ und /p/ wurden aus den aufgenommenen Sprachaufnahmen zwei Sätze ausgewählt, die ein passendes Wort mit den gesuchten Plosiven (Plosiv im Onset) enthält. So wurden die Sätze „Warum hast du es denn gepachtet?“ und „Warum hat er es noch nicht gekauft?“ ausgewählt und zur weiteren Bearbeitung aus den Gesamtaufnahmen ausgeschnitten. Zur Bearbeitung der Audioaufnahmen wurde mit praat [14] gearbeitet. Zielplosive in diesen Aussagen sind das /p/ in „gepachtet“ und das /k/ in „gekauft“. Der Datensatz umfasste dann vier Aufnahmen pro Sprecherin, die jeweils einen Satz enthalten (2 Zeitpunkte der Aufnahmen x 2 Sätze pro Aufnahme).

Mit der Software WebMAUS Basic [15] wurden TextGrid-Dateien zu jeder Aufnahme erstellt. Die Grenzen der Trägerwörter und der Zielplosive wurden manuell überprüft. Die Grenzen wurden so angepasst, dass die davor und danach liegenden Vokale sauber vom Plosiv abgegrenzt sind. Anschließend wurde innerhalb des Plosivs der Bereich der Aspiration ab dem Verschluss mit einer Grenze markiert. Mithilfe eines Praat Skriptes wurde die Dauer der Aspiration und die Gesamtlänge der Äußerung erfasst. Beide Messwerte von allen Probandinnen wurden mit weiteren Angaben (Nummer der Probandin, der Altersgruppe, Zeitpunkt der Aufnahme, Zyklusphase, Zielwort und Plosiv) vollständig als Tabelle in einer csv-Datei ausgegeben. Zum Schluss wurden die VOT-Messwerte wie folgt normalisiert. Aus den Äußerungsdauern jeden Satzes wurden jeweils eine mittlere Dauer berechnet. Jede VOT-Messung wurde durch die Einzeläußerungsdauer geteilt und mit dem jeweiligen Äußerungsmittelwert multipliziert. Die so aufbereiteten Daten wurden mit der Software R (4.4.1) visualisiert und statistisch ausgewertet. Dafür wurden zusätzlich die Pakete dplyr, emmeans, forcats, ggplot2, gridExtra, lattice, lme4, lmerTest, Matrix, MuMIn und sjstats verwendet.

## 5 Ergebnisse

Die Auswertung zu der Frage, ob die VOT der Plosive Unterschiede in der fertilen oder lutealen Phase aufweist, bestätigt diese nicht. Wie in Abbildung 1 zu sehen ist, liegen die Mittelwerte der VOT zu beiden Zeitpunkten im Zyklus nah beieinander. Diese Beobachtung wurde anhand eines t-test bestätigt: Die Mittelwerte für fertil = 66 ms und luteal = 65 ms unterscheiden sich kaum ($p = .75$). Einzig die Werte der VOT der Plosive variieren in Abhängigkeit der Artikulationsstelle. So liegt der Mittelwert des Plosivs /k/ in beiden Phasen bei 68 ms während der Wert für /p/ sich bei 64 ms bewegt. Dieser Unterschied mit einer Signifikanz von $p = .035$ zwischen den Ergebnissen von /k/ und /p/ ist aber auf die unterschiedlichen Artikulationsorte zurückzuführen [9] und ist ein erwartbares Ergebnis, welches sich lediglich bestätigt hat. Mit diesen Ergebnissen können wir unsere Hauptthese nicht bestätigen. Unsere Ergebnisse lassen keine Veränderung der VOT-Werte während des Zyklus erkennen.

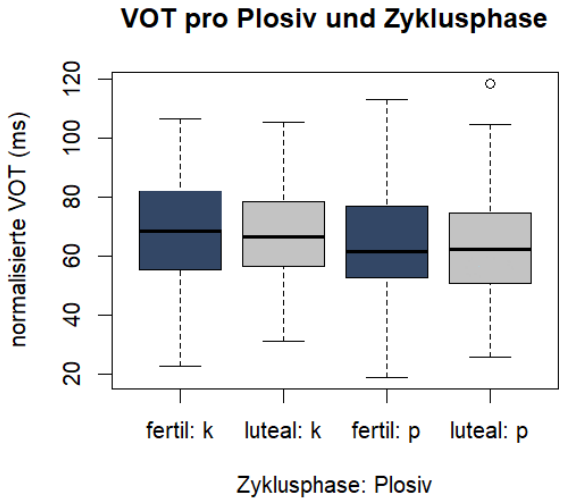



**Abbildung 1 -** Verteilung der VOT von p und k auf die Zyklusphasen 1

In einem weiteren Schritt haben wir die Plosive getrennt voneinander betrachtet und das Alter bzw. die unterschiedlichen Altersgruppen der Probandinnen sowie die Zyklusphasen mit in die Betrachtung einbezogen. Bei /k/ ergab die Verteilung der Werte auf die drei Altersgruppen Folgendes: Wie in Abbildung 2 erkennbar ist, steigt der Mittelwert der VOT mit dem Alter der Probandinnen an. Die Unterschiede zwischen den Altersgruppen betragen jeweils 2 ms. Der Wert für F1 liegt normalisiert im Mittel bei 66 ms, steigt bei F2 auf 68 ms und bis zu 70 ms für die Gruppe F3. Dieses Ergebnis ist auch dann erkennbar, wenn zusätzlich zu den Altersgruppen ebenfalls die Zyklusphasen mit in die Betrachtung einbezogen werden: Der Anstieg hin zu F3 lässt sich sowohl in der fertilen als auch in der lutealen Phase beobachten. Allerdings sind die Unterschiede in der fertilen Phase deutlicher ausgeprägt als in der lutealen. In der fertilen Phase betragen die Abstände in den normalisierten Werten zwischen den Altersgruppen jeweils ca. 2 ms. Der Unterschied in den Werten der Lutealphase dagegen ist nicht linear. Zwischen dem Wert der F1- und der F2-Gruppe liegt ein Abstand von 3 ms. Der Abstand vom F2-Wert zum F3-Wert dagegen ist mit ca. 2 ms geringer. Bei dem Plosiv /k/ sind die Unterschiede in den Messwerten der VOT für die einzelnen Altersgruppen zwischen den Zyklusphasen sehr gering.

Auch bei /p/ sind Unterschiede zwischen den Altersgruppen erkennbar, siehe Abbildung 3. Während bei /k/ die Werte der VOT mit dem Alter steigen, liegen für /p/ die Mittelwerte der F2-Gruppe niedriger als die der anderen beiden Gruppen. Beim Hinzufügen der Zyklusphasen mit in die Betrachtung ist erneut erkennbar, dass die normalisierten Mittelwerte der F2-Gruppe niedriger sind als die der anderen beiden. In der Fertilphase verzeichnet die Altersgruppe F3 den höchsten Wert, danach kommt der Wert der Gruppe F1, gefolgt von der Altersgruppe F2 (F3>F1>F2). In der Lutealphase lässt sich eine andere Reihenfolge feststellen. Da liegt der Wert der F1-Gruppe am höchsten, danach folgt der Mittelwert der F3-Gruppe und auch hier zuletzt die Gruppe F2 (F1>F3>F2). Während bei der Gruppe F1 der Wert in der lutealen Phase höher ist als in der fertilen Phase, ist bei den anderen Altersgruppen genau das Gegenteil zu beobachten: Bei F2 und F3 liegt jeweils der Wert der fertilen Phase über dem der lutealen. Im Vergleich mit den anderen Altersgruppen verzeichnet F3 die größte Spanne an Veränderung innerhalb einer Altersgruppe zwischen den Zyklusphasen.

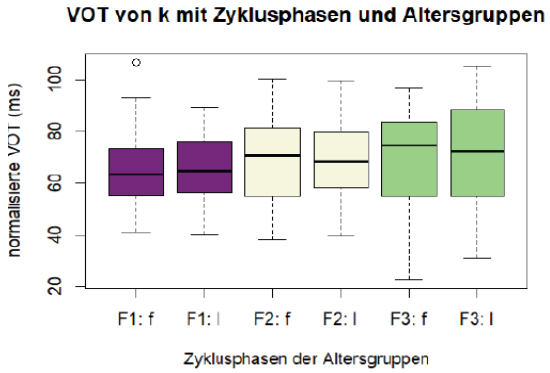



**Abbildung 2 -** VOT des Plosivs k verteilt auf die Zyklusphasen und die Altersgruppen

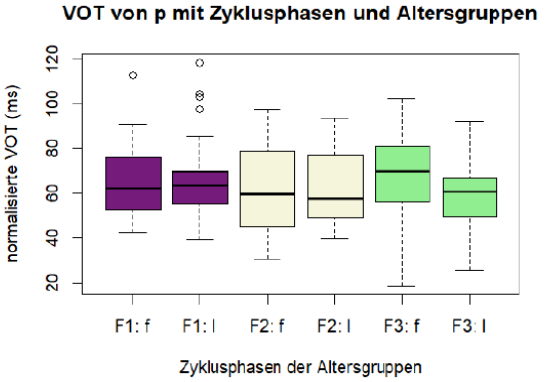



**Abbildung 3 -** VOT des Plosivs p verteilt auf die Zyklusphasen und die Altersgruppen

# 6 Diskussion

Die Ergebnisse zeigen, dass für die Anzahl von 79 Sprecherinnen kein Unterschied in der VOT von Plosiven zu unterschiedlichen Zeitpunkten im Zyklus zu erkennen ist. Nach unseren Erkenntnissen ist an der Bildung der Plosive nicht erkennbar, in welchem hormonellen Zustand sich die Probandinnen zum Zeitpunkt der Aufnahme befinden, siehe Abbildung 1. Interessant bei den Messwerten ist auch der geringe Abstand der VOT der Plosive /p/ und /k/. In den Ergebnissen sind höhere Werte der VOT für den Plosiv /k/ erkennbar, was nach Pompino-Marschall [9] durchaus zu erwarten ist, da mit nach hinten verlagertem Artikulationsort auch die Lösung des Plosivs länger wird.

Die Informationen zu den Zeitpunkten im Zyklus, ob demnach die fertile oder die luteale Phase bei der Probandin vorliegt, beruhen auf den Selbstaussagen der Probandinnen. Zwar haben alle Probandinnen einen regelmäßigen Menstruationszyklus angegeben, dennoch können naturbedingte Schwankungen nicht ausgeschlossen werden. Zu dem Zeitpunkt der Analyse der Daten lagen keine exakten Hormonwerte vor, die eine Einbeziehung des individuellen Hormonspiegels in die Untersuchung ermöglicht hätten.

Aufgrund der Anzahl der Probandinnen und den Ergebnissen, dass die Mittelwerte für die Plosive während des lutealen und des fertilen Zeitpunkts sich kaum unterscheiden, wird weiterhin die Erkenntnis bestätigt, dass die Hormonzyklusphasen in dieser Untersuchung keine Auswirkungen auf die VOT der Plosive /p/ und /k/ hat. Obwohl die Hauptthese damit nicht bestätigt werden konnte, ergeben die Daten aber eine andere Erkenntnis bezüglich des Alters der Probandinnen.

Besonders bei /k/ ist erkennbar, dass das Alter der Probandinnen eine Auswirkung auf die VOT hat. In einem Rahmen von 2 Millisekunden stieg die VOT mit dem Alter der Probandinnen an. Bei /k/ variiert diese Erkenntnis zusätzlich je nach hormoneller Phase: der Anstieg ist in der fertilen Phase deutlicher zu erkennen als in der Lutealphase. Die Ergebnisse für /p/ dagegen zeichnen ein abweichendes Bild für die Verteilung auf die Altersgruppen. Dort ist vorrangig erkennbar, dass die Werte der VOT für die zweite Altersgruppe F2 niedriger sind als für die anderen beiden Gruppen.

Auch bei /p/ ist das Ergebnis vielfältiger, wenn in die Betrachtung auch die Zyklusphasen hinzugefügt werden. In der Verteilung der VOT von /p/ auf die Altersgruppen und die Zyklusphasen ist besonders ein Unterschied in der F3-Gruppe zwischen der fertilen und der lutealen Phase zu sehen. Innerhalb der Altersgruppen liegt sowohl bei der F2- als auch bei der F3-Gruppe der fertile Wert höher, während bei der Gruppe F1 der Wert der Lutealphase über

dem der fertilen liegt. Hierbei muss aber festgehalten werden, dass die vorliegende Analyse des Effektes von Alter und dessen Interaktion mit der Zyklusphase nur eine beschreibende ist und keine statistische Auswertung gemacht wurde. Die Vergleichbarkeit der Altersgruppen wird jedoch dadurch erschwert, dass die Verteilung der Probandinnen auf die Gruppen nicht gleichmäßig erfolgt. In der Altersgruppe der 18-25-Jährigen sind die meisten Probandinnen aufgenommen worden und mit steigendem Alter nimmt die Anzahl der Probandinnen ab. Das Alter der Probandinnen kann Auswirkungen auf den Hormonhaushalt haben, was die Unterschiede erklären kann. Auch sind sicherlich Unterschiede in den gesellschaftlichen Generationen vorhanden, die die Sprechweise und Stimmbildung der Probandinnen beeinflussen. Eine Betrachtung der Ergebnisse aus soziophonetischer Perspektive würde sich anbieten.

# 7 Fazit

Die Hypothese, dass der Hormonhaushalt einen Einfluss auf die Bildung der stimmlosen Plosive hat und dies in der VOT nachweisbar ist, hat sich mit den hier behandelten Messwerten von 79 deutschsprachigen Frauen im Alter von 18 bis 45 Jahren nicht bestätigt. Der hormonelle Zustand der Probandinnen ist an den Ergebnissen nicht erkennbar. Da aufgrund der Ergebnisse kein Unterschied in den Zyklusphasen zwischen den Plosiven p und k gefunden werden konnte, bleibt die Aussage bestehen, dass die Zyklusphase keine Auswirkung auf die VOT von stimmlosen Plosiven hat.

Jedoch deuten die Ergebnisse darauf hin, dass die VOT möglicherweise vom Alter abhängig ist. Die Vergleichbarkeit der Ergebnisse wird durch die unterschiedliche Verteilung der Probandinnen auf die Altersgruppen erschwert. Die Unterschiede in den Altersgruppen sind möglicherweise durch den unterschiedlichen Hormonhaushalt des jeweiligen Alters sowie durch Unterschiede der gesellschaftlichen Generationen bedingt, weshalb die Ergebnisse für eine genauere Untersuchung noch einmal aus soziophonetischer Perspektive betrachtet werden sollten. Im Rahmen des Projekts werden außerdem weitere Analysen durchgeführt und neben der VOT der stimmlosen Plosive auch die Realisation der stimmhaften Plosive und der fortis-lenis Kontrast betrachtet. Die Ergebnisse aus der Studie von Whiteside [5] konnten vorerst nicht bestätigt werden.

# 8 Literatur


[1] ABITBOL, B., J. ABITBOL AND P. ABITBOL: Sex hormones and the female voice. In: Journal of Voice. Vol. 13, Issue 3, p. 424 – 446, 1999.

[2] HIGGINS, M. B. AND J. H. SAXMAN: Variations in vocal frequency perturbation on across the menstrual cycle. In: Journal of Voice, Vol. 3, Issue 3, p. 233 – 243, 1989.

[3] AMIR O., T. BIRON-SHENTAL, C. MUCHNIK AND L. KISHON-RABIN: Do oral contraceptives improve vocal quality? Limited trial on low-dose formulations. In: Obstet Gynecol, Vol. 101, Issue 4, p. 773 – 777, 2003.

[4] KISHON-RABIN L., S. PATAEL, M. MENAHEMI AND N. AMIR: Are the perceptual effects of spectral smearing influenced by speaker gender? In: J Basic Clin Physiol Pharmacol, Vol. 15, Issue 1-2, p. 41 – 55, 2004.

[5] WHITESIDE, S. P., A. HANSON AND P. E. COWELL: Hormones and temporal components of speech: Sex differences and effects of menstrual cyclicity on speech. In: Neuroscience Letters, Vol. 367, Issue 1, p. 44 – 47, 2004.

[6] BRYANT, G. A. AND M. G. HASELTON: Vocal cues of ovulation in human females. In: Biology Letters, Vol. 5, Issue, 1, p. 12 – 15, 2009.

[7] FISCHER, P. JI. KRUEGER, T. GREITEMEYER, C. VOGRINIC, D. FREY, M. HEENE, M. WICHER AND M. KAINBACHER: The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. In: Psychol Bull, p. 517 – 537, 2011.

[8] MENDES-LAUREANO, J., M. F. S. SÁ, R. A. FERRIANI, R. M. REIS, L. N. AGUIAR-RICZ, F. C. P. VALERA, D. S. KÜPPER AND G. S. ROMÃO: Comparison of fundamental voice frequency between menopausal women and women at menacme. In: Maturitas, Vol. 55, Issue 2, p. 195 – 199, 2006.
[9] POMPINO-MARSCHALL, B.: Einführung in die Phonetik. De Gruyter, 1995.
[10] KARLSSON, F., E. ZETTERHOLM AND K. P. SULLIVAN: Development of a gender difference in voice onset time. In: Proc. Of the 10th Australian international conference on speech science & technology, p. 316 – 321, 2004.
[11] SWARTZ, B. L.: Gender difference in voice onset time. In: Perceptual and motor skills, Vol. 75, Issue 3, p. 983 – 992, 1992.
[12] WADNERKAR, M. B, P. E. COWELL AND S. P. WHITESIDE: Speech across the menstrual cycle: A replication and extension study. In: Neuroscience Letters, Vol. 408, Issue 1, p. 21 - 24, 2006.
[13] MORRIS, R. J., C. R. MCCREA AND K. D. HERRING: Voice onset time differences betweenadult males and females: Isolated syllables. In: Journal of phonetics, Vol. 36, Issue 2, p. 308 – 317, 2008.
[14] BOERSMA, P. AND D. WEENINK: Praat. Version 6.3.02, 2022.
[15] SCHIEL, F.: A statistical Model for Predicting Pronounciation. In: Proc. Of the ICPhS, paper 195, 2015. (MAUS (Munich Automatic Segmentation))
[16] ROMEO, R., V. HAZAN AND M. PETTINATO: Developmental and gender-related trends of intra-talker variability in consonant production. In: Journal of the Acoustical Society of America, Vol. 134, Issue 5, p. 3781 – 3791, 2013.

# ANFORDERUNGEN AN DIE GESANGSAUSSPRACHE IM 19. JAHRHUNDERT

## (AM BEISPIEL VON WAGNERS "DER RING DES NIBELUNGEN")

*Ursula Hirschfeld*

*Martin-Luther-Universität Halle-Wittenberg / Sprechwissenschaft und Phonetik*
*ursula.hirschfeld@sprechwiss.uni-halle.de*

**Kurzfassung:** Im 19. Jahrhundert bildeten sich nicht nur Normen für die geschriebene und gesprochene deutsche Sprache heraus, sondern auch für die Gesangsaussprache. Eine besondere Rolle spielte hierbei Richard Wagner, der der Textverständlichkeit einen wichtigen Stellenwert beimaß. Im Rahmen eines Projektes, an dem Expert/-innen aus Musik-, Theater- und Sprechwissenschaft beteiligt sind, werden seine Vorstellungen von einer regelhaften und verständlichen Aussprache für die Uraufführung des „Ring des Nibelungen" (1876) untersucht und in historisch informierten, konzertanten Vorstellungen auf der Bühne umgesetzt. Der Beitrag geht auf ausgewählte Aspekte der Entwicklung und Untersuchung von Aussprachenormen ein und beschreibt die Forschungsarbeiten zu Wagners Anforderungen an eine verständliche Gesangsaussprache.

## 1 Entwicklung und Untersuchung von Aussprachenormen

Im 19. Jahrhundert bildeten sich Normen für die geschriebene und gesprochene deutsche Sprache heraus. Ansatzpunkte für die Beschreibung der „richtigen" oder „hochdeutschen" Aussprache waren die Schreibung, die Sprechweise in einzelnen Regionen bzw. die der Gebildeten einzelner Regionen. Es erschienen zahlreiche Publikationen, die bei diesen unterschiedlichen Ansätzen vor allem Regeln für eine allgemein verständliche, schrifttreue und möglichst dialektfreie Aussprache in öffentlichen bzw. offiziellen Situationen und für gesellschaftliche und künstlerische Bereiche (wie Schule, Kirche, Bühne) formulierten. Dies ist insbesondere nach der Gründung des Deutschen Reiches und dem Erscheinen des Rechtschreibdudens (1871) zu beobachten.

Doch bereits zu Beginn des 19. Jahrhunderts (1803) verfasste Goethe als Theaterdirektor in Weimar „Regeln für Schauspieler" [1], in denen er eine „reine und vollständige" Aussprache jedes einzelnen Wortes forderte. Goethe orientiert sich an der Schreibung: „Vollständig aber ist die Aussprache, wenn kein Buchstabe eines Wortes unterdrückt wird, sondern wo alle nach ihrem wahren Werte hervorkommen." (§ 4). Er wandte sich gegen eine dialektale Sprechweise auf der Bühne: „Wenn mitten in einer tragischen Rede sich ein Provinzialismus eindrängt, so wird die schönste Dichtung verunstaltet und das Gehör des Zuschauers beleidigt. Daher ist das Erste und Notwendigste für den sich bildenden Schauspieler, daß er sich von allen Fehlern des Dialekts befreie und eine vollständige reine Aussprache zu erlangen suche. Kein Provinzialismus taugt auf die Bühne! Dort herrsche nur die reine deutsche Mundart, wie sie durch Geschmack, Kunst und Wissenschaft ausgebildet und verfeinert worden." (§ 1). Diese Ausführungen können als eine erste Definition einer überregionalen Standardaussprache betrachtet werden.

Ein erstes allgemein anerkanntes, auf Beobachtungen der Aussprache von Schauspielern beim klassischen Versdrama beruhendes Regelwerk wurde 1898 von Theodor Siebs herausgegeben. Siebs leitete eine Kommission aus Sprachwissenschaftlern und Vertretern deutscher Bühnen, mit denen er Beratungen zu Ausspracheregeln abhielt, deren Ergebnisse in der „Deut-

sche Bühnenaussprache“ veröffentlicht wurden [2]. Die erste Auflage dieses Regelwerkes enthielt neben einleitenden Vorträgen zwar eine zusammenhängende Aussprachelehre, jedoch noch kein Wörterverzeichnis. Der „Siebs“ erschien bis 1969 in 19 Auflagen und ist teilweise noch heute ein in der Schauspiel- und Gesangsausbildung genutztes Nachschlagewerk.

Seit den 1950-er Jahren ist die Untersuchung und Festlegung von Aussprachenormen ein Forschungsschwerpunkt der halleschen Sprechwissenschaft. Es wurden auf der Grundlage empirischer Untersuchungen (überwiegend zur Rundfunkaussprache) drei Aussprachewörterbücher erarbeitet:

a) „Wörterbuch der deutschen Aussprache“ 1964 [3] – unter der Leitung von Hans Krech, der sich auch mit „Richard Wagner als Sänger und Sprecher“ beschäftigt hat [4];

b) „Großes Wörterbuch der deutschen Aussprache“ (1982) [5] und

c) „Deutsches Aussprachewörterbuch“ (2009) mit einer digitalen Auskopplung des Wörterverzeichnisses in der „Deutschen Aussprachedatenbank“ [6].

Der Fokus lag dabei immer auf der jeweils aktuellen Abbildung des Entwicklungsstandes der gesprochenen Standardsprache. In allen drei Wörterbüchern gibt es auch einen Abschnitt zur Gesangsaussprache. Das 19. Jahrhundert wurde erst im Rahmen der 2018 begonnenen Kooperation des Forschungsprojektes *Wagner-Lesarten* mit der Abteilung Sprechwissenschaft und Phonetik der Martin-Luther-Universität Halle-Wittenberg einbezogen. Da es damals noch nicht die Möglichkeit gab, gesprochene Sprache aufzunehmen und zu konservieren, wurden und werden ca. 150 Quellen aus dem 19. Jahrhundert für eine wissenschaftliche Auseinandersetzung mit der Beschreibung bzw. Festlegung von Aussprachenormen hinzugezogen.

## 2 Projekt zur Aussprache in Wagners „Der Ring des Nibelungen“ (1876)

Für das 150. Jubiläum der Uraufführung von Richard Wagners „Der Ring des Nibelungen“ bei den Bayreuther Festspielen im Jahr 1876, das im Jahr 2026 begangen wird, wurde und wird der vierteilige Opernzyklus historisch informiert (und konzertant) aufgeführt: *Rheingold* 2021 (in Köln im Rahmen von *Wagner-Lesarten*) und 2023 (in Dresden im Rahmen von *The Wagner Cycles*), *Walküre* 2024 (in Prag), *Siegfried* 2025 (im Rahmen von *The Wagner Cycles* in Prag) und *Götterdämmerung* 2026 (im Rahmen von *The Wagner Cycles*). Historisch informiert bedeutet nicht nur, dass sich die musikalischen Charakteristika dem Ursprung annähern, dass z. B. damals verwendete Instrumente nachgebaut werden, sondern auch, dass die damalige Gesangsaussprache rekonstruiert wird.

Bereits 2017 begannen wissenschaftliche und künstlerische Vorarbeiten, die in einem Projekt der Kunststiftung NRW, von Concerto Köln, Kent Nagano und wissenschaftlichen Partnern wie der Hochschule für Musik und Tanz Köln, dem Projekt *Wagner-Lesarten* (Laufzeit von 2018 bis 2023), fortgesetzt und intensiviert wurden [7]. Seit 2023 läuft das Nachfolgeprojekt *The Wagner Cycles* [8] unter der künstlerischen Leitung des Projektinitiators und Dirigenten Kent Nagano sowie des Intendanten der Dresdner Musikfestspiele Jan Vogler. Musik-, theater- und sprechwissenschaftlichen Untersuchungen sind fester Bestandteil des Projekts (vgl. dazu auch den Bericht von Wissmann zur Walküre-Aufführung [9]).

Ein wichtiges Ergebnis sprechwissenschaftlicher Forschung ist die Dissertation von Ulrich Thilo Hoffmann (2024), die sich mit Ausspracheregeln und Normen für das Sprechen und Singen auf der Bühne im 19. Jahrhundert beschäftigt und Vokale und Diphthonge in den Mittelpunkt stellt [10]. Auf Wagner wird hier nicht explizit eingegangen. Aber es wurden und werden weitere Untersuchungen vorgenommen, die vor allem Wagners Anforderungen an die Gesangsaussprache gelten.

## 3 Richard Wagners Anforderungen an die Gesangsaussprache

Richard Wagner hat sich vielfach mit Fragen nach der „richtigen“ Aussprache des Deutschen und mit dem Spannungsfeld von Singen und Sprechen befasst. Für ihn war die klangliche Abgrenzung der deutschen von der italienischen Sprache wesentlich. Statt des „italienischen Modells“ sollte die deutsche Sprache mit ihren spezifischen Vokal-Konsonant-Verhältnissen, ihren differenzierten Wortbetonungen und einer angemessenen Gliederung der Sätze Ausgangspunkt des Singens sein. Eine reine und deutliche („scharfe“) Aussprache der Schauspieler/innen und Sänger/innen (damals oft noch in Personalunion) war ihm äußerst wichtig, wichtiger als die Gesangsqualität [11]. Immer wieder beklagte er sich über deren mangelndes Textverständnis und ihre schlechte Textverständlichkeit, z. B. in seinem „Bericht an Seine Majestät den König Ludwig II. von Bayern über eine in München zu errichtende deutsche Musikschule“: „Schon dieser einzige Umstand der gänzlich vernachlässigten und undeutlichen Aussprache unserer Sänger ist von der abschreckendsten Bedeutung für das Zustandekommen eines wahrhaft deutschen Styles in der Oper.“ [12].

In den von ihm geleiteten Proben entwickelte er neue Zugänge zu einer verständlichen Gesangsaussprache: Der Text sollte zunächst rhythmisch sprechend erarbeitet werden – mit der entsprechenden situativen und emotionalen Prägung im Sinne eines Schauspiels, und erst danach sollte die Musik dazukommen. Wagner vertrat die Auffassung, dass den Text nur richtig singen kann, wer ihn auch richtig sprechen kann. Das richtige (Aus-)Sprechen betraf dabei nicht nur die Bildung der Vokale und Konsonanten, sondern auch das Sprechtempo sowie die rhythmisch-melodische Struktur [11]. Wagner geht in seinen Schriften aber nicht genauer auf phonetische Details ein, die Beschreibung der für ihn so wichtigen Verständlichkeit der Gesangsaussprache bleibt weitgehend allgemein [11].

Wagner orientierte sich beim Komponieren an der Struktur der gesprochenen deutschen Sprache und berücksichtigte Längen und Kürzen, Hervorhebungen und Senkungen, Strukturierungen und Pausen bereits in der Partitur. Am 08.09.1850 schreibt er an Franz Liszt, er habe sich „…bemüht, den sprechenden ausdruck der rede so sicher und scharf abzuwägen und zu bezeichnen, daß die sänger nur nöthig haben sollten, *in dem angegebenen tempo genau die noten nach ihrem werthe zu singen,* um dadurch allein schon den sprechenden ausdruck in ihrer hand zu haben. Ich ersuche daher die sänger inständigst, jene redenden stellen in meiner oper zu allernächst genau im tempo – wie sie geschrieben stehen – zu singen; sie mögen sie durchgehends lebhaft, mit scharfer aussprache vortragen, so haben wir schon *viel* gewonnen; – wenn sie von dieser basis aus weitergehend mit verständiger freiheit, eher befeuernd als zurückhaltend, das peinliche des tempo's ganz verschwinden lassen und nur noch den eindruck einer erregten, poetischen redeweise hervorbringen können, – so haben wir *Alles* gewonnen.“ [13]

Für die Proben an den Ring-Opern 1875/76 hat Wagner den Gesangslehrer und Musikpädagogen Julius Hey, dessen „Sängerbildungsideal“ der hallesche Sprechwissenschaftler Hans Krech in seiner Dissertation 1941 [14] untersucht hat, nach Bayreuth geholt. Hey hatte bereits 1868 begonnen, ein mehrbändiges Regel- und Übungswerk – „Deutscher Gesangsunterricht“ [15] – für die Ausbildung von Sängerinnen und Sängern zu erarbeiten. Der ersten Band, der „Sprachlichen Theil“, der sich mit der Gesangsaussprache beschäftigt und den Hey während und nach den Proben und den Aufführungen des „Rings“ mit engem Bezug zu Wagners Vorgaben weitergeführt und 1882 veröffentlicht hat, trägt den Untertitel „Anleitung zu einer naturgemässen Behandlung der Aussprache, als Grundlage für die Gewinnung eines vaterländischen Gesangstyles“ [15]. Dieses Lehrbuch enthält sehr detaillierte Vorgaben sowie Übungen zur Artikulation von Vokalen und Konsonanten und u. a. Beschreibungen von dynamischen und rhythmischen Mustern in der deutschen Sprache. Wie schon Goethe hat sich Hey an der Schrift orientiert. Je nach Silben- und Wortkontext unterscheidet er bei einzelnen Vokalen und

Konsonanten oft mehrere Aussprachevarianten, z. B. eine unterschiedliche Aussprache unterschiedlich geschriebener gleicher Diphthonge. Ulrich Thilo Hoffmann hat wichtige Passagen für die Proben im Rahmen des Ring-Projektes in einem „Gesangsaussprache-Profil nach Julius Hey“ (2024) zusammengefasst [16]. Heys „Deutscher Gesangsunterricht“ (Teil I) ist bis heute in gekürzter und leicht geänderter Form unter dem Titel „Die Kunst des Sprechens“ („*Der kleine Hey*“) in zahlreichen Auflagen erschienen [17].

Hey verweist in der Einleitung zum „Deutschen Gesangsunterricht“ [15] unter Bezug auf Wagner mehrfach auf die undeutliche Aussprache vieler Sänger. Als Grund dafür sieht er „die Vernachlässigung des Sprachstudiums, welches vor und mit dem Beginne der Tonbildung die gründlichste Behandlung erfordert und die instrumentelle Entwickelung des zu bildenden Organs erst in die zweite Linie stellt. Denn nur auf diesem Wege wird das Ergebnis eines abgerundeten, mit wirkungsvollster Klangenergie erzeugten Tones, und die durch kunstgerechte Verschmelzung bei der Wortbildung sich ergebende vollste Verständlichkeit des Textgesangs zu erreichen sein.“ (S. 6), wobei Hey unter „Wortbildung“ die korrekte Aussprache versteht.

Hey hat sich mit der bis dahin erschienenen Literatur, meist Lehrwerken, zur Gesangsaussprache beschäftigt, er verweist auf einige wenige Autoren, die aus seiner Sicht „als einzig einschlägig“ (S. 2) anzuerkennen waren, so vor allem auf seinen Lehrer Friedrich Schmitt („Große Gesangschule für Deutschland“, 1854, und „Neues System zur Erlernung der deutschen Aussprache nebst neuer Eintheilung des A B C“, 1868), Gustav Engel („Die Consonanten der deutschen Sprache“, 1874) sowie Christian Gottfried Nehrlich („Der Kunstgesang“, 1859).

## 4 Beschreibung und Aneignung einer verständlichen Gesangsaussprache (nach Hey)

Wesentliche Angaben zu konkreten Aussprachemerkmalen, die den Vorstellungen Wagners von einer verständlichen Gesangsaussprache entsprechen, können vor allem dem ersten Teil von Heys „Deutschem Gesangsunterricht“ entnommen werden. Der folgende kurze Überblick zeigt Heys Strukturierung, allgemeine Ausspracheregeln für Vokale und Konsonanten sowie Beispiele für den Lehr- und Lernprozess [15].

Hey bezeichnet die Vokale wie folgt und behandelt sie in dieser Reihenfolge: 1. A – neutraler Grundvokal; 2. Ä – heller Mischvokal, 3. E – Nebenvokal, 4. I – heller Grundvokal, 5. AI und EI – helle Diphthonge, 6. AU – neutraler Diphthong, 7. ÄU / EU – dunkle Diphthonge, 8. O – Zwischenvokal, 9. Ö – dunkler Mischvokal, 10. Ü – dunkler Mischvokal, 11. U – dunkler Grundvokal.

Für die Vokale gelten folgende Regeln:

a. Lange und kurze Vokale sind deutlich zu unterscheiden – mit Kürzungen und Dehnungen nach den vorgegebenen Notenwerten.
b. Zwischen hell und dunkel, (ganz) offen, halbgeschlossen und geschlossen ist ebenfalls deutlich zu unterscheiden.
c. Lange Vokale sind in der Regel geschlossen, kurze Vokale offen.
d. Die kurzen Ü-, Ö- und U-Vokale haben die gleiche Qualität / Spannung wie die langen, sie sind geschlossen.

Hey unterscheidet bei den Konsonanten mehrere Gruppen und behandelt sie in dieser Reihenfolge.: 1. die Gruppe der Klinger / Liquide (L, N – Ng, M, R, W, J); 2. die tonlosen Konsonanten / Zischergruppe: Ch – vorderer Rauschlaut; S – Säusellaut; Z – scharfer Zischlaut; sanfte und verschärfte Rauschlaute: Sch (St, Sp); einfache und zusammengesetzte Blaselaute: F, V,

Pf. Es folgen 3. die Gaumenkonsonanten, und die angrenzenden Hauch- und Zischlaute (Ck / K; Q, gg, G, gh; Ch; H; D – T, B – P).

Für die Konsonanten gelten folgende Regeln:

a. Alle Konsonanten sollen „scharf“ und deutlich artikuliert werden.
b. Konsonanten dürfen nicht gedehnt werden.
c. In Konsonantenverbindungen sind alle aufeinanderfolgenden Konsonanten deutlich zu realisieren.
d. Gleiche oder ähnliche aufeinanderfolgende Konsonanten werden nicht verbunden, sondern nacheinander realisiert.
e. Es wird konsequent in allen Silbenpositionen ein Zungenspitzen-R realisiert (d. h. es werden keine vokalisierten R-Laute verwendet.
f. Es gibt keine Auslautverhärtung (d. h. es werden auch im Wort- und Silbenauslaut stimmhafte Verschlusslaute und -frikative gebildet). Im Gegensatz dazu bemerkt Wagner (in „Oper und Drama“): „Der Sänger, der aus dem Vokale den vollen Ton zu ziehen hat, empfindet sehr lebendig den bestimmenden Unterschied, den energische Konsonanten – wie K, R, P, T –, oder gar verstärkte – wie Schr, Sp, St, Pr –, und schlaffere, weiche – wie G, L, B, D, W, – auf den tönenden Laut äußern. Ein verstärkter Ablaut – nd, rt, st, st – giebt da, wo er wurzelhaft ist – wie in „Hand“, „hart“, „Hast“, „Kraft“ – dem Vokale mit solcher Bestimmtheit die Eigenthümlichkeit und Dauer seiner Kundgebung an, daß er diese letztere als eine kurze, lebhaft gedrängte geradesweges bedingt, und daher als charakteristisches Merkmal der Wurzel zum Reime – als Assonanz – sich bestimmt (wie in „Hand“ und „Mund“)“.

Breiten Raum widmet Hey der „Sprachlichen Tonbildung“ (S. 121-143) sowie dem „Dynamischen und rhythmischen Element in der deutschen Sprache“ (S. 144-170). Letzteres Kapitel geht auf prosodische Merkmale ein und behandelt: A. Klangstärke (Silben- Wort- und Satzbetonung), B. Hebung und Senkung des Sprechtones, C. Klangfarben der Sprache (Modulation) und D. Sprachrhythmus – und hier u.a. die rhythmische Satzgliederung.

Neben der Darstellung dieser fachlichen Grundlagen vermittelt Hey auch seine Erfahrungen als praktizierender Gesangslehrer, der nicht nur die der Musik und Sprache gleichermaßen zugrundeliegenden Strukturen und Merkmale der Aussprache anschaulich und mit vielen Beispielen und Bezügen zu anderen Sprachen und Dialekten beschreibt, sondern er gibt auch Anleitungen zur Aneignung dieser Strukturen und Merkmale mit zahlreichen Übungsbeispielen, die aus Einzelwörtern, Beispielen aus den Wagner-Opern und kurzen Texten in reimloser Gedichtform bestehen, z. B. für die „Momentanlaute B – P“ (S. 117 f.). Für diese beiden „Lippendrücker – Explosivlaute im wirklichsten Sinne“ wird zunächst ausführlich die Bildung beschrieben und darauf hingewiesen, dass es der „muthigsten“ Ausdauer von Seiten des Lehrers wie des Schülers bedarf, um zu einer völlig geläuterten Sprachreinigung, d. h. zur genauen Unterscheidung beider Konsonanten, zu kommen. Es folgen Ausführungen zum Auftreten beider Konsonanten in unterschiedlichen Silben- und Wortkontexten, Textbeispiele (z. B. mit einem Stabreim aus „Rheingold“: *Garstig glatter, glitschiger Glimmer, wie gleit ich aus*!). Abschließend gibt es kurze Texte von Hey, in denen sich die behandelten Laute häufen – und die sich meist durch einen nur geringen Sinngehalt auszeichnen, z. B. (S. 119 [18]):

**P**lum**p** **b**richt der **b**e**p**ackte **B**auer
Die Lau**bp**racht fal**bp**rangend **b**eim **B**irn**b**aum;
**P**rom**p**t **b**läut der er**p**robte **P**ächter
Den Die**b** im **b**aum**b**uschigen **P**arke,
Mit **B**am**b**us beim **P**um**pb**runn‘!

# 5 Literatur


[1] GOETHE, J. W.: Regeln für Schauspieler. Berliner Ausgabe. Kunsttheoretische Schriften und Übersetzungen, Band 17. Berlin, 1960 ff. http://www.zeno.org/Literatur/M/Goethe,+Johann+Wolfgang/Theoretische+Schriften/Regeln+f%C3%BCr+Schauspieler

[2] SIEBS, T.: Deutsche Bühnenaussprache. Berlin, Köln, Leipzig, 1898.

[3] KRECH, E.-M., E. KURKA, H. STELZIG, E. STOCK, U. STÖTZER, R. TESKE: Wörterbuch der deutschen Aussprache. Leipzig, 1964.

[4] KRECH, H.: Richard Wagner als Sänger und Sprecher. In: KRECH, E.-M. (Hg.): Beiträge zur Sprechwissenschaft III, S. 51-65. Berlin, 2012 (Nachdruck, Original: 1955).

[5 ] KRECH, E.-M., E. KURKA, H. STELZIG, E. STOCK, U. STÖTZER, R. TESKE: Großes Wörterbuch der deutschen Aussprache. Leipzig, 1982.

[6] KRECH, E.-M., E. STOCK, U. HIRSCHFELD, L. C. ANDERS: Deutsches Aussprachewörterbuch. Berlin, 2009. Deutsche Aussprachedatenbank: https://dad.sprechwiss.uni-halle.de/dokuwiki/doku.php/start)

[7] WAGNER-LESARTEN. https://wagner-lesarten.de/

[8] THE WAGNER CYCLES. https://www.musikfestspiele.com/de/der-ring/ueber-das-projekt

[9] WISSMANN, F.: „Vielleicht hin und wieder noch mehr Konsonanten“ – Wagners *Ring* in historischer Aufführungspraxis. In: wagner**spectrum** 39, S. 113-126. Würzburg, 2024. https://wagnerspectrum.de/aktuell.html

[10] HOFFMANN, T. U.: Aussprachenormen für das Sprechen und Singen auf der Bühne im 19. Jahrhundert. Mit besonderer Berücksichtigung der Aussprache von Vokalen und Diphthongen. Berlin, 2024. https://www.frank-timme.de/de/programm/produkt/aussprachenormen-fur-das-sprechen-und-singen-auf-der-buhne-im-19-jahrhundert?file=/site/assets/files/6864/9783732989010.pdf

[11] HIRSCHFELD, U. / MÜLLER, K. H.: „Wer g nicht von ch zu unterscheiden vermag, ist ein undeutscher Barbar … .“ – Richard Wagner und die (Gesangs-)Aussprache des Deutschen im 19. Jahrhundert. Köln, 2018. https://musiconn.qucosa.de/api/qucosa%3A34371/attachment/ATT-0/

[12] WAGNER, R.: Bericht an Seine Majestät den König Ludwig II. von Bayern über eine in München zu errichtende deutsche Musikschule. In: Richard Wagner: Werke, Schriften und Briefe, S. 3762 (vgl. Wagner-SuD Bd. 8, S. 135). http://www.digitale-bibliothek.de/band107.htm

[13] WAGNER, R.: Brief an Franz Liszt. In: Sämtliche Briefe: Bd. 3: Briefe Mai 1849 bis Mai 1851. Richard Wagner: Werke, Schriften und Briefe, S. 9445, (vgl. Wagner-SB Bd. 3, S. 387-388), http://www.digitale-bibliothek.de/band107.htm

[14] KRECH, H.: Julius Hey und sein Sängerbildungsideal ‚Deutscher Gesangs-Unterricht‘. Diss. Halle, 1941 (Mskr.).

[15] HEY, J.: Deutscher Gesangsunterricht. I. Sprachlicher Theil. Mainz 1882.

[16] HOFFMANN, T. U.: Gesangsaussprache-Profil nach Julius Hey. Schriften der Richard-Wagner-Akademie 1. Dresden, 2024.

[17] HEY, J.: Der kleine Hey. Die Kunst des Sprechens. Mainz u. a. 2006. (52. Auflage).

[18] WAGNER, R.: Oper und Drama. In: Richard Wagner: Werke, Schriften und Dichtungen, S. 1979 (vgl. Wagner-SuD Bd. 4, S. 229). http://www.digitale-bibliothek.de/band107.htm

# "SHOUT OUT THE WINDOW" - ACOUSTIC AND ELECTROGLOTTOGRAPHIC ANALYSIS OF SHOUTING VOICE TRAININGS

*Lara Höffner*[1], *Anna Wessel*[1], *Sven Grawunder*[1,2]

[1]*Martin Luther University Halle-Wittenberg, Germany,* [2]*Max Planck Institute for evolutionary Anthropology, Leipzig, Germany*

*hoeffner@student.uni-halle.de, {wessel, grawunder}@sprechwiss.uni-halle.de*

**Abstract:** This small-scale trial aims to make a start on improving the realism of acoustic and electrophysiological shouting voice measurements by incorporating more realistic practice scenarios and improving the practicability of the measurement setting. Aiming for a reference corpus, the results reflect expected trends for loud shouting voice but also shed light on ways of improvement for data acquisition during voice training sessions.

## 1 Shouting

Shouting is part of everyday life for many speech-intensive professions such as actors, singers, teachers, sales persons, and sports coaches, so that these make out a good portion of patients seeking voice treatment. Nevertheless, there is little research into this voice practice, let alone the voice training practice; hence specific measurements need to be adjusted for such extreme vocal performances. Therefore, the objective of this study is at least twofold, namely aiming at data of shouting voices in realistic training sessions and exploring robust parameters in acoustic and electroglottographic signal analysis. In this way, it can be examined what kind of experimental set-up and analysis pipeline can be effectively used for generating a shouting voice reference corpus.

The term shouting voice and some synonyms are described differently in the literature following different perspectives (therapy, training, coaching, performance). Lagier et al. [1, p. 141ff.] differentiate shouting as a specific form of loud voice. In both forms, the subglottic pressure increases, but shouting is defined by a sudden onset of vocal production. "A shout is a sudden loud outburst, a very specific form of communication motivated by an emotional or situational context. [...] Whatever the motivation behind a shout, it differs greatly in character from the speaker's conversational voice" [1, p. 141]. Hacki [2, p. 814] defines the shouting voice as the "maximum frequency-related and intensity-related limiting power of the chest register" , with 90 dB given as the standard value for the lower intensity limit. This value is also described by others [3, p. 114] [4, p. 81]. Although those definitions differentiate between modal and shouting voice, the voice quality and the physiological processes of the shouting voice are usually neglected. More recent definitions refrain from specifying a lower intensity limit and focus more on the communicative intention or the production method of the shouting voice. Like Jokinen [5, p. 1]: "In shouting, speakers use increased vocal effort to convey spoken messages over distance or above environmental noise". Lemke [6, p. 67] describes the phenomenon in its ideal and goal for teaching as *the maximal vocal performance with the least effort and the lowest tension of the laryngeal muscles*. Aderhold & Wolf [7, p. 31] specify this *least effort* as *without additional or excessive tension*. Refering to those definitions, Häußler [8, p. 23] provides a definition that takes into account the aspects of intensity, physiology, voice quality as well as the goal. She defines shouting voice as *a physiological development of the speaking*

*voice with the aim of achieving the maximum vocal intensity and range with the least possible effort and free from over-tension*. In contrast, modal voice is a physiological voice phenomenon that is used in the majority of utterances in communication situations [9, p. 176].

# 2 Method

## 2.1 Participants

The participants were 13 speech science students from Martin-Luther-University Halle- Wittenberg (MLU) who volunteered to take part in the study. The age range was between 22 and 33 years, with the average age of 26. The inclusion criteria were at least four semesters of speech training in the context of the speech science degree programme at MLU and no severe vocal restrictions due to chronic or current illnesses, although mild cold symptoms were tolerated. It should be noted that all of the test subjects were fellow students of the first author. In terms of gender distribution, the group of participants was heterogeneous: seven female, five male, and one diverse.

## 2.2 Procedure

Measurements were conducted using a calibrated sound level meter (SLM, IEC 651 Typ 2), a microphone (BEHRINGER ECM8000), placed at a controlled distance of 30 cm and an electroglottograph (EGG, ElectroGlottograh EG-2 by GLOTTAL.COM) [cf. 10]. As the speech signal is recorded via the microphone and the SLM at a distance of 30 cm, but the EGG signal is recorded directly at the larynx, there is a time delay between the two signals, which can be approximately 12 - 15 ms [11, p. 1174ff.]. This time delay can cover an entire period and is taken into account in the evaluation of the results; exact temporal synchronisation was not carried out. To ensure consistent positioning of the electrodes at the level of the larynx, a coloured marker was placed on the subject's neck to illustrate any slippage of the collar and to readjust it if necessary.

The task included a text sequence in modal (reading) voice and in three shouted sequences, with voice exercises in between: one exercise with the goal to relax the voice and one shouting voice exercise. As speech production generally relates to meaningful utterances with a communicative intention, it was not considered reasonable to have syllables phonated in isolated cases. For this reason the syllable onsets are later isolated from the words in order to analyse the results in detail [12, p. 33f.]. The text was created by means of short utterances of suitable context, and a selection of different syllabic onsets ([h-, v-, ʔ-]) in combination with German vowel phonemes. Here, only the vocal /a/ is analysed. The data collection was preceded by a pre-test.

The task bouts of the session recording were segmented in a PRAAT [13] TextGrid. Subsequently, the WEBMAUS Basic service, version 3.14 [14] was applied for forced alignment of the training text. The initial vowels, like for /a/ from the words <Arbeit>, <alles>, <Alma>, <an> and <also>, were targeted using a customised PRAATscript and analysed for intensity, fundamental frequency (f0) and contact area of the vocal folds using the quasi contact quotient (QCQ) [cf. 15], closing quotient and band energy difference (BED, 0 to 2kHz vs 2 to 4kHz, [16]).

## 2.3 Parameters

The phenomenon of the shouting voice differs from the modal voice in some parameters. The modal voice tends to be associated with everyday speech at a short distance from one or more

communication partners. The shouting voice tends to be associated with greater distance, increased ambient noise, emotional outbursts such as anger or joy, and distress signals [17, 18].

In order to determine the difference between shouting and modal voice in a realistic speech training setting, four parameters have been described in the literature as parameters of the shouting voice: *intensity (Int), fundamental frequency* (f0), *quasi-contact quotient* (QCQ) and *band energy difference* (BED).

The signal intensity, which manifests itself as loudness, is probably the most prominent characteristic of the shouted voice. An increase in signal intensity is also associated with a high vocal effort [5]. Age can be an influencing factor on intensity [19, p.14]. The intensity is calculated by subtracting the reference value of the intensity in dB of the calibration section recorded with the microphone (refIntdB) from the intensity of the vowel section in the audio signal (intdB). The calibration intensity (calibrInt) is added to this: $Intensity = calibrInt + (IntdB - refIntdB)$

Shouting is often associated with an increase in f0, i.e. the f0 decreases slightly with an increase in intensity and becomes significantly higher with the use of the shouted voice [cf. 3, p. 109]. It has been also described that an increase in intensity often increases the f0, though this can have negative consequences for larynx, vocal folds and vocal tract [6, p. 54ff.]. f0 is calculated using the duration of the glottal period ($T$) in seconds, using $f0 = 1/T$.

While the vocal folds in the modal voice ideally vibrate in a relaxed and regular manner, the vocal folds in shouting are very tense and vibrate with an increased amplitude, which is also reflected in the increasing amplitude of the epithelial movement. In addition, the vocal folds change their length and mass [20, p. 38 f.] [19, p. 14]. The modal voice has a prolonged opening phase of the glottis in relation to the closing phase. In shouting, this relationship is reversed and the closing phase is prolonged in relation to the opening phase, which means that the glottis is only open briefly during loud phonation [20, p. 39] [21, p. 44] [17, p. 3056]. Therefore QCQ is derived from the EGG signal and calculated by means of the duration of the contact phase divided by the total time of the glottal cycle.

The configuration of the vocal tract has an influence on the loudness of a voice. By amplifying the resonance chamber in the vocal tract, certain frequency components can be inhibited or amplified, which means that the setting of the vocal tract has a specific influence on the position of the formants and thus on the distinction between modal and shouting voice [20, p. 39] [22, p. 28]. The position of the vocal tract not only influences the perceptual intelligibility of the shouting voice in relation to the vowel formants, but also the characterisation of the singer's formant, which can also be associated with the speaker's formant [cf. 16, p. 555]. There is disagreement in literature about the exact position of the singer's formant, but there is agreement that it is located above the vowel formants (F1-F2) and approximately between 2.5-3.5 kHz. Titze & Story [23, p. 2234ff.] describe the singer's formant as a cluster of the 3rd, 4th and 5th formants. This formant is associated with characteristics such as radiance and carrying capacity [24, p. 84]. In addition, an increasing spectral energy in the region of the singer's formant as an effect of increasing loudness is described [16, p. 556]. This can be explained by the fact that the epilarynx can be responsible for the formant positions from the 3rd formant onwards [23, p. 2234ff.] and that the larynx has a tendency to move upwards when shouting [25, p. 3439]. Hence, the band energy difference (BED) between 0-2 kHz and 2-4 kHz is calculated to analyse the change in the singer's formant.

## 3 Results

For the data presented here, the vowels /a, e, i/ were isolated from the read and shouted text sequences (n=495) and analysed for the different acoustic parameters. As expected, we find

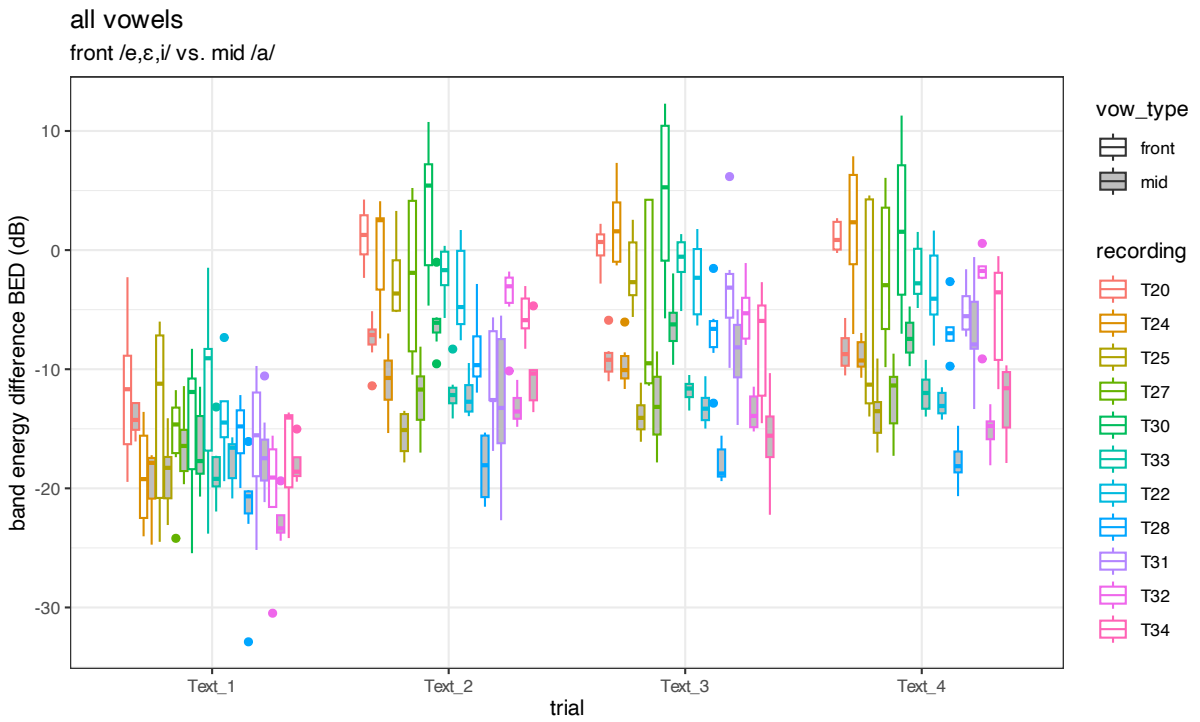



**Figure 1** – band energy differences between 0-2 kHz and 2-4 kHz with boxplots for front vs. mid vowels for each participant

for the mid /a/ vowel significant differences between the modal and shouting voice in all the analysed parameters on a group level and individually. On a group level, the results are largely consistent with previous research, with respect to differences between shouted and modal voice. In the shouted voice, which are all trials except for the first (Text_1) the intensity, f0 and QCQ increased statistically significantly ($\alpha = 0.05$), which has been tested respectively by means of ANOVAs. This indicates an increase in loudness (intensity), an increased average speaking pitch and a prolonged contact phase of the vocal folds. The prolonged contact phase indicates a prolonged closing phase of the vocal folds within a glottal cycle. The BED decreases statistically significantly in the shouted voice compared to the modal voice (s. Figure 1), which indicates a more pronounced singer's formant. However, we see lower BED values especially for the mid open vowel /a/.

Next to the expected higher intensity in the more open /a/ vowels other differences to the front vowels /e i/ are mainly seen in the values for BED. On an individual level, we see e.g. for QCQ, that individual trajectories reveal specific strategies and perhaps also challenges. So is there for participant T24 a larger difference between speaking (Text_1) and shouting, especially with respect to the last trial (Text_4). In addition, this seems to be dependent on the vowel (s. Figure 2), with a drop for vowels /i, e/ in trial 4.

## 4 Discussion

The text created for the experimental setup consists of meaningful utterances and does not require the production of isolated sentences, syllables or vowels. It was possible to shout with a *gestus* (after Bertold Brecht) and situational reference. However, it should be noted that the *gestus* and situational reference were relatively unspecific and should be further developed in similar experimental designs in the sense of more distance and directedness. The homogeneity in the results of the statistical analysis supports the hypothesis that the shouted voice can also be investigated in a specific setting using speech training methods. Although attention was paid to the proximity to speech training practice during the preparation, the embodiment, i.e.

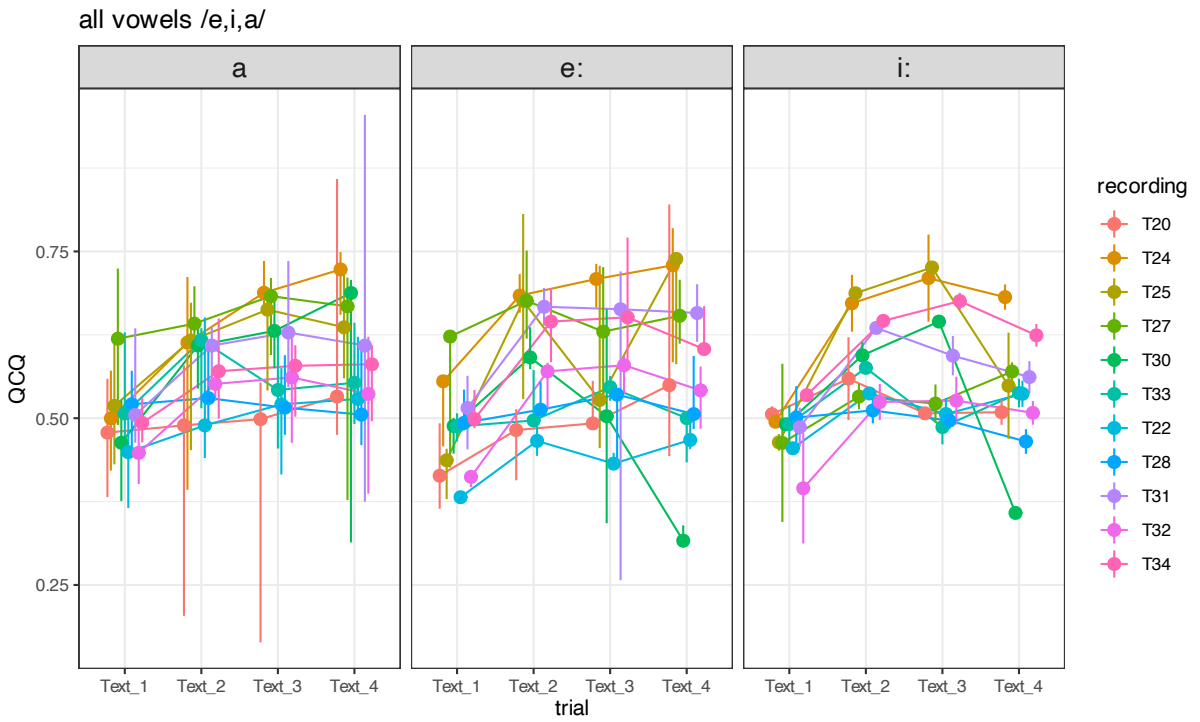



**Figure 2** – QCQ values in vowels for individual trials and participants

inter alia the entire involvement of posture and movement, could not be taken into account. In further studies, this aspect could be examined in more detail, e.g. with video recordings and the usage of body-related exercises. The aspect of comprehensibility is also of great importance in a speech-artistic communication situation and could be taken into account in a further analysis of the data.

There is a discrepancy between the analysis of the parameters carried out in this study and the phonetically specialised literature and the literature on speech training. The phonetically specialised literature describes that the shouted voice is associated with an increased f0 [26, p. 41] [18, p. 4]. In the speech training literature, the optimal pitch should also be used as a guide when shouting, even if the shouted voice is often raised [6, p. 54ff.]. The optimal pitch refers to a pitch range in which the voice 'works economically', i.e. with relatively low muscle potential and a high radiation effect. It is located in the lower range of the speaking voice pitch [27, p. 24]. At 224.8 Hz, the average f0 of the male subjects in this study is significantly higher than the norm range given e.g. in the manual by Nawka & Wirth [4, p. 78 f.] of 87/98 Hz to 131 Hz. Likewise, the average f0 of the female subjects, at 335.9 Hz, is also well above the norm of 175/196 Hz to 262 Hz. As intelligibility also decreases with an increase of f0 in shouting and negative consequences for the larynx, vocal folds and vocal tract are also described [6, p. 54ff.] [28, p. 11], the question arises as to what increase is still physiological and defines the shouting voice and at what point it is a questionable phonation tendency. Further scientific research is needed to determine whether the shouted voice should be based exactly on the optimal pitch or which f0 increase is most effective and can still provide a natural and nuanced expression.

With regard to the measurement technique, it should be noted that the upward vertical laryngeal movements during shouting were stronger than assumed in the design of the experimental set-up. As a result, the collar and hence the EGG electrodes would lose their position at the level of the thyroid cartilage. The vertical elevation of the larynx during shouting has been also described before [cf. e.g. 25].

The experimental set-up involves repeated measurements. The individual measurement processes will have influenced the subsequent repetitions. An attempt was made to reduce

this transfer effect by using exercises in between. However, quantifying this effect is beyond the scope of this paper. Therefore, it can only be speculated how large it is. However, as the statistical analyses of variance did not show any significant differences between the mean values of the individual shouted voice parameters between the shouted trials, this effect is considered tolerable.

The implementation of the experiments encountered various challenges that influenced the validity of the results. The use of the EGG proved to be problematic due to the strong vertical movement of the larynx with regard to the positioning of the electrodes. The questionnaire for acquiring the personal data was not designed unambiguously enough. For example, there should have been more precise definitions of gender or speaking and singing experience.

The preparation of the material for the measurements must also be critically examined. The *gestus* of the text was ambiguous. A situation with more width and distance should have been specified. In addition, a major aspect that is highly relevant, especially from a speech training perspective, is a detailed investigation of the relationship between the vocal function of the shouted voice and shouted voice exercises. In this way, the improvement and training of a well-functioning shouted voice can also be described using phonetic parameters and this knowledge could have a transfer into practice. The use of the experimental set-up presented in this study could, for example, be supplemented with individual feedback and/or biofeedback and carried out with a different group of test subjects and a control group. A useful target group could be teachers, who need to be able to use their shouted voice effectively in their everyday work. It would also be interesting to look at the effect of phrase length on the effectiveness of a shouted voice. Finally, the individually identifiable parameters could be analysed qualitatively. For example, which parameters are involved if an increase in the intensity of the shouted voice is not possible or only possible to a reduced extent. With this knowledge, individually customised and needs-oriented shouting exercises can be selected and applied.

## 5 Conclusion

The experiment design enables both individual and group-level analysis of the shouting voice, allowing for acoustic and EGG recordings in a setting that closely mirrors speech training practice. The collected data can be used for further research, such as studying the influence of shouting voice exercises on vocal function, examining the impact of phonetic context on vocal function, and as a reference corpus.

The test set-up is similar to speech training practice, which offers valuable new insights into this working area and has not been carried out in this way before. The parameters intensity, f0, QCQ and BED, which differ between the modal voice and the shouted voice, could be described.

However, currently it remains unanswered to what extent the measurements described above can constitute an effective shouted voice, what correlations occur within the parameters, to what extent the shouted voice should be orientated towards the so-called indifference (tone) level and how shouted voice exercises influence the shouted voice function. The experimental setup presented can be seen as preparation for investigating the influence of shouted voice exercises on the vocal function of the shouted voice and also offers further scope for investigating the influence of the phonetic context on the shouted voice. The generated data can be used as a basis for a reference corpus for further shouting voice analyses in a speech training context.

The collected data provides more material than could be analysed in this study. For example, the other voicing onsets such as /h/ and /v/ as the quality of the glottal onsets themselves could be analysed. Further, there is a potential effect of the in-between exercises on the shouted

voice. And, of course, at the perceptual level, e.g. a comprehensibility test, would also be worth considering. In addition, associations with qualitative measures and individual observations should be carried out. Further studies could also look at the extent to which the parameters described here actually serve as an indication of an effective shouted voice. In this context, the consideration of the EGG signals with regard to the degree of vocal fold collision via the closure quotient could be of interest [cf. 15, 29].

## Acknowledgements

We are grateful to all students of the dept. of speech science and phonetics for their participation and for helpful comments to an earlier version of this research at PEVOC (Santander 2024).

## References

[1] LAGIER, A., T. LEGOU, C. GALANT, B. AMY DE LA BRETÈQUE, Y. MEYNADIER, and A. GIOVANNI: *The shouted voice: A pilot study of laryngeal physiology under extreme aerodynamic pressure*. *Logopedics Phoniatrics Vocology*, 42(4), pp. 141–145, 2017. doi:10.1080/14015439.2016.1211735.

[2] HACKI, T.: *Tonhöhen- und Intensitätsbefunde bei Stimmgeübten*. *HNO*, 47(9), pp. 809–815, 1999. doi:10.1007/s001060050464.

[3] SCHNEIDER-STICKLER, B. and W. BIGENZAHN: *Stimmdiagnostik: Ein Leitfaden für die Praxis*. Springer Vienna, 2013. doi:10.1007/978-3-7091-1480-3.

[4] NAWKA, T. and G. WIRTH: *Stimmstörungen: für Ärzte, Logopäden, Sprachheilpädagogen und Sprechwissenschaftler*. Deutscher Ärzte-Verlag, 5th edn., 2008.

[5] JOKINEN, E., R. SAEIDI, T. KINNUNEN, and P. ALKU: *Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task*. *Computer Speech & Language*, 53, pp. 1–11, 2019. doi:10.1016/j.csl.2018.06.002. Publisher: Elsevier.

[6] LEMKE, S. (ed.): *Sprechwissenschaft, Sprecherziehung : ein Lehr- und Übungsbuch. Unter Mitarb. von Philine Knorpp*, vol. 4 of *Leipziger Skripten*. Lang, 2nd edn., 2012.

[7] ADERHOLD, E. and E. WOLF: *Sprecherzieherisches Übungsbuch*. Henschel, 18th edn., 2018.

[8] HÄUSSLER, R.: *Die Rufstimme. Konzept, physiologische Grundlagen und praktische Anwendung.*, 2022. Unpublished BA thesis, Martin-Luther-Universität Halle-Wittenberg.

[9] ZRAICK, R. I., W. MARSHALL, L. SMITH-OLINDE, and J. C. MONTAGUE: *The effect of task on determination of habitual loudness*. *Journal of Voice*, 18(2), pp. 176–182, 2004-06-01. doi:10.1016/j.jvoice.2003.09.005.

[10] ŠVEC, J. G. and S. GRANQVIST: *Tutorial and Guidelines on Measurement of Sound Pressure Level in Voice and Speech*. *Journal of Speech, Language, and Hearing Research*, 61(3), pp. 441–461, 2018. doi:10.1044/2017_JSLHR-S-17-0095.

[11] HÉZARD, T., V. FRÉOUR, R. CAUSSÉ, T. HÉLIE, and G. P. SCAVONE: *Synchronous multimodal measurements on lips and glottis: Comparison between two human-valve oscillating systems*. *Acta Acustica united with Acustica*, 100(6), pp. 1172–1185, 2014. doi:10.3813/AAA.918796.

[12] ANDERS, L. C.: *Spektrale Analysen des Stimmklangs*. In *Sprechen als soziales Handeln. Hallische Schriften zur Sprechwissenschaft und Phonetik*, vol. 2, pp. 30–39. Verlag Werner, 1997.

[13] BOERSMA, P. and D. WEENINK: *Praat: Doing phonetics by computer [Computer Program]*. 2024. URL `https://www.praat.org`.

[14] KISLER, T., U. D. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 45, pp. 326–347, 2017-09. doi:10.1016/j.csl.2017.01.005.

[15] HERBST, C. T.: *Electroglottography – an update*. *Journal of Voice*, 34(4), pp. 503–526, 2019. doi:10.1016/j.jvoice.2018.12.014.

[16] BELE, I. V.: *The speaker's formant*. *Journal of Voice*, 20(4), pp. 555–578, 2006. doi:10.1016/j.jvoice.2005.07.001.

[17] MITTAL, V. K. and B. YEGNANARAYANA: *Effect of glottal dynamics in the production of shouted speech*. *The Journal of the Acoustical Society of America*, 133(5), pp. 3050–3061, 2013. doi:10.1121/1.4796110.

[18] MITTAL, V. K. and A. K. VUPPALA: *Significance of automatic detection of vowel regions for automatic shout detection in continuous speech*. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5. 2016. doi:10.1109/ISCSLP.2016.7918393.

[19] KIESE-HIMMEL, C.: *Körperinstrument Stimme*. Springer Berlin Heidelberg, 2016. doi:10.1007/978-3-662-49648-0.

[20] FÖCKING, W. and M. PARRINO: *Praxis der Funktionalen Stimmtherapie*. Springer, 2015.

[21] HACKI, T.: *Klassifizierung von Glottisdysfunktionen mit Hilfe der Elektroglottographie*. *Folia Phoniatrica et Logopaedica*, 41(1), pp. 43–48, 1989. doi:10.1159/000265931.

[22] HAMMER, S. S. and A. TEUFEL-DIETRICH: *Stimmtherapie mit Erwachsenen : was Stimmtherapeuten wissen sollten*. Praxiswissen Logopädie. Springer, 6th edn., 2017.

[23] TITZE, I. R. and B. H. STORY: *Acoustic interactions of the voice source with the lower vocal tract*. *The Journal of the Acoustical Society of America*, 101(4), pp. 2234–2243, 1997. doi:10.1121/1.418246.

[24] SCHUTTE, H. K. and W. SEIDNER: *Physiologische grundlagen*. In J. WENDLER, W. SEIDNER, and U. EYSHOLDT (eds.), *Lehrbuch der Phoniartie und Pädaudiologie*, pp. 71–90. Thieme, 5th edn., 2015.

[25] TRAUNMÜLLER, H. and A. ERIKSSON: *Acoustic effects of variation in vocal effort by men, women, and children*. *The Journal of the Acoustical Society of America*, 107(6), pp. 3438–3451, 2000-06-01. doi:10.1121/1.429414.

[26] DIMOS, K., L. HE, and V. DELLWO: *Shouting affects temporal properties of the speech amplitude envelope*. *JASA Express Letters*, 4(1), p. 015202, 2024. doi:10.1121/10.0023995.

[27] NEBERT, A. U.: *Der Tonhöhenumfang der deutschen und russischen Sprechstimme*. Peter Lang, 2013.

[28] XUE, Y., M. MARXEN, M. AKAGI, and P. BIRKHOLZ: *Acoustic and articulatory analysis and synthesis of shouted vowels*. *Computer Speech & Language*, 66, p. 101156, 2021. doi:10.1016/j.csl.2020.101156.

[29] SERRY, M. A., C. E. STEPP, and S. D. PETERSON: *Exploring the mechanics of fundamental frequency variation during phonation onset*. *Biomechanics and Modeling in Mechanobiology*, 22(1), pp. 339–356, 2023. doi:10.1007/s10237-022-01652-8.

# THE INFLUENCE OF TEXT STYLE TRANSITIONS ON RHYTHM PATTERNS IN FIXED-ORDER READING TASKS

*Neda Mousavi*[1], *Olga Fikiel*[1], *Baldur Neuber*[1], *Sven Grawunder*[1,2]

[1]*Martin Luther University Halle-Wittenberg, Germany,* [2]*Max Planck Institute for evolutionary Anthropology, Leipzig, Germany*

*neda.mousavi@sprechwiss.uni-halle.de, sven.grawunder@sprechwiss.uni-halle.de*

**Abstract:** In this study we conducted an experiment in which participants were asked to read texts in different styles that were presented in a fixed order for 12 German participants. We used five different texts representing different styles - descriptive, narrative, conversational, reflective, and instructional - to examine the effects of style on rhythmic speech patterns. In order to explore the hierarchical structure of our data, where multiple readings (texts) are nested within participants, we apply a mixed-effects model; however, regarding the large number of metrics to measure rhythm, first we apply a dimensionality reduction model, principal component analysis (PCA), to reduce the number of response variables in our model. Based on the results of the mixed-effects models for PC1 and PC2, we observe significant influences of the text style on the rhythmic properties of speech. For PC1, tasks t2, t3, t4 and t5 all show significant positive effects compared to the baseline task t1. For PC2, tasks t2, t3 and t5 show significant negative effects, while task t4 has no significant effect. The variability between speakers is also present, but more pronounced in PC1 than in PC2.

## 1 Introduction

This study is based on creating a German read speech corpus collected in a controlled laboratory environment where participants were asked to read texts of different styles, ranged from descriptive and narrative to conversational, reflective and instructional. The research follows two main perspectives. The first examines the influence of text style on verbal patterns, with a particular focus on rhythmic features. Following previous literature (e.g. [1]), this study hypothesizes that the stylistic differences in texts significantly influence the rhythmic features of the resulting speech. The second perspective focuses on the variability of speaker behavior within the same task setting considering differences in speech realization that are influenced by the task material and individual speaker characteristics. While at first view the setting may seem like a straightforward speech elicitation task for recording read speech, the different styles of the texts introduce an additional layer of complexity. This complexity drives speakers to produce varying realizations of the same reading task, reflecting differences in their linguistic behavior. It is important to emphasize that although laboratory speech is sometimes criticized for its controlled nature, it is still a widely accepted and practical method for studying natural speech because it is inherently produced by humans and reflects natural behavior [2]. Therefore, in addition to text style as a factor influencing the speech produced in reading tasks, speakers are expected to show different levels of proficiency and individual characteristics in their reading performance.

Another factor that can influence the results of such studies is the order in which the texts are presented to the speakers. Previous research on experimental design (e.g. [3]) has shown that

the order of tasks can significantly influence participants' performance. In varying-sequence trials, where the order of tasks differs from participant to participant, carry-over effects can be investigated by analyzing how the influence of one task persists and affects subsequent tasks. Statistical methods, such as mixed-effects models, are commonly used to investigate these effects[4]. Conversely, researchers can use a fixed sequence to investigate how participants adapt to or master tasks over time. In this study, we used a fixed sequence of tasks that allows us to examine how participants adapt to task material and reading texts of different styles over time and to analyze performance trends in learning, adaptation, and potential fatigue during the sequence.

## 2 Data and Method

The dataset contains recordings of the voice of 12 participants, 3 males (Sp01, Sp09, Sp12) and 9 females, reading five different texts, shown in figure 1, each representing a different style. The first text is a narrative in which past events are reported in formal and descriptive language. The second text contains a personal letter in a conversational and emotive style. The third text is a traditional fairy tale with a narrative structure and imaginative language. The fourth text is based on metaphorical and poetic language about love. Finally, the fifth text is a recipe with direct and clear instructions. The following table shows the texts, with a symbol in the first column representing each text style used later in the plots.

| | |
|---|---|
| Text 1 | Christian August besaß ein Haus in Dornburg. Das kleine verträumte Dorf, malerisch in den Elbauen südlich von Magdeburg gelegen, war jedoch auch in geografischer Hinsicht das letzte Anhalts. Also musste er, um Haus, Hof und Familie zu unterhalten, sich auswärts um Arbeit bemühen. Er wurde, wie für Personen seines Standes üblich, Offizier. Als Garnisonskommandant von Stettin durfte er immerhin ein ordentliches Haus und später gar das Schloss bewohnen. |
| Text 2 | Liebe Christa, es regnet. Es ist windig, auf dem Fensterbrett rascheln gelbe Blätter, und ich bin traurig alles, wie es sich für den November schickt. Abends halte ich Dir lange Reden Du weisst schon diese Kopfkissen Monologe, aber mit dem Briefeschreiben will es nicht recht gehen, und für den Fall, dass ich heute auch wieder steckenbleibe, will ich Dir erst mal schönsten Dank sagen für die Seghers Essays. |
| Text 3 | Es war einmal ein König, der hatte einen Sohn, der warb um die Tochter eines mächtigen Königs, die hieß Jungfrau Maleen und war wunderschön. Weil ihr Vater sie einem andern geben wollte, so ward sie ihm versagt. Da sich aber beide vom Herzen liebten, so wollten sie nicht voneinander lassen, und die Jungfrau Maleen sprach zu ihrem Vater: Ich kann und will keinen anderen zu meinem Gemahl nehmen. Da geriet der Vater in Zorn und ließ einen finstern Turm bauen, in den kein Strahl von Sonne oder Mond fiel. Als er fertig war, sprach er: Darin sollst du sieben Jahre lang sitzen, dann will ich kommen und sehen, ob dein trotziger Sinn gebrochen ist. Für die sieben Jahre ward Speise und Trank in den Turm getragen, dann ward sie und ihre Kammerjungfer hineingeführt und eingemauert und also von Himmel und Erde geschieden. |
| Text 4 | Die Wege der Liebe sind selten die kürzesten Verbindungen zwischen zwei Punkten: die gehen hin und her, kreuz und quer, winden sich als Serpentinen, drehen sich wie Spiralen, führen über Berg und Tal, durch Lust und Qual, scheinen Labyrinthe besonders dem, der sie geht oder kriecht oder fliegt. |
| Text 5 | Quark zu Kartoffeln oder als Brotaufstrich: fünfhundert Gramm Quark, Milch, Salz, Paprika, Kümmel. Den Quark mit so viel Milch verrühren, dass er geschmeidig wird. Durch ein Sieb streichen oder elektrisch rühren. Mit Salz, Paprika und Kümmel abschmecken. Zum Anrühren kann Sahne oder außer der Milch, wenig Öl verwendet werden. Als Würze eignen sich auch Kräuter, Zwiebel, Meerrettich, Rettich, Tomaten und Paprikamark. |

**Figure 1** – Texts in different styles

Automatic segmentation of the recordings was performed using the WebMAUS system [5], which generated Praat TextGrids [6]. A Praat script was then developed to add a new tier to the TextGrids that labeled segments as "v" for vowels and "c" for consonants. The script then removed consonant boundaries, creating vowel and consonant intervals that served as the primary

units for rhythm measurement. The quantification of speech rhythm involved measuring eight different rhythmic metrics: percentage of vocalic intervals and standard deviation of consonant intervals [7], nPVI for vocalic and rPVI for consonant intervals [8], VarcoC and CV-rate [9], and the mean values of consonant and vocalic intervals. At this stage, the selection of appropriate rhythmic metrics posed a major methodological challenge given the great variability in the literature. Since our aim is to analyze the effects of text style on rhythmic patterns, a crucial question arises: how do we determine the most appropriate metrics for this study? Should we include as many metrics as possible, or would it be more effective to focus on a specific set based on a theoretical framework for defining rhythms? This question is further complicated by the fact that we want to use a mixed-effects model to account for the hierarchical structure of the dataset, in which texts are nested in speakers. In such a model, each rhythmic metric would have to be analyzed separately, necessitating the running of multiple models. This raises concerns about the efficiency of the analysis and the potential for statistical problems, such as multicollinearity and an increased risk of type I errors. In order to overcome such challenges, we chose an approach in which we used principal component analysis (PCA) to reduce the dimensions of our metrics and lower the number of response variables before applying the mixed-effects model. In this way, we summarize the most important information from all metrics into a smaller set of uncorrelated components, simplifying the analysis while preserving the essential data variability. This approach of using PCs as response variables in a mixed-effects model has been successfully applied in other areas of research, for example in the work of [10]. Importantly, while PCA reduces the number of variables by combining rhythmic metrics into principal components, the contribution of each individual metric to these components can still be analyzed. This means that even if the analysis focuses on the principal components, we can still interpret how specific rhythmic metrics influence the overall patterns and outcomes, providing insight into the role of each individual metric.

## 3 Results

After applying PCA to reduce the dimensionality of the rhythmic metrics, we focused on the first two principal components (PC1 and PC2) to investigate the variations between text styles and speakers. Based on the loading plot extracted from PCs, PC1 is primarily determined by consonant-related metrics, with the standard deviation of consonant intervals (sdC) and the raw PVI of consonants (rPVI-C) showing the strongest positive contributions. In addition, the mean consonant (mean-C) and vowel duration (mean-V) play a significant role, while CV-rate and the percentage of vowel intervals (percent-V) make a negative contribution, indicating an inverse relationship. PC2 is dominated by vowel-related features, with nPVI-V having the largest positive influence, followed by percent-V and mean-V. In contrast, VarcoC shows a negative contribution, indicating that PC2 captures vowel rhythmicity, with an inverse relationship to consonantal variability. The plot below shows the distribution of tasks based on the principal components, with centroids (average positions) representing the rhythmic profile of each task across different speakers. These centroids highlight the core tendencies of each text style, and since the tasks are known, we can treat the centroids as predefined clusters within the PCA space.

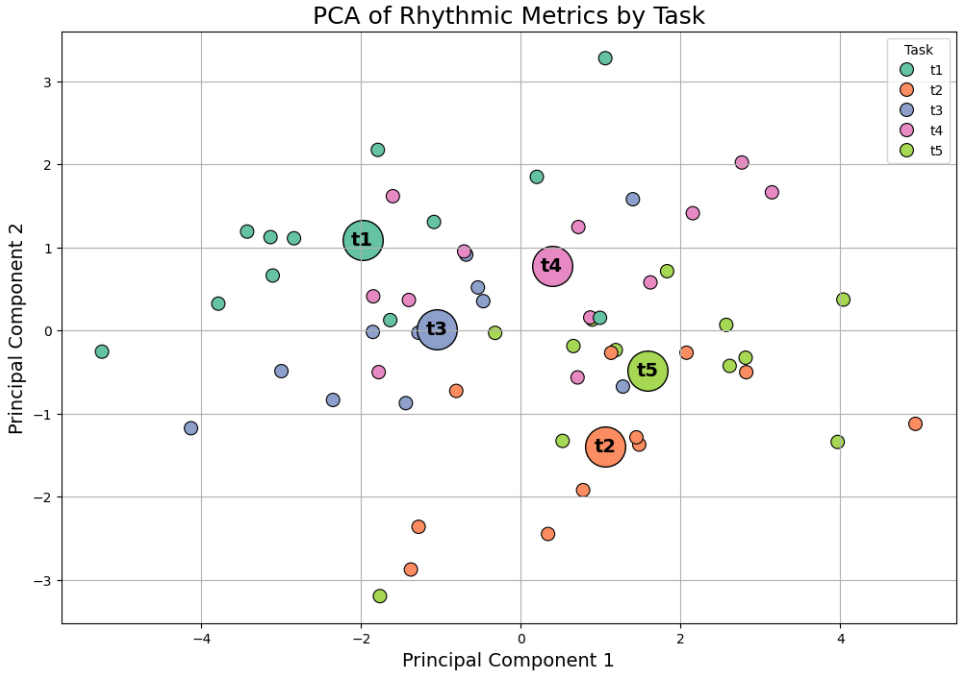



**Figure 2** – Distribution of Tasks and Centroids in PCA Space

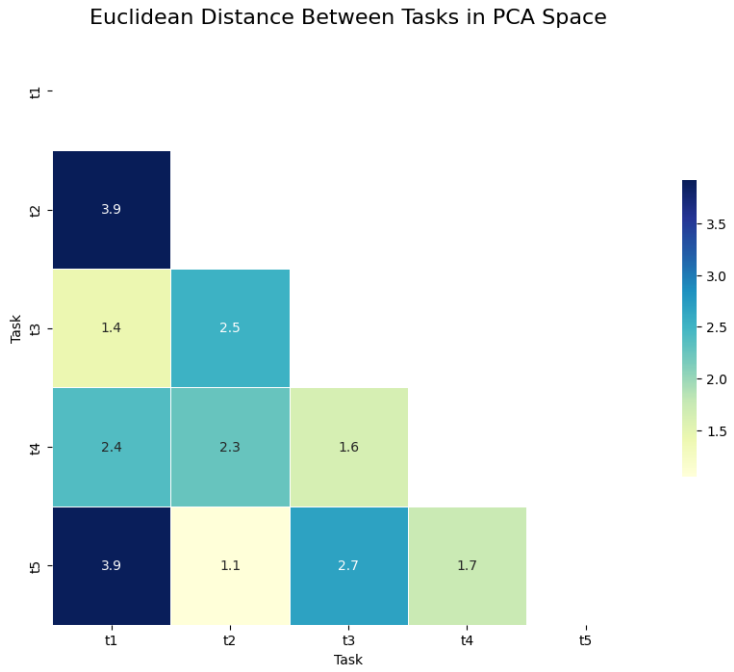



**Figure 3** – The Euclidean distance between tasks in PCA Space

In the next step, we evaluated how well the tasks were clustered and separated in PCA space using two metrics: Euclidean distance and silhouette value. Euclidean distance measures how far apart the tasks are and gives us a sense of how distinct they are in PCA space. The silhouette value is an additional measure that indicates how well the tasks are separated and provides insight into the overall effectiveness of the clustering. The Euclidean distance heatmap in figure (1) highlights the differences between the tasks based on the first two main components (PC1 and PC2). Larger distances, such as 3.9 between Task 1 (historical narrative) and Task 5 (recipe), indicate different rhythmic structures likely due to their different styles - formal and

descriptive versus directive and precise. In contrast, smaller gaps, such as 1.05 between Task 2 (personal letter) and Task 5, reflect greater similarity, possibly due to a common conversational tempo or simpler sentence structures. However, the silhouette score for tasks (-0.04) indicates significant overlap in rhythmic patterns between tasks, suggesting that the separation between them in the PCA space is not strong. The same procedure was applied to the speakers. The negative Silhouette Score of -0.2239 indicates a considerable overlap in rhythmic patterns between speakers, suggesting that the PCA space does not effectively distinguish between them. Comparatively, rhythmic patterns appear to be more similar across speakers than across tasks, though both exhibit significant overlap. This suggests that while text style may have a stronger influence on rhythm than speaker-specific characteristics, it is still not substantial enough to produce clear separations in the PCA space.

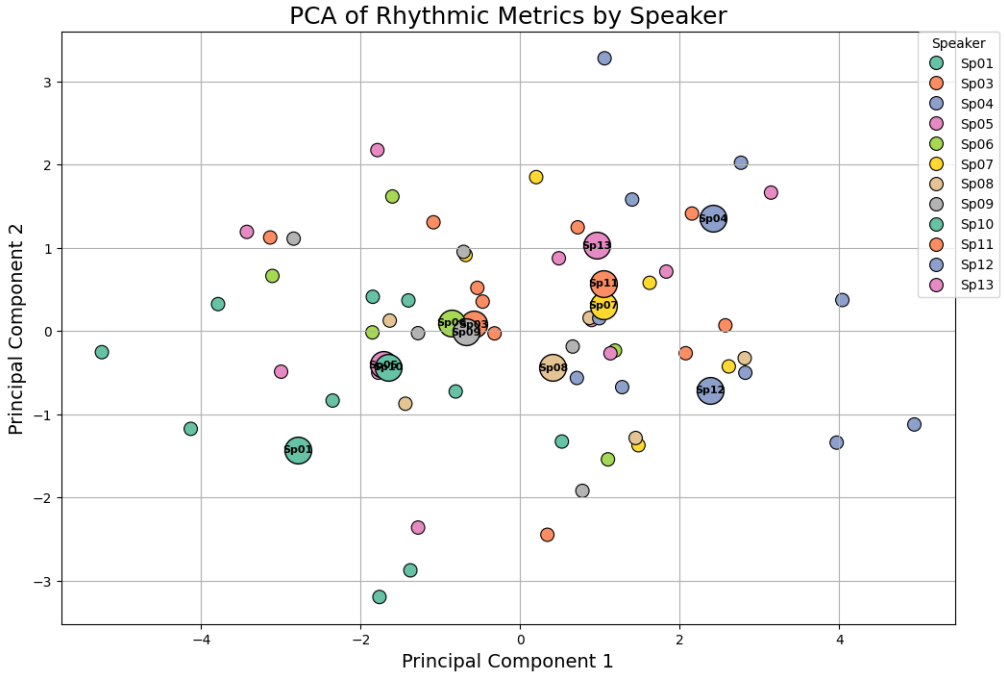



**Figure 4** – Distribution of speakers and Centroids (regarding different tasks per each one) in PCA Space

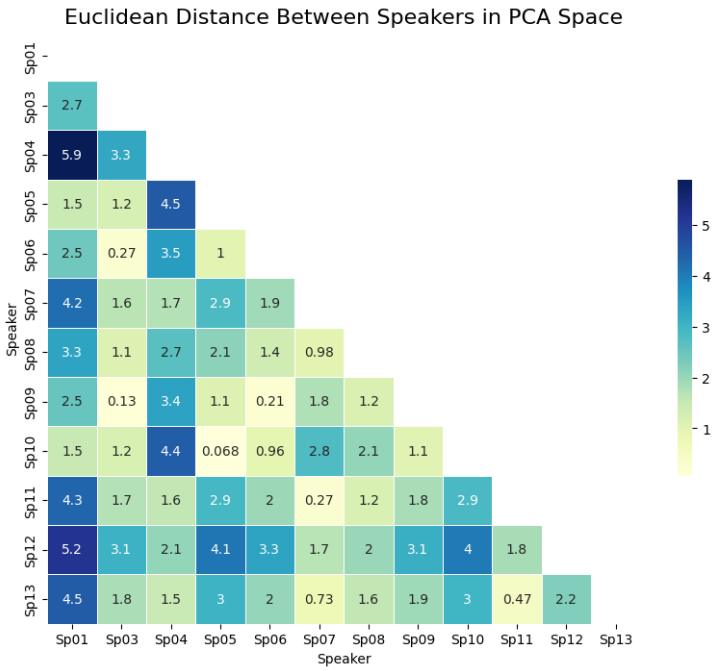



**Figure 5** – The Euclidean distance between speakers in PCA Space

Although PCA provided an efficient method for reducing the dimensionality of the rhythmic metrics, the resulting silhouette scores for both task (-0.04) and speaker (-0.2239) indicate overlap in clusters. This suggests that the principal components of rhythmic metrics did not capture enough variation to clearly distinguish between different tasks or speakers. However, speaker-specific variations in rhythmic patterns can be seen in the following plots, which show the variation in PC1 (top) and PC2 (bottom) across tasks for all speakers. In the PC1 plot, several speakers show large shifts between tasks, particularly noticeable in Task 1 and Task 5, where the lines representing the speakers are either far apart or change significantly, indicating high variability in rhythmic patterns for these tasks. In contrast, in the PC2 plot, the lines for most speakers follow a more uniform trend with smaller differences between tasks, especially after Task 2, indicating greater consistency in rhythmic patterns. The largest deviation in PC2 occurs in Task 1, where some speakers show strong differences in their PC2 scores.

We then used the PCs to investigate how the tasks influence the rhythmic patterns through a mixed-effects model. Treating the PCs as response variables, we analyzed the effects of task and speaker, accounting for the hierarchical structure of the data. Speaker was treated as a random effect to capture individual variability, while text style was treated as a fixed effect to assess its influence independently of differences between speakers. This approach allowed us to determine whether the rhythmic patterns presented by the PCs varied significantly between tasks while accounting for speaker-level differences.

The mixed linear models for both PC1 and PC2 show significant effects of the reading tasks on the rhythmic patterns, also taking speaker variability into account. For PC1 (Table 1), all tasks (t2 to t5) show a significant positive effect compared to the baseline task (Task 2), with Task 5 showing the largest effect, followed by Task 2, Task 4 and Task 3, indicating an increase in PC1 scores when participants move from Task 1 to these tasks.

In contrast, the model for PC2 (Table 2) shows significant negative effects for Tasks 2, 3, and 5, with Task 2 having the largest negative impact, followed by Task 5 and Task 3. Task 4 shows a small, non-significant effect. Speaker variability plays a more substantial role in PC1 (variance = 2.575) than in PC2 (variance = 0.514), meaning that differences between speakers have a greater impact on PC1 than on PC2. Overall, while different tasks significantly influence

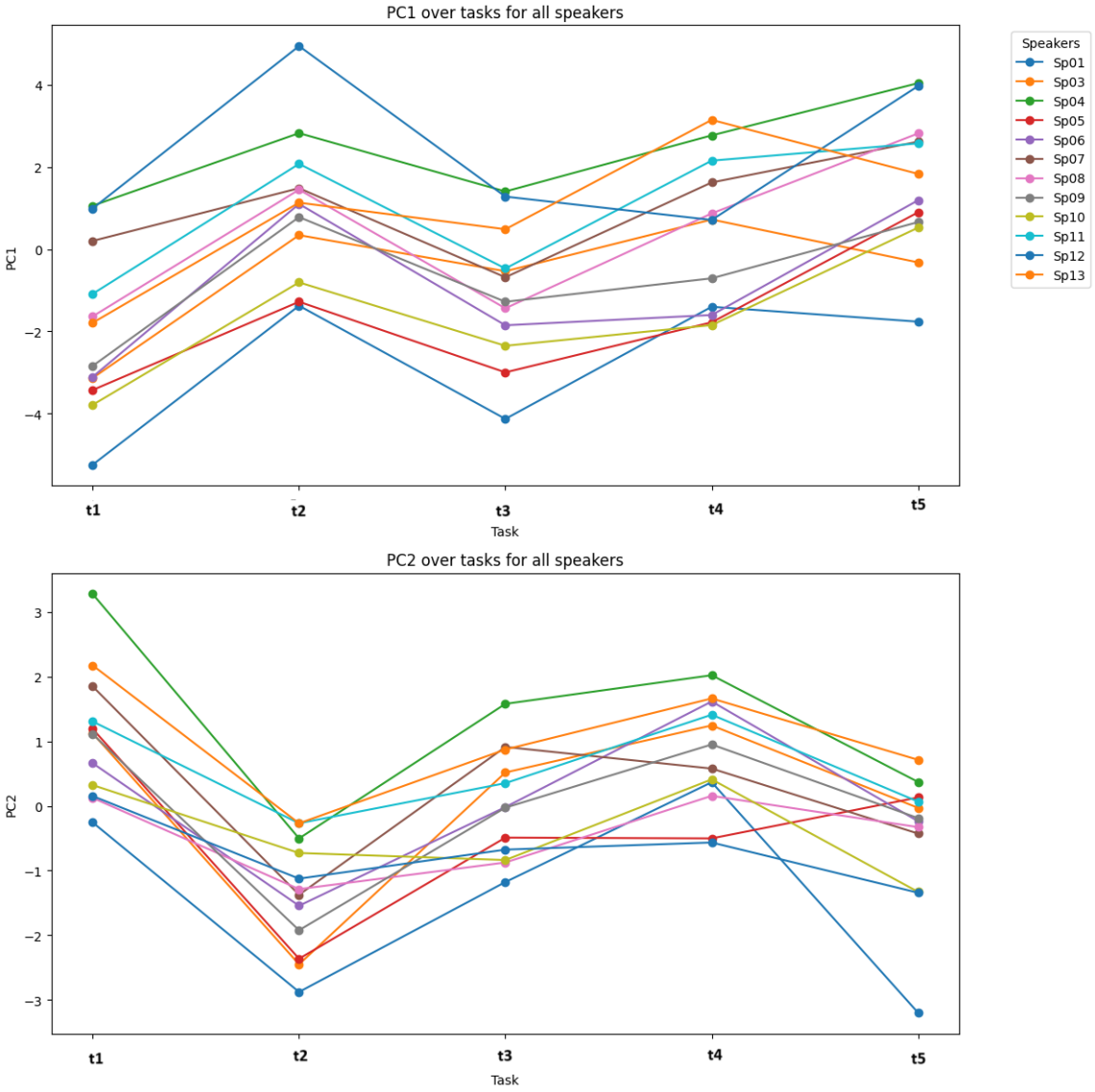



**Figure 6** – Speaker-Specific Variations in PC1 and PC2 Across Tasks

**Table 1** – Summary of Mixed-Effects Model Results for PC1

| **Fixed Effects** | **Estimate (Coef.)** | **Std. Error** | **z-value** | **p-value** |
|---|---|---|---|---|
| Intercept | -1.983 | 0.519 | -3.821 | 0.000 |
| Task 2 | 3.039 | 0.331 | 9.181 | 0.000 |
| Task 3 | 0.936 | 0.331 | 2.827 | 0.005 |
| Task 4 | 2.372 | 0.331 | 7.165 | 0.000 |
| Task 5 | 3.569 | 0.331 | 10.784 | 0.000 |

| Random Effects | Group Variance | Std. Dev |
|---|---|---|
| Speaker (Group) | 2.575 | 1.591 |

**Table 2** – Summary of Mixed-Effects Model Results for PC2

| **Fixed Effects** | **Estimate (Coef.)** | **Std. Error** | **z-value** | **p-value** |
|---|---|---|---|---|
| Intercept | 1.085 | 0.268 | 4.055 | 0.000 |
| Task 2 | -2.478 | 0.240 | -10.312 | 0.000 |
| Task 3 | -1.074 | 0.240 | -4.472 | 0.000 |
| Task 4 | -0.307 | 0.240 | -1.277 | 0.202 |
| Task 5 | -1.568 | 0.240 | -6.528 | 0.000 |
| Random Effects | | Group Variance | Std. Dev | |
| Speaker (Group) | | 0.514 | 0.472 | |

rhythmic patterns in distinct ways, speaker variability also contributes, particularly for PC1.

# 4 Discussion

In this study, the influence of text style on speech reading was investigated by comparing rhythmic patterns, also taking into account the influence of speaker variability. The results indicate a significant effect of text style, although the response variable was not an individual rhythmic metric, but the principal components derived from a set of rhythmic features by principal component analysis. Overall, the results provided valuable insights into the effects of different test materials on speaking time behavior. However, they also made it clear that the experimental design itself is a key factor, which raises additional questions for future research. One emerging question is how the order of reading tasks might influence each other and how such effects can be measured or controlled. To clarify this, future studies will need to repeat the experiment with different task orders across participants to better understand the influence of task order on speaking behavior. Another important approach for future research is the analysis of individual speaker behavior. Although this study involved a relatively small group of participants, differences between speakers were nevertheless observed. While the PCA visualization did not fully capture all of this variability, speaker-specific differences remain an interesting topic for further investigation using other methods of analysis. Some speakers are close to each other, while others occupy a more peripheral position, as can be seen in Figure 2. In addition, there is an interesting trend in how the speakers' behavior changes in different tasks. Figure 3 shows that speakers exhibit similar patterns in the early tasks, while their differences become more pronounced in later tasks. Modeling these evolving distinctions represents a potentially promising direction for future research. Future studies could expand the participant pool to include a more diverse range of speakers, incorporating factors such as gender, age, or language proficiency to better account for individual variability. Additionally, exploring more complex text genres or tasks—such as spontaneous speech or conversational settings—could offer deeper insights into how task structure influences rhythmic speech patterns.

# References


[1] GIBBON, D.: *Speech rhythms: learning to discriminate speech styles*. In *Proc. Speech Prosody*, vol. 2022, pp. 302–306. 2022.

[2] WAGNER, P., J. TROUVAIN, and F. ZIMMERER: *In defense of stylistic diversity in speech research. Journal of Phonetics*, 48, pp. 1–12, 2015.

[3] KENWARD, M. G. and B. JONES: *15 design and analysis of cross-over trials. Handbook of Statistics*, 27, pp. 464–490, 2007.

[4] STURDEVANT, S. G. and T. LUMLEY: *Statistical methods for testing carryover effects: A mixed effects model approach*. *Contemporary Clinical Trials Communications*, 22, p. 100711, 2021.

[5] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 45, pp. 326–347, 2017.

[6] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer [computer program]*. 2024. Retrieved 27 October 2024 from `http://www.praat.org/`.

[7] RAMUS, F., M. NESPOR, and J. MEHLER: *Correlates of linguistic rhythm in the speech signal*. *Cognition*, 73(3), pp. 265–292, 1999.

[8] GRABE, E. and E. L. LOW: *Acoustic correlates of rhythm class*. *Laboratory phonology*, 7(515-546), 2002.

[9] DELLWO, V., P. KARNOWSKI, and I. SZIGETI: *Rhythm and speech rate: A variation coefficient for deltac*. 2006.

[10] MARTIN, M. and B. GOVAERTS: *Limm-pca: Combining asca+ and linear mixed models to analyse high-dimensional designed data*. *Journal of Chemometrics*, 34(6), p. e3232, 2020.

# SEMANTIC TRANSPARENCY AFFECTS THE PHONETIC SIGNAL

*Annika Schebesta*[1], *Jessica Nieder*[2]

[1]*Universität Siegen,* [2]*Universität Passau*
*annika.schebesta@uni-siegen.de*

**Abstract:** Previous research has highlighted the impact of phonological factors and morphological structure on compound pronunciation, but the role of semantics in this interplay remains largely unexplored. Building on psycholinguistic findings that demonstrate semantic effects in compound processing, this study examines whether these effects extend to the production of English nominal triconstituent compounds. Using a state-of-the-art computational model to derive semantic transparency measures, we assessed the predictability of compound meanings from their constituents. Our regression model of experimental speech data revealed a notable duration difference between the first and second constituent and the third constituent for opaque compounds. For semantically transparent compounds, we find a shortening of the third constituent, diminishing the larger duration difference observed for less transparent compounds. These findings underscore the importance of semantic considerations in phonetic analyses of compound constituents, complementing prior research on morphological and phonological correlates of phonetic variation.

## 1 Introduction

Over the past years a growing body of linguistic evidence has shown that the phonetic signal of word forms interacts with their internal organization [e.g. 1, 2, 3, 4, 5, 6, 7], challenging earlier models of lexical processing that posit a strict separation of phonetics and morphology [8, 9]. More concretely, previous empirical research has yielded an effect of the presence of morphological boundaries on the duration and the quality of linguistic units.

Recent studies have demonstrated that phonetic factors can in fact influence different aspects of morphological structure in significant ways and vice versa. For example, research by Plag et al. [5] shows significant differences in acoustic duration between morphemic and non-morphemic word-final S in English. Additionally, word-final S varies according to the type of morphological boundary. For instance, plural S is significantly longer than 3rd person singular S and clitic S.Tomaschek et al. [7] reinforce these results by demonstrating that the discriminative capability of word-final S segments is reflected in their acoustic characteristics. They find that segments with higher discriminative capability tend to be articulated with longer durations, whereas segments with lower discriminative capability show overall shorter durations.

Similarly, Sproat and Fujimura [10] investigate the gradient velarization of /l/. Their results of articulatory measurements indicate that the categorical distinction of the allophones [l] and [ɬ] is inappropriate. Instead, the velarization of [l] appears to be gradient depending on the morphological boundary at which /l/ appears.

Compounds such as *blackbird* or *toothpaste* provide an intriguing testing ground for assessing the relationship of the phonetic signal and morphological structure [11]. Bell et al. [11] examined consonant durations at compound-internal boundaries in English compounds, finding

that duration correlates positively with paradigmatic support and negatively with paradigmatic diversity.

The production and processing of larger morphological constructions has been the focus of Schebesta and Kunter [6] who investigate constituent durations in English left-branching, e.g. *healthcare law*, and right-branching, e.g. *corner drugstore*, triconstituent nominal compounds (NNN). Their statistical analysis reveals that the morphological organization of NNN compounds and the bigram frequency of two neighboring constituents enter a significant interaction. In particular, increasing N1N2 bigram frequencies have a significant shortening effect on N2 constituents in left-branching NNN, while N2 duration increase evidently in right-branching NNN. Similarly, high N2N3 bigram frequencies have a significantly stronger shortening effect on right-branching N3 constituents. The findings indicate that speakers make use of varying constituent durations in order to underpin the intended branching direction in cases where bigram frequencies interfere: The shortening of constituent durations is used to support the branching direction as determined by the morphological structure of the compounds.

Recent advances in computational linguistics and natural language processing have led to a number of new computational models that were proven useful for modelling various aspects of language [12, 13, 14, 15]. Crucially, in these models semantics plays a pivotal role, as it provides a deeper understanding of meaning that is essential for accurately capturing and processing language [16, 17, 18, 19, 20]. One notable example is the use of pre-trained models like BERT (Bidirectional Encoder Representations from Transformers) [15] which use contextual information from the sentential context in which words appear to generate rich semantic representations of words and phrases. These representations are not only crucial for improving performance on traditional NLP tasks but also allow for linguistically relevant applications such as the processing of compound words [16]. In the context of this study, the work by Buijtelaar and Pezzelle [16] demonstrates how measures such as semantic transparency, as derived from BERT models, can be effectively used to analyse and predict human semantic similarity judgements of compound words.

Inspired by recent work [20, 21], we apply the methodology of Buijtelaar and Pezzelle [16] to experimental data from a production study focused on NNN compounds by Schebesta [22]. Nieder et al. [20] used semantic measures taken from a computational model to successfully explain semantic effects in priming data that was collected in an earlier experimental study [23]. Similarly, Nieder et al. [21] detects semantic priming effects in auditory masked-priming data described in Ussishkin et al. [24] through computationally obtained semantic measures.

Following up on this, we aim to explore how semantic transparency influences the phonetic signal of English nominal triconstituent compounds. Our code and data is openly available at `https://osf.io/vfwes/`.

## 2 Experimental data

The data used in this study was elicited by Schebesta [22] in a production experiment with native speakers of North American English at the University of Alberta, Edmonton. The experimental study investigates the impact of branching direction, as determined by morphological structure, on the phonetic signal of English NNN compounds. It tests the prediction that morphologically embedded compound constituents show more phonetic reduction than free constituents.

In order to disentangle the effects of lexical bigram frequency and morphological structure on the duration of NNN compound constituents [6], the NNN in the production experiment were designed in such a way that the bigram frequencies of the neighboring nominal constituents were very low i.e., < 12 hits on COCA [25] so that an interaction of lexical bigram frequency and morphological structure is ruled out. In total, 25 W1W2 pairs e.g. *account service* were

formed as the basis of NNN compounds. Each W1W2 pair was accompanied by a preceding or a following nominal constituent, yielding 50 different NNN e.g. *guest account service* and *account service assistant*. All NNN were spelled with spaces in order to prevent the orthographic form of the NNN from provoking a particular branching direction.

The 50 NNN compounds were embedded in short text passages that contained a context sentence and a carrier sentence. The morphological structure, that is, the branching direction of the NNN compounds was determined by the semantics of the context sentence, so that each NNN compound occurred once as left-branching and once as right-branching in the production experiment. The resulting 100 NNN compounds were produced by 42 participants (37 female, 5 male) who were recorded reading aloud the 100 text passages.

Using R [26], a linear mixed-effects regression model [27] predicted constituent durations and segment durations. The statistical model incorporated a number of phonological (number of phonemes, number of syllables, pitch measurements), lexical (phonological neighborhood density, lexical unigram frequency), and extra-linguistic (speech rate, number of repetitions) noise variables as well as two random intercepts for the speaker and the constituent.

The statistical analysis revealed that N1 and N2 constituent durations are equally long while N3 constituents are significantly longer, irrespective of branching direction. Following these results, the branching direction of the NNN compounds is not reflected in their phonetic signal, whereas the majority of the noise variables has the expected impact on constituent durations. This raises the question whether the semantics of the NNN compounds, that has shown an effect on processing in earlier studies, can explain the duration pattern of the compound constituents.

# 3 Method

## 3.1 Computational workflow

To investigate potential effects of the semantics of NNN on their acoustic signal, we implemented a computational workflow using IPython within the Jupyter Notebook environment. For acquiring word embeddings from BERT, we used the Hugging Face transformers library [28] to interact with the `BERTbase` model trained on English data [15]. Building upon the work of Buijtelaar and Pezzelle [16], we used their `nocontext_vector` function to obtain static meaning representations for each N of our NNN compounds. Notably, retrieving static embeddings from BERT in this way results in a meaning representation that is aggregated across multiple contexts.

In the next step, we calculated the semantic transparency (ST) of our compounds ($c$) following the methodology provided in Buijtelaar and Pezzelle [16]. ST quantifies how transparent the compound's meaning is based on the meanings of its constituents. Crucially, this calculation is based on cosine similarity between the left and the right constituent of the compound [see 16, for details on the original calculations]. Our study, however, involves triconstituent compounds, which differ in the number of constituents from the NN compounds analysed in their study. To accommodate for this difference, we extended the methodology to account for the additional constituent, ensuring that the semantic relationships among all three constituents of our compounds were appropriately captured and reflected in our ST calculations. To do so, we first created a combined vector for the embedded constituents by summing the vectors of the first embedded constituent ($\text{embed}_1$) and the second embedded constituent ($\text{embed}_2$) and taking the average of this combined vector. For ST, the cosine similarity between this embedded compound vector and the free constituent was calculated. The formula for semantic transparency as adapted from Buijtelaar and Pezzelle [16] is given below.

$$\mathrm{ST}(c) = \frac{\mathbf{v}_{\text{free}} \cdot \left( \frac{\mathbf{v}_{embed_1} + \mathbf{v}_{embed_2}}{2} \right)}{\|\mathbf{v}_{\text{free}}\| \cdot \left\| \frac{\mathbf{v}_{embed_1} + \mathbf{v}_{embed_2}}{2} \right\|}$$

The ST measure was then added to the experimental data and subsequently given as input to the R environment [26] for further statistical modelling.

### 3.2 Statistical workflow

For our statistical modelling procedure, we opted for a linear mixed-effect regression model using the R package `lme4` [27]. The preliminary regression model is provided in (1). It contains the dependent variable DURATION which corresponds to the constituent durations as annotated in Praat [29]. The statistical analysis includes an interaction of MEMBER × SEMANTIC TRANSPARENCY × BRANCHING, where MEMBER has the three levels `N1`, `N2` and `N3` corresponding to the three compound constituents, and BRANCHING contains the levels `left-branching` and `right-branching`. Equipped in this way, the regression model is able to predict any effect of SEMANTIC TRANSPARENCY on individual members of the NNN compounds or on members from NNN with a particular branching direction.

The regression model is complemented with a set of noise variables, such as LGUNIGRAMFREQ (the frequency of individual constituents log-transformed to the base 10), the SPEECH RATE at which a NNN compound was produced, and REPETITION which indicates the number of repetitions of the contained W1W2 pairs. Moreover, three Principal Component Analyses [30] were performed with (i) six different pitch measurements (generated with Praat), (ii) four predictors that inform about the phonological form and neighborhood of constituents, and (iii) two predictors that inform about the phonological form of NNN compounds ((ii) and (iii) from the *English Lexicon Project* [31]) in order to minimize the risk of collinearity. The most informative of the resulting principal components are PC1PITCH, PC2PITCH and PC3PITCH (i), PC1PHON and PC2PHON (ii), and PC1NNNPHON (iii), all of which are included in the regression model. In addition, the random intercepts (1 | SPEAKER) and (1 | CONSTITUENT) are included in the statistical analysis to account for speaker-specific variation and a potential influence of features of individual constituents.

(1) DURATION ∼ (1 | SPEAKER) + (1 | CONSTITUENT)
+ MEMBER × SEMANTIC TRANSPARENCY × BRANCHING
+ LGUNIGRAMFREQ
+ SPEECH RATE
+ REPETITION
+ PC1PITCH + PC2PITCH + PC3PITCH
+ PC1PHON + PC2PHON
+ PC1NNNPHON

As the distribution of residuals of the preliminary model did not meet the normality assumption of linear regression, DURATION was Box-Cox-transformed [32] using an exponent of $\lambda = 0.02$. Additionally, 29 outliers from a total of 10,710 observations from 3,573 NNN compounds were excluded from the data set ($\hat{=}$ 0.27%). With the modified dependent variable and the reduced data set, the final regression model was tested for correlated predictors using the `vif()` function from the `car` package [33], which did not reveal any collinearity. The regression model was not further reduced in order to minimize the risk of overfitting and Type I errors [34, 35].

## 4 Results

The results from our linear mixed-effect regression model ($R^2_{marg.}$ 0.562, $R^2_{cond.}$ 0.797, AIC -96542.034) show that N1 and N2 constituents are significantly shorter than N3 constituents, irrespective of branching direction. The noise variables affect constituent durations as expected: phonologically longer constituents with fewer phonological neighbors and constituents with a pitch accent display longer durations, and repetitions of constituents and a higher speech rate lead to shorter constituent durations.

The interaction of MEMBER × SEMANTIC TRANSPARENCY reveals that only N3 constituents are significantly affected by SEMANTIC TRANSPARENCY. As can be observed in the left and middle panel of Figure 1, N1 and N2 durations (measured in seconds) remain the same regardless of increasing SEMANTIC TRANSPARENCY. For N3 durations, displayed in the right panel, we observe a significant facilitatory effect for greater values of semantic transparency, i.e. more transparent NNN.

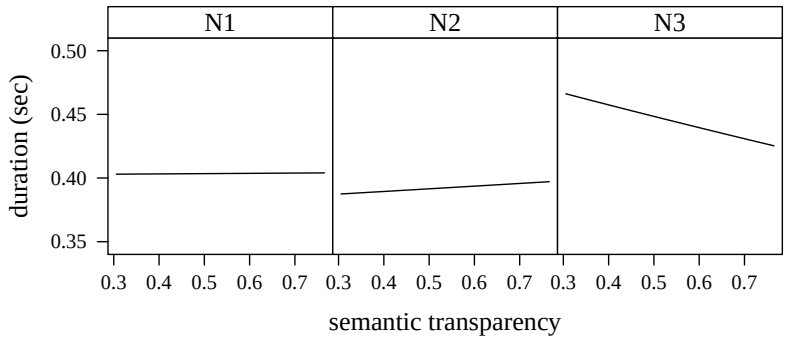



**Figure 1** – Constituent duration by semantic transparency and member

## 5 Discussion

In this study, we tested the effect of semantics on the phonetic signal of NNN compounds. Previous work on NNN compounds has mainly focused on an interaction of morphological structure and the acoustic signal, leaving possible effects of semantics aside.

Recent work has shown the potential of computational models in detecting semantic effects in post-hoc analyses of experimental data [20, 21]. Following this line of research, we extracted distributional meaning representations, i.e. word embedding vectors, for our NNN compounds from a large language model. Using these embeddings, we calculated the semantic transparency of the compounds by assessing the meaning difference between the embedded compound vector and the vector of the free constituent through a cosine similarity calculation [16].

In our statistical model, we observed an effect for opaque vs. transparent compounds: In more opaque NNN, the first and second constituents, N1 and N2, were significantly shorter than the third constituent N3. While this duration pattern remained for semantically transparent NNN, the overall difference in duration between N1N2 and N3 became less pronounced as N3 constituent durations decreased. We interpret this effect as follows. In semantically opaque compounds, speakers indicate the opaqueness through a lengthening of the third constituent, emphasizing the difference in meaning between the embedded compound and the free constituent. In semantically transparent compounds, on the other hand, speakers do not highlight the relationship of the constituents as their meaning emerges naturally from the whole NNN. Thus, previously observed duration differences for opaque compounds become less pronounced

as N3 durations decrease.

Our results provide evidence that the semantic relationship of compound constituents needs to be taken into account for studies focusing on an interaction of morphological structure and the acoustic signal. We show that including measures such as semantic transparency in a post-hoc analysis of experimental data can reveal previously unseen patterns. This can help in further assess and enrich data seeking to explain the morphology-phonetics interaction to gain a more comprehensive understanding of language processing.

## References


[1] BEN HEDIA, S. and I. PLAG: *Gemination and degemination in English prefixation: Phonetic evidence for morphological organization*. *Journal of Phonetics*, 62, pp. 34–49, 2017. doi:https://doi.org/10.1016/j.wocn.2017.02.002. URL https://www.sciencedirect.com/science/article/pii/S0095447017300372.

[2] BLAZEJ, L. J. and A. M. COHEN-GOLDBERG: *Can we hear morphological complexity before words are complex? Journal of Experimental Psychology: Human Perception and Performance*, 41(1), pp. 50–68, 2015. doi:10.1037/a0038509. URL https://doi.org/10.1037/a0038509.

[3] HAY, J. and I. PLAG: *What Constrains Possible Suffix Combinations? On the Interaction of Grammatical and Processing Restrictions in Derivational Morphology*. *Natural Language & Linguistic Theory*, 22(3), pp. 565–596, 2004. doi:10.1023/B:NALA.0000027679.63308.89. URL https://doi.org/10.1023/B:NALA.0000027679.63308.89.

[4] KUNTER, G. and I. PLAG: *Morphological embedding and phonetic reduction: the case of triconstituent compounds*. *Morphology*, 26, pp. 201–227, 2016. doi:10.1007/s11525-016-9284-5.

[5] PLAG, I., J. HOMANN, and G. KUNTER: *Homophony and morphology: The acoustics of word-final S in English*. *Journal of Linguistics*, 53(1), p. 181–216, 2017. doi:10.1017/S0022226715000183.

[6] SCHEBESTA, A. and G. KUNTER: *Constituent durations in English NNN compounds: A case of strategic speaker behavior? Journal of Phonetics*, 94, 2022. doi:10.1016/j.wocn.2022.101164.

[7] TOMASCHEK, F., I. PLAG, M. ERNESTUS, and R. H. BAAYEN: *Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning*. *Journal of Linguistics*, 57(1), pp. 123–161, 2021. doi:10.1017/S0022226719000203.

[8] BERMÚDEZ-OTERO, R.: *The architecture of grammar and the division of labor in exponence*. In *The Morphology and Phonology of Exponence*. Oxford University Press, 2012. doi:10.1093/acprof:oso/9780199573721.003.0002. URL https://doi.org/10.1093/acprof:oso/9780199573721.003.0002. https://academic.oup.com/book/0/chapter/270627764/chapter-pdf/39032501/acprof-9780199573721-chapter-2.pdf.

[9] LEVELT, W. J. M., A. ROELOFS, and A. S. MEYER: *A theory of lexical access in speech production*. *Behavioral and Brain Sciences*, 22(1), pp. 1–38, 1999. doi:10.1017/S0140525X99001776.

[10] SPROAT, R. and O. FUJIMURA: *Allophonic Variation in English /l/ and its Implications for Phonetic Implementation*. *Journal of Phonetics*, 21, pp. 291–311, 1993.

[11] BELL, M. J., S. BEN HEDIA, and I. PLAG: *How morphological structure affects phonetic realisation in English compound nouns*. *Morphology*, 31, pp. 87–120, 2021. doi:10.1007/s11525-020-09346-6.

[12] BAAYEN, R. H., Y.-Y. CHUANG, E. SHAFAEI-BAJESTAN, and J. P. BLEVINS: *The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning*. *Complexity*, 2019.

[13] JOULIN, A., E. GRAVE, P. BOJANOWSKI, M. DOUZE, H. JÉGOU, and T. MIKOLOV: *Fasttext.zip: Compressing text classification models*. *arXiv*, 2016. URL `https://arxiv.org/abs/1612.03651`.

[14] JOULIN, A., E. GRAVE, P. BOJANOWSKI, and T. MIKOLOV: *Bag of tricks for efficient text classification*. *arXiv*, 2016. URL `https://arxiv.org/abs/1607.01759`. `1607.01759`.

[15] DEVLIN, J., M.-W. CHANG, K. LEE, and K. TOUTANOVA: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In J. BURSTEIN, C. DORAN, and T. SOLORIO (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, 2019. doi:10.18653/v1/N19-1423.

[16] BUIJTELAAR, L. and S. PEZZELLE: *A Psycholinguistic Analysis of BERT's Representations of Compounds*. In A. VLACHOS and I. AUGENSTEIN (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2230–2241. Association for Computational Linguistics, Dubrovnik, Croatia, 2023. doi:10.18653/v1/2023.eacl-main.163.

[17] FEIFEI, X., Z. SHUTING, and T. YU: *BERT-based Siamese Network for Semantic Similarity*. *Journal of Physics: Conference Series*, 1684(1), p. 012074, 2020. doi:10.1088/1742-6596/1684/1/012074. URL `https://dx.doi.org/10.1088/1742-6596/1684/1/012074`.

[18] HEITMEIER, M., Y.-Y. CHUANG, and R. H. BAAYEN: *Modeling morphology with linear discriminative learning: considerations and design choices*. *Frontiers in Psychology*, 12, 2021. doi:10.3389/fpsyg.2021.720713.

[19] MIKOLOV, T., E. GRAVE, P. BOJANOWSKI, C. PUHRSCH, and A. JOULIN: *Advances in pre-training distributed word representations*. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[20] NIEDER, J., Y.-Y. CHUANG, R. VAN DE VIJVER, and R. H. BAAYEN: *A Discriminative Lexicon Approach to Word Comprehension, Production and Processing: Maltese Plurals*. *Language*, 99(2), pp. 242–274, 2023.

[21] NIEDER, J., R. VAN DE VIJVER, and A. USSISHKIN: *Emerging roots: Investigating early access to meaning in Maltese auditory word recognition*. *PsyArXiv*, 2024. doi:https://doi.org/10.31234/osf.io/hwumf.

[22] SCHEBESTA, A.: *NNN compounds in English: Investigating the interface of morphology, lexical frequency and the phonetic signal*. Ph.D. thesis, Universität Siegen, 2024. [Unpublished doctoral dissertation].

[23] NIEDER, J., R. VAN DE VIJVER, and H. MITTERER: *Priming Maltese plurals: Representation of sound and broken plurals in the mental lexicon*. *The Mental Lexicon*, 16(1), pp. 69–97, 2021.

[24] USSISHKIN, A., C. R. DAWSON, A. WEDEL, and K. SCHLUTER: *Auditory masked priming in Maltese spoken word recognition*. *Language, Cognition and Neuroscience*, 30(9), pp. 1096–1115, 2015.

[25] DAVIES, M.: *The Corpus of Contemporary American English (COCA)*. `https://www.english-corpora.org/coca/`, 2008.

[26] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

[27] BATES, D., M. MÄCHLER, B. BOLKER, and S. WALKER: *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software*, 67(1), pp. 1–48, 2015. doi:10.18637/jss.v067.i01.

[28] WOLF, T., L. DEBUT, V. SANH, J. CHAUMOND, C. DELANGUE, A. MOI, P. CISTAC, T. RAULT, R. LOUF, M. FUNTOWICZ, J. DAVISON, S. SHLEIFER, P. VON PLATEN, C. MA, Y. JERNITE, J. PLU, C. XU, T. L. SCAO, S. GUGGER, M. DRAME, Q. LHOEST, and A. M. RUSH: *Huggingface's transformers: State-of-the-art natural language processing*. 2020. `1910.03771`.

[29] BOERSMA, P. and D. WEENINK: *Praat: Doing phonetics by computer*. `http://www.praat.org/`, 2017.

[30] JAMES, G., D. WITTEN, T. HASTIE, and R. TIBSHIRANI: *An introduction to statistical learning with applications in R*. Springer, New York, 6th printing 2015 edn., 2013.

[31] BALOTA, D. A., M. J. YAP, M. J. CORTESE, K. A. HUTCHISON, B. KESSLER, B. LOFTIS, J. H. NEELY, D. L. NELSON, G. B. SIMPSON, and R. TREIMAN: *The English Lexicon Project*. *Behavior Research Methods*, 39(3), pp. 445–459, 2007.

[32] BOX, G. and D. R. COX: *An analysis of transformations*. *Journal of the Royal Statistical Society*, 26(2), pp. 211–256, 1964.

[33] FOX, J. and S. WEISBERG: *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edn., 2019. URL `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

[34] HARRELL, F. E.: *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer eBook Collection Mathematics and Statistics. Springer, New York, NY, 2001. doi:10.1007/978-1-4757-3462-1.

[35] WINTER, B.: *Statistics for linguists: An introduction using R*. Routledge Taylor & Francis Group, New York and London, 2020. doi:10.4324/9781315165547. URL `https://www.taylorfrancis.com/books/9781315165547`.

# *EI VERBIBBSCH*: LIFESPAN (IN-)STABILITY OF EAST CENTRAL GERMAN LENITION

*Simon Oppermann*

*Leipzig University, Leipzig, DE*
*simon.oppermann@uni-leipzig.de*

**Abstract:** This paper investigates lenition of fortis plosives across the lifespan of East Central German (ECG) speakers, specifically of 213 distinct individuals appearing in the zoo-docusoap "Elefant, Tiger & Co." between 2003 and 2019. As proxy for lenition, the fraction of locally voiced frames per word-initial pre-vocalic, word-initial pre-sonorant, word-medial intervocalic and word-final fortis and lenis plosive is measured in Praat. This method identifies a slight community trend of decreased voicing in fortis plosives, although individual speakers show highly individual trajectories, with many exhibiting linguistic stability over time. A pattern already observed in the use of other ECG variables, monophthongisation and coronalisation, this might point towards a generational change within the ECG regiolect. The adverse process of fortition in pre-sonorant lenis plosives is a newly emerging feature, especially amongst young ECG speakers. Incidentally, the distance between word-initial pre-vocalic and pre-sonorant lenis plosives is expanding across all recording periods. However, the results underscore the complex interplay between individual linguistic stability and community-level variation. Further methodological approaches are needed to disentangle the influences of e.g. speaker age, socioeconomic status, and the respective communicative situation.

## 1 Introduction

In 2015, the interjection *ei verbibbsch* was chosen as the most endangered Saxon word of the year. This rather mild and intentionally whimsical curse, originally only documented in and around Leipzig, usually acts as an exclamation of astonishment. *Verbibbsch* originated as a minced oath for *verdammich* '[God] damn me', presumably based on the family name *Pippich* [1]. It's a handy phrase to highlight one of the best-known characteristics of the East Central German (ECG) varieties: The lenition of all Standard High German (SHG) fortis plosives and the resulting merger of all SHG fortis and lenis plosives: *Pippich* → *bibbsch* (SHG /p/ → ECG /b/). Note that in this paper, corresponding NHG phonemes and the term "merger" are used as contemporary points of reference, not as predecessors or historical processes.

In the ECG language area, many of the old base dialects were replaced by a new regiolect [2], levelling, on the one hand, the horizontal differences between neighbouring dialects and on the other, the vertical differences between dialectal and standard varieties. Previous research suggests that this societal trend towards dialect levelling is not directly reflected at the individual level of ECG speakers, neither for monophthongisation [3], nor coronalisation [4]. Instead, the most common lifespan pattern seemed to be no significant change at all. This paper aims to add a third variable to this collection, by now looking at the aforementioned lenition.

## 2 A brief history of lenition

In the ECG base dialects, correspondents of NHG fortis plosives /p, t/ tend to be merged with their lenis counterparts NHG /b, d/. Historically, the additional merger of NHG /k/ and /g/ was exclusive to the northern Upper Saxon base dialects [5]. The respective isogloss is depicted as a thick black line in Fig. 1. Whether these mergers result from Inner-German Consonant

Weakening, which led to lenition of MHG fortis consonants first in the Upper German varieties and then reached the ECG varieties in the 15th century [6], or remnants of older, pre-OHG lenis consonants unaffected by the High German consonant shift [7], remains to be thoroughly disentangled. According to Becker & Bergmann [8], the city of Leipzig is located well within the (Southwest) Osterländisch base-dialectal region (see Fig. 1). However, these fine base-dialectal divisions must be treated with utmost care, as the underlying isoglosses are highly terraced and seldom overlap [2], [5]. Although the old base-dialectal structure was postulated to have shown considerable differences, it was already barely traceable by the middle of the 20th century [2]. The remaining standard-deviating variables differ only slightly within the contemporary ECG region and can be subsumed as local variations of one ECG regiolect [9]. Here, all fortis plosives can appear as lenis, even in those areas with base-dialectal distinction of /k/ and /g/ [9] – and in all phonological contexts [10].

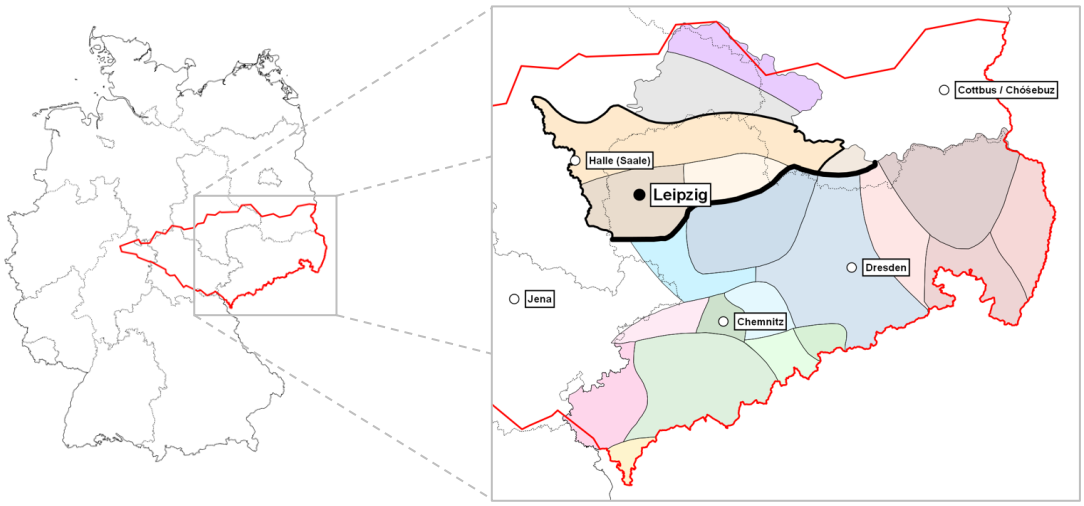



**Fig. 1** - ECG area [11] in red, with supposed Upper Saxon base-dialectal regions [8] shaded following [12]. The medium-thick black line highlights the Osterländisch base-dialects shaded in yellowish-brown. The thick black line represents the isogloss dividing northern lenis from southern fortis velar plosives (e.g. *gind–kind* 'child'). Created with regionalsprache.de [13].

Lenition is often used as an umbrella term describing various ways of reduction in articulatory complexity. At which point a fortis consonant can be considered lenis is just as hard to define as what articulatory aspects create the fortis/lenis opposition in the first place. In his overview, Hahn [14] lists, amongst others, various suggestions like articulatory strength, intensity, duration of aspiration, voicing, VOT, and the duration of either the entire plosive or just its burst phase.

## 3 Individual variation across the lifespan

Although language change is often driven by the respective speech community, the trajectories of its members must not necessarily align with those of the community as a whole. Past longitudinal panel studies have discovered a multitude of highly individual lifespan trajectories: The individual language changes can align with the community change (lifespan change) but can also proceed in the opposite direction (retrograde change). Stable individuals within a changing community can indicate this community change to be the result of each generation speaking differently to the previous (generational change). The opposite constellation can point towards change being a factor of ageing, repeating cyclically as each individual ages (age-grading). In their meta-analysis of 46 English real-time panel studies, Buchstaller & Wagner [15] discovered stability to be the most common pattern across the lifespan. Phonetic-phonological variables were the most stable, accounting for only 19% of the total change observed. While a handful of panel studies have been conducted in German-speaking communities [16]–[21], a comparable

meta-study remains a desideratum. However, previous analyses confirmed Buchstaller & Wagner's [15] findings with most ECG speakers exhibiting stability across their lifespan in their use of monophthongisation [3] and coronalisation [4].

## 4 Elephant, Tiger & Co…rpus

**Fig. 2** - An elephant.

This study is part of the IVaL-project, which as of October 2024 is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project number 549020991). Its main aim is to investigate individual variation across the lifespan (hence IVaL) of ECG speakers. Data basis are episodes of the German zoo-docusoap "Elefant, Tiger & Co.", produced by the public broadcast station MDR. In this show, the employees of Leipzig Zoo are accompanied during their work, they speak to the camera, to each other and to the animals (Fig. 2), entirely without a script. Although they're aware of being filmed, they do not know that their speech is being analysed, circumnavigating the observer's paradox. The protagonists are, apart from the narrator, non-professional speakers, mostly from the ECG area (Fig. 1). Many of them appear regularly, and some have done so since the beginning of the show in 2003. To date, 1102 regular episodes (each ~25min long) have aired, as well as more than 500 additional re-cuts, 40 special episodes and 92 podcast episodes. To manage this overwhelming plethora of data, only the regular episodes, grouped into five evenly spaced recording periods, will be analysed.

## 5 Method

As transcription of these episodes is still ongoing, only four of these recording periods, 2003/04, 2008/09, 2013/14 and 2018/19 can be considered for this paper. Out of 392 total episodes aired within these periods, 305 have already been transcribed, featuring appearances of 213 uniquely identifiable speakers. 25 of them have made appearances in all four recording periods, 31 in three, 38 in two, and 119 in just one recording period. Orthographic time-aligned transcripts of each episode are automatically force-aligned with MAUS [22], embedded within the BAS web services [23]. As the limited extent of this paper doesn't allow to test all proposed articulatory aspects of the fortis/lenis contrast (see chapter 2), for now the focus shall lie on voicing. The fraction of locally voiced pitch frames of each token corresponding to NHG /p, t, k, b, d, g/ is measured in Praat [24]. If this fraction is ≥ 50%, the token is categorised as voiced. As in the ECG regiolect lenition can affect plosives in all positions (see chapter 2), three different phonological contexts will be compared: Word-initial, word-medial and word-final. The word-initial context is split to pre-vocalic (#_V) and pre-sonorant settings (#_S), since following consonantal sonorants /n, l, r/ tends to facilitate the adverse process of fortition in word-initial lenis plosives, especially in younger ECG speakers [9], [25]. Word-medially, only intervocalic settings (V_V) are considered. The word-final settings (_#) are not differentiated by preceding sound. An overview of all tokens gathered is given in Tab. 1.

**Tab. 1** - Number of all tokens, grouped by their phonological context, mainly word-initial pre-vocalic (#_V), word-initial pre-sonorant (#_S), word-medial intervocalic (V_V) and word-final (_#).

| Phoneme | #_V | #_S | V_V | _# | Other | Total |
|---|---|---|---|---|---|---|
| /p/ | 3,186 | 1,786 | 2,212 | 356 | 5,319 | 12,859 |
| /t/ | 5,329 | 1,268 | 11,334 | 49,943 | 32,864 | 100,738 |
| /k/ | 17,223 | 4,821 | 6,765 | 2,609 | 15,029 | 46,447 |
| /b/ | 14,720 | 2,782 | 13,904 | 2,797 | 9,985 | 44,188 |
| /d/ | 82,215 | 3,801 | 8,706 | 7,763 | 10,615 | 113,100 |
| /g/ | 25,556 | 4,850 | 8,150 | 1,624 | 8,570 | 48,750 |
| Total | 148,229 | 19,308 | 51,071 | 65,092 | 82,382 | 366,082 |

## 6 Results

Overall, there seems to be a clear distinction in use of voice between fortis and lenis plosives: As can be seen in Fig. 3, the relative number of voiced tokens was higher for lenis plosives than for the fortis plosives, across all recording periods. This holds true for all phonological contexts, although this contrast is least prominent in word-final setting. Pearson's chi-squared tests at $\alpha = .05$ with plosive type (fortis/lenis) as dependent and percentage of voiced tokens as independent variable showed significant differences between fortis and lenis plosives for all phonological contexts: Word-initial pre-vocalic, $\chi^2$ (1, $N$ = 148,229) = 5,653.13, $p < .001$; word-initial pre-sonorant, $\chi^2$ (1, $N$ = 19,308) = 409.19, $p < .001$; intervocalic, $\chi^2$ (1, $N$ = 51,071) = 4,191.31, $p < .001$, and word-final, $\chi^2$ (1, $N$ = 65,092) = 583.56, $p < .001$. Unsurprisingly, all plosives were most voiced in intervocalic positions. Fortis plosives were least voiced in word-initial pre-vocalic positions, lenis plosives in word-final position.

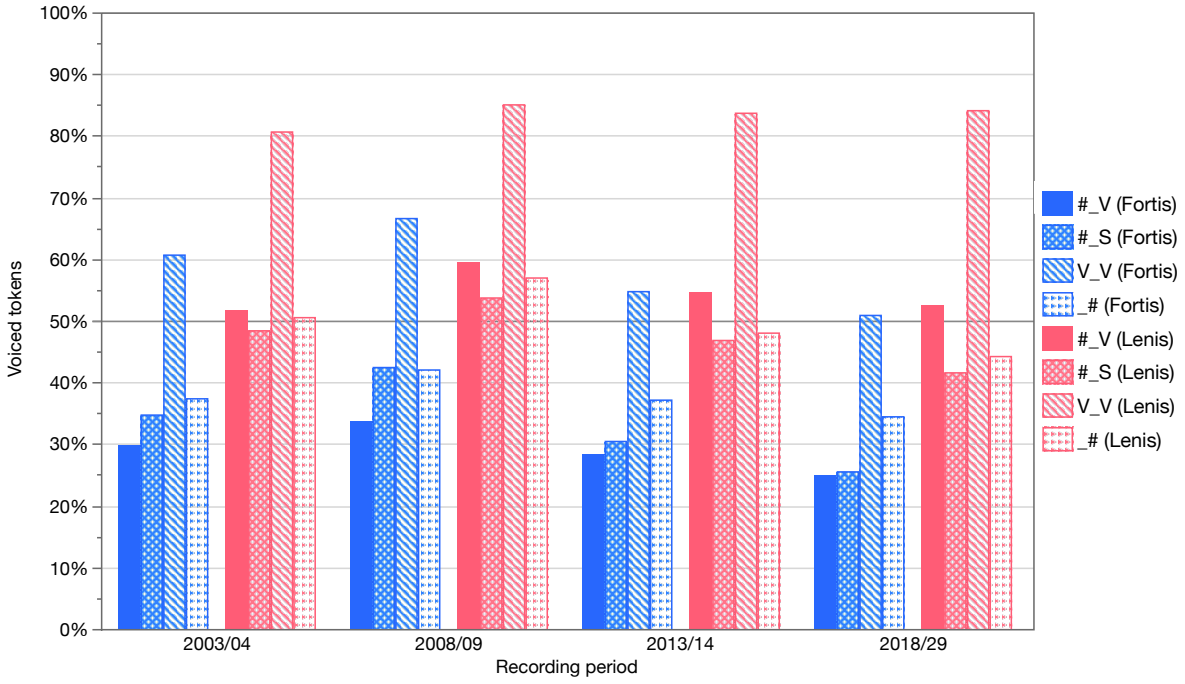



**Fig. 3** - Community trends in percentage of voiced tokens, divided by recording period, fortis/lenis-status of corresponding NHG phoneme and phonological context.

Contrary to their fortis counterparts, word-initial pre-sonorant lenis plosives were less voiced than in pre-vocalic contexts, with this contrast continuously growing across the recording

periods. On average, fortis plosives become less voiced in all positions, with net averages decreasing between approximately 5–10%. Lenis plosives seem to be more stable, especially in intervocalic position. Other net community trends are harder to discern, except for a significant increase in voicing in 2008/09 across all sub-classes.

Let's now zoom in on the individual level and focus on the word-initial, pre-vocalic plosives (Fig. 4). Here, once again, lifespan variation is highly individual. Note that in Fig. 4, only the lifespan trajectories of those 25 speakers are depicted, who appear in all four recording periods. Most of the more prominent speakers, like Speakers 1, 3, 4, 9, 10, 15, 38, 65, show only slight, if any variation, others, like Speakers 2, 27, 75, 94, show oscillation. The 2008/09 net increase is not mirrored, except in Speakers 2 and 27.

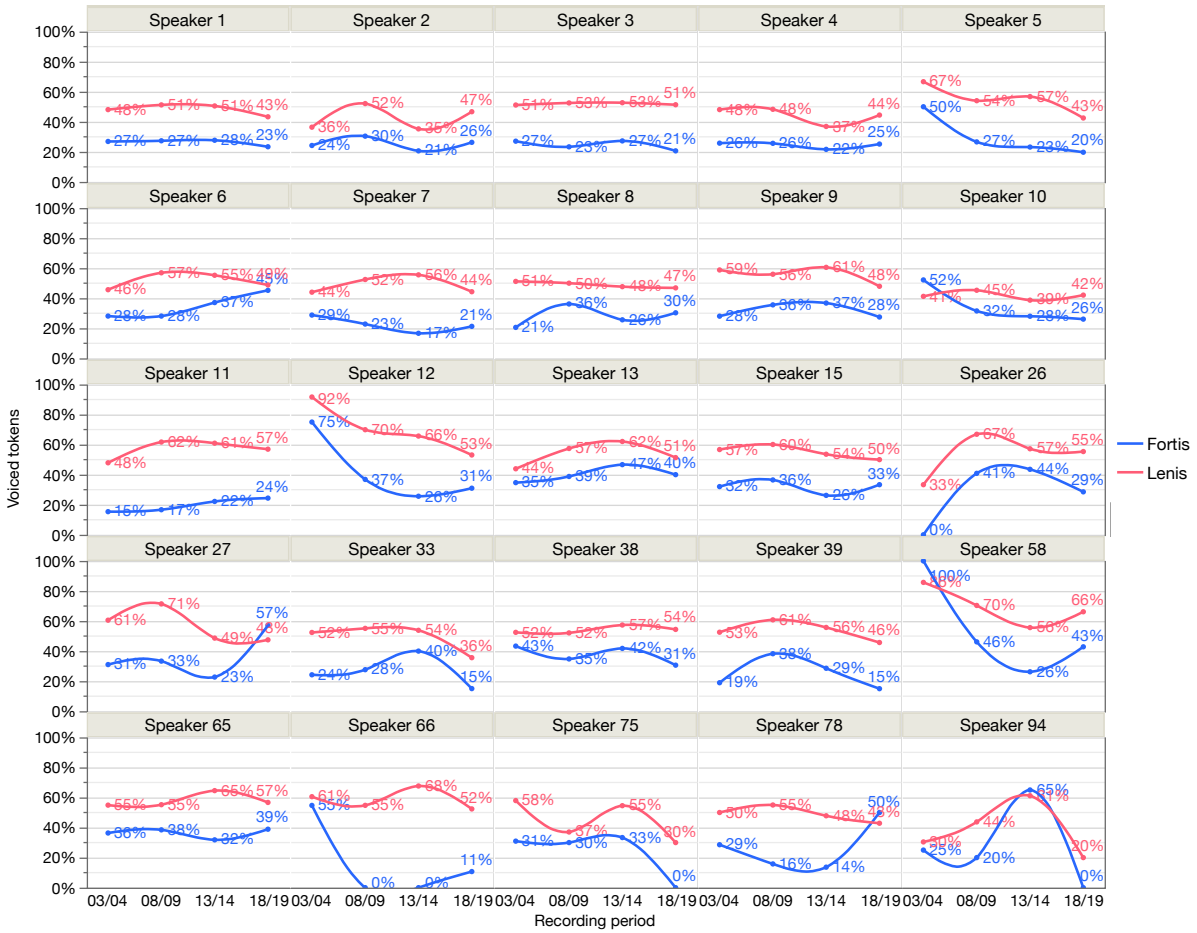



**Fig. 4** - Individual trajectories of the 25 speakers who appear in all four recording periods (just #_V).

All individuals show clear distinction in voicing between fortis and lenis plosives. Notable exceptions are Speakers 10, 27, 58, 78, 94, who at some point in time voiced their fortis plosives more than their lenis plosives. Speaker 66 shows greatest distinction between fortis and lenis. On the one hand, his prominent distinction in voicing could be due to his different socioeconomical status, as he is not only the zoo's director, but also a professor for veterinary medicine. In fact, he and Speaker 38, who works as veterinarian, are the only non-zookeepers within this smaller subsample. On the other hand, when his distinction is greatest (2008/9 and 2013/14), he has made the least appearances resulting in lower number of tokens. Low token numbers often correlate with different recording settings, as people how only appear in the background won't be equipped with their own lavalier microphones, potentially resulting in poorer recording quality. However, this holds less true for Speaker 66, as he as the zoo director almost exclusively appears in planned interview situations, seldom without a designated microphone.

Low token number was already presumed to be a significant factor in [4], where speakers displayed the most extreme values in recording periods with lowest token numbers. Note that the respective token number tends to be reflected in the speaker's assigned code, as the assignment of speaker numbers was based largely on the total number of their appearances. The lower their number, the more tokens could be sampled.

Three other notable trajectories are the ones by Speakers 5, 10, and 12, who display continuous reduction of voicing in fortis plosives. Throughout the recording periods, Speaker 12 experiences a socially upward socioeconomic trajectory within the workplace, starting as apprentice and then rising to become head keeper. As tempting as it may be to interpret her reduction of fortis plosive voicing as a general reduction of standard-divergent variants and connect this to her socially upwards trajectory, one must factor in the aforementioned correlation of low token numbers and extreme measurements. In 2018/19, around 450 fortis and 650 lenis tokens could be assigned to Speaker 12, in 2003/04, it was just a tenth as much. To make matters even more complex, in [4] she displayed the adverse trajectory in relation to use of standard-divergence.

## 7 Discussion & Outlook

A quick glance at the individual trajectories displayed in Fig. 4 reveals almost as many different patterns as there are speakers, making general individual patterns very difficult to differentiate. The community exhibited a slight net decrease in voicing of fortis plosives. As previously observed in [3] and [4], this was barely mirrored on the individual level. Instead, most speakers remained largely stable, with only minor variations. This could indicate towards a generational change, where individual speakers display linguistic stability, but each new generation behaving differently to the previous. For future studies, this will be tested by additional apparent-time studies at the different recording periods, as stratification by respective speaker age could show potential generational influence.

In previous tests, the respective communicative situations within "Elephant, Tiger & Co." have shown to influence the use of coronalisation [4] and other variables like centralisation of rounded back vowels. The average use of standard-divergent variants was highest when individuals were speaking towards colleagues and lowest when speaking towards the camera; when speaking towards animals, it was somewhere in the middle. This can now be confirmed for lenition of fortis plosives: Colleague 43.16%, camera 37.9%, animal 39.17%, $\chi^2$ (2, $N$ = 151,251) = 268.69, $p < .001$). However, a more detailed differentiation is needed, as these influences are once again highly individual.

As voicing tends not to be the sole correlate of the German fortis-lenis-contrast, other, e.g. durational aspects (see chapter 2) should considered as well. In Swiss German, this distinction is distinguished by closure duration [26]; Hahn [14] was able to show that across all German-speaking countries, burst duration of velar fortis plosives were more than twice as long as of their lenis counterparts. A quick test shows that in the IVaL-corpus, lenis plosives ($Mdn$ = 30 ms, $M$ = 42.16 ms, $SD$ = 34.19 ms) are on average much shorter than fortis plosives ($Mdn$ = 41.9 ms, $M$ = 66.75 ms, $SD$ = 132.9 ms), $t$(366126) = -80.62, $p < .001$.

For once, the usual ceterum censeo of empirical linguistic studies, which postulates that reliable results can only be achieved with more data [27], can be rephrased. With roughly 2.5 million tokens (Ei verbibbsch!), the "Elefant, Tiger & Co."-corpus has reached a quite formidable size. Further reliable insights into individual variation across the lifespan shall now be achieved with additional methodological approaches.

## Acknowledgements

I'm highly grateful for our current and past student assistants Laura Döring, Pia Henkel, and Maren Schurer, as well as countless students over the past decade (!), without whom this enormous treasure trove of data would not have existed. This publication is part of the research project "Individual Variation Across the Lifespan" (IVaL), funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project number 549020991). Image credit of Fig. 2: BirdHunter591 (iStock).

## References

[1] SÄCHSISCHE AKADEMIE DER WISSENSCHAFTEN ZU LEIPZIG (Ed.): *Wörterbuch der obersächsischen Mundarten. Band 4: S–Z*. Berlin: Akademie, 1996.

[2] SIEBENHAAR, B.: *Ostmitteldeutsch: Thüringisch und Obersächsisch*., In: J. HERRGEN and J. E. SCHMIDT (Eds.): *Deutsch: Sprache und Raum – Ein internationales Handbuch der Sprachvariation*, Berlin: De Gruyter Mouton, 2019, pp. 407–435. DOI: 10.1515/9783110261295.

[3] OPPERMANN, S. and B. SIEBENHAAR: *What's that phthong? Automated classification of dialectal mono- and standard diphthongs*. In *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, 2023, pp. 3637–3642.

[4] OPPERMANN, S. and B. SIEBENHAAR: *Tracing coronalisation across the lifespan* [Poster], In *Phonetik und Phonologie im deutschsprachigen Raum (P&P19)*, Bern, 2023. DOI: 10.13140/RG.2.2.28391.42403.

[5] BERGMANN, G.: *Upper Saxon*. In: C. V. J. RUSS (Ed.): *The Dialects of Modern German*, London: Routledge, 1990.

[6] MOSER, V.: *Frühneuhochdeutsche Grammatik. 1. Band: Lautlehre, 3. Teil: Konsonanten*. Heidelberg: Carl Winter Universitätsverlag, 1951.

[7] CZAJKOWSKI, L.: *Schreibsprachen im Übergang: Untersuchungen zum Sprachwandel im niederdeutsch-ostmitteldeutschen Übergangsraum im Spätmittelalter und in der Frühen Neuzeit*. Hildesheim et al.: Olms, 2021.

[8] BECKER, H. and G. BERGMANN: *Sächsische Mundartenkunde. Entstehung, Geschichte und Lautstand der Mundarten des obersächsischen Gebietes*, 2nd ed. Halle (Saale): Max Niemeyer, 1969.

[9] ROCHOLL, M. J.: *Ostmitteldeutsch – eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum*. Hildesheim et al.: Olms, 2015.

[10] SCHIRMUNSKI, V. M.: *Deutsche Mundartkunde. Vergleichende Laut- und Formenlehre der deutschen Mundarten.* Frankfurt am Main et al.: Peter Lang, 2010.

[11] LAMELI, A.: *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Berlin: De Gruyter, 2013. DOI: 10.1515/9783110331394.

[12] FORSCHUNGSZENTRUM DEUTSCHER SPRACHATLAS, R. HÜNECKE, and E. KOCH: *Mundarten und Regionalsprache in Sachsen. Informationsportal zur Sprachgeographie*, 2023. https://dsa.info/sachsen/ (accessed Oct. 28, 2024).

[13] SCHMIDT, J. E., J. HERRGEN, R. KEHREIN, and A. LAMELI (Eds.): *Regionalsprache.de (REDE III). Forschungsplattform zu den modernen Regionalsprachen des Deutschen.* Marburg: Forschungszentrum Deutscher Sprachatlas, 2020–. https://regionalsprache.de/ (accessed Oct. 26, 2024).

[14] HAHN, M.: *Zwischen Prozess und Produkt: Zur Lenisierung velarer Plosive im Deutschen*., In: M. HAHN, A. KLEENE, R. LANGHANKE, and A. SCHAUFUß (Eds.): *Dynamik in den deutschen Regionalsprachen: Gebrauch und Wahrnehmung. Beiträge aus dem Forum Sprachvariation*, vol. 250–251, Hildesheim et al.: Olms, 2020, pp. 87–124.

[15] BUCHSTALLER, I. and S. E. WAGNER: *Life-span changes and the history of English*. In: R.

HICKEY (Ed.): *New Cambridge History of the English Language*, Cambridge, UK: Cambridge University Press, to appear.
[16] BAUSCH, K.-H.: *Wandel im gesprochenen Deutsch: Zum diachronen Vergleich von Korpora gesprochener Sprache am Beispiel des Rhein-Neckar-Raums*. Mannheim: Institut für Deutsche Sprache, 2000.
[17] SIEBENHAAR, B.: *Sprachwandel von Sprachgemeinschaften und Individuen*. In: A. HÄCKI BUHOFER (Ed.): *Spracherwerb und Lebensalter*, Tübingen & Basel: A. Francke, 2003, pp. 313–325.
[18] LAMELI, A.: *Standard und Substandard: Regionalismen im diachronen Längsschnitt*. Stuttgart: Steiner, 2004.
[19] BIEBERSTEDT, A., J. RUGE, and I. SCHRÖDER (Eds.): *Hamburgisch: Struktur, Gebrauch, Wahrnehmung der Regionalsprache im urbanen Raum*. Frankfurt am Main: Peter Lang Edition, 2016.
[20] VERGEINER, P. C., D. WALLNER, and L. BÜLOW: *Language change in real-time*. In: *The Coherence of Linguistic Communities*, 1st ed., New York: Routledge, 2022, pp. 281–300. DOI: 10.4324/9781003134558-21.
[21] BEAMAN, K. V.: *Language Change in Real- and Apparent-Time: Coherence in the Individual and the Community*, 1st ed. New York: Routledge, 2024. DOI: 10.4324/9781003267331.
[22] SCHIEL, F.: *A Statistical Model for Predicting Pronunciation*. In *Proceedings of the 18th International Congressof Phonetic Sciences*, Glasgow, 2015.
[23] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*., *Computer Speech & Language*, vol. 45, pp. 326–347, 2017, DOI: 10.1016/j.csl.2017.01.005.
[24] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer* [Computer program]. Available at http://praat.org/ (accessed: Nov. 30, 202)
[25] VORBERGER, L.: *Plau, krau und krün? – Zur Verteilung und Variation der anlautenden Fortisierung von Plosiven vor Sonoranten*., *LO*, vol. 128, no. 04, pp. 101–117, 2024, DOI: 10.13092/lo.124.10632.
[26] LADD, D. R. and S. SCHMID: *Obstruent voicing effects on F0, but without voicing: Phonetic correlates of Swiss German lenis, fortis, and aspirated stops*. In *Journal of Phonetics*, vol. 71, pp. 229–248, 2018, DOI: 10.1016/j.wocn.2018.09.003.
[27] SIEBENHAAR, B.: *Instrumentalphonetische Analysen zur Ausgestaltung des Sprechlagenspektrums in Leipzig*. In *ZDL*, vol. 81, no. 2, pp. 151–190, 2014, DOI: 10.25162/zdl-2014-0005.

# MetaHumans as holistic Personae for implementation into Virtual Reality

*Miriam Oschkinat[1], Denise Bischof[1], Melanie Weirich[2], Stefanie Jannedy[1]*

*[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft Berlin,*
*[2]Friedrich Schiller Universität Jena*
*oschkinat@leibniz-zas.de*

**Abstract:** The scope of this study comprises preparatory work for the implementation of so-called *MetaHumans* into virtual reality (VR) environments, thereby creating immersive experiment designs optimized for socio-phonetic objectives. This comprehensive creation process requires the design of MetaHumans as holistic *Personae* with different appearances and testing the perception of their visual appearance as rated by naive individuals. Further, different rooms/scenarios need to be created in VR. The rated personae can then be animated with a typical addressee motion (such as nodding or thinking motions) and subsequently be implemented into the VR scenarios. The present study is concerned with the persona creation process and the raters' perception of these Personae. For the purpose of testing the effect of addressee gender on the participants' speech in the final experiment design, a set of 5 middle-aged MetaHumans was created. These were designed such that their perceived gender would symmetrically align on a continuum from *feminine* to *masculine*. Meanwhile, other factors such as perceived age, naturalness, and personality ratings gathered across a benevolence/dominance spectrum were intended to be kept constant across MetaHumans. Ratings from 453 raters across different age groups and across genders situated the 5 MetaHumans along a gender continuum, varying in masculinity and femininity. Other rating factors did not show unexpected differences between MetaHumans. Interactions with raters' age highlighted the role of hairstyle for gender and age perception. These ratings set the ground for interpreting the participants' speech dependent on addressee gender within the virtual environment scenarios.

## 1 Introduction

Speech production is inherently variable. Despite limitless degrees of freedom, a large amount of this variability can be attributed to factors either rooted in physiology (nature), such as the systematic difference in fundamental frequency between male and female speakers, or in learned behavior (nurture), such as speaker accents that expose a regional or social affiliation to a certain group. Variation in speech production requires adaption in speech perception, that is the listener's ability to precisely map different acoustic realizations to learned representations of phonemes, syllables, or words [1] for successful verbal interaction. In sociophonetic research, the learned behavior and associated phonetic patterns that constitute inter- and intra-speaker variability are of key interest.

### 1.1 The study of systematic speech variation

In studying systematic inter-speaker variability, it was shown that different speakers vary in their speech depending on age [2], gender [3, 4], social class [5] or other categories that are relevant in the individual's environment [6]. This variation is associated with the (subconsciously) self-ascribed affiliation to a certain group of speakers, manifesting in the linguistic and fine phonetic detail in their speech.

Aside from such inter-individual variation, individual speakers are able to flexibly adapt their speaking style within their range of possibilities and frequently do so depending on who they talk to [7], including factors such as the relationship to the addressee [8] or the addressee's

and the situational formality [9], as well as the topic of conversation [10]. This intra-individual variation based on functional and situational parameters is also referred to as *speech register* [11]. The current study is situated in the scope of investigating fine-phonetic detail in speech register.

### 1.2 Person perception

A significant amount of human speech springs from the desire to communicate and is therefore generally directed to another person, the addressee. Thereby, the perception of the other person plays a significant role in shaping our interpretation and reception of spoken communication [12]. Our pre-existing attitudes, biases, and prior experiences influence how we process and understand verbal messages [i.e., 13]. This subjective filtering can lead to selective listening, misinterpretation, or overemphasis on certain aspects of speech. Additionally, cultural background, social context, and power dynamics all contribute to how we perceive and respond to spoken language. Understanding these factors is crucial for effective communication, as it allows speakers to tailor their message and listeners to critically evaluate the information presented. Furthermore, research in social psychology has shown that person perception can significantly impact speech processing, particularly in cross-cultural contexts or when dealing with unfamiliar accents or speaking styles. Therefore, recognizing the complex interplay between speaker, listener, and linguistic variables is essential for fostering clear and productive dialogue.

### 1.3 Investigating speech register with novel experimental designs

A significant challenge in studying speech register poses the controllability of experimental setups: Communication in the field is difficult to replicate, and laboratory settings inevitably alter communicative behavior. In previous investigations of our group, speech register was examined with a special interest in addressee formality by using pre-recorded video stimuli of an addressee, aiming for an ecologically valid method for testing intra-individual variation with respect to situational and functional parameters [11]. In this setup, participants were tested in the lab (lab condition) or at home (home condition) by looking at and talking to a video of a female addressee who was dressed in different guises varying in formality. The addressee was motioning a typical listener's reaction while the participant spoke (such as light nodding, head tilting, etc.) [14]. In speech analysis, Duran et al. [14] found participants to change their speech depending on the formality of the addressee (formal or informal guise). Specifically, participants slowed down their speech rate and dispersed their vowel formants in the formal compared to the informal setting [15]. Female speakers also lowered their F0 and showed a narrower F0 range while the effect was non-significant for the male speaker group. Additionally, and even more intriguingly, they also found an effect of space: Differences in F0 between formal and informal conditions were significant for the speakers in the lab condition, but not in the home condition, this effect sparking the assumption that the place of conversation impacts speech register as well.

### 1.4 Scope of the study

The larger scope of this project thus aims at testing the effect of *addressee formality*, *age*, and *gender* as influencing factors on speech register, and additionally examines the effect of space/location on the participant's speech. Locations will be varied in formality and other factors such as publicness and crowdedness while keeping the geometry of the place, the addressee, and the topic of conversation constant. By creating the entire experimental setup in virtual reality, we achieve maximum controllability. Addressees are created as digital MetaHumans and will be animated and implemented into virtual reality rooms, creating an immersive experience in these experimental setups. The current study is concerned with a subpart of this project: The following sections describe the creation of different MetaHumans as addressees and their

evaluation including the factors of interest as validation for implementation into virtual reality Scenarios. Personae are stereotyped versions of people types that pattern together by virtue of their way of being or social attributes and characteristics [16]. We created 5 MetaHumans who share features of their appearance, in the sense that they could be perceived as being related. The superordinate project and all related testing will take place in Berlin, Germany. Therefore, in choosing features for the MetaHumans, we aimed for five middle-aged Caucasian/white Personae with medium blond hair with some grey strands and greenish eyes.

## 2 Experiment

### 2.1 Stimuli

*Personae* were created as MetaHumans using the browser-based MetaHuman creator (https://metahuman.unrealengine.com/ last visited 14.10.2024) from *Epic Games* for the Unreal Engine (version 5.5). MetaHuman is a complete framework for creating and animating highly realistic digital humans which can be implemented in any Unreal Engine environment, such as self-created video-game-like places and scenarios.

The five MetaHumans were created with the aim to be perceived equally distributed along a gender continuum from female (*weiblich)* to male (*männlich*) while keeping other factors such as age, naturalness, and personality ratings constant across MetaHumans. The MetaHuman application provides a set of example characters to serve as a starting base, which can then be morphed with other existing example characters in each one of their features. The gender continuum was achieved by creating one male and one female MetaHuman with similar features from example characters. Subsequently, features from the characters that composed our female/male character were morphed together to create MetaHumans with more male or female features along a continuum. Facial parameters that were changed were e.g. the shape and height of the eyebrows, the lips, the jaw, and the cheeks/cheekbones. The most feminine person wears make-up and a chignon showing some loose hair on the sides (Fig. 1A). The supposedly second-most female MetaHuman was wearing a hairstyle with side bangs (Fig. 1B). Another person was created with the exact same face and features as in Fig. 1B, but with the male hairstyle (Fig.1C). For the supposedly most and second-most male MetaHuman, the same short-hair hairstyle was chosen (Fig. 1D, 1E). The most masculine person had a short stubble. All five MetaHumans were dressed in a white shirt and were portrayed with neutral facial expressions (like headshots for a passport picture) in front of a light-grey background. Pictures were created by fitting them into a screenshot frame with a resolution of 144 dpi (cf. Fig. 1).


**Figure 1** – Five MetaHumans intended to be perceived from female (A) to male (E). MetaHuman B and C differ exclusively in hairstyle. Colored frames added for visualization in the Figures further below. Pictures were presented without colored frames.

## 2.2 Questionnaire

Questionnaires were created with easyfeedback (*easy-feedback.de*). Each questionnaire contained pictures of one MetaHuman, thus a total of five questionnaires were implemented. Those questionnaires contained questions about the MetaHuman's perceived age in years (slide control from 15 to 99), their gender (slide control from 0 *masculine* to 28 *feminine*), and their naturalness from 0 *very unnatural* to 28 *very natural*. Additionally, raters were instructed to evaluate the MetaHuman's personality with 8 selected attributes on a benevolence/dominance spectrum (the single attributes are presented in Figure 3d). The personality attributes were presented in pairs along with the picture of the respective MetaHuman and rated on a seven-point scale (1-7). The interpretation of the scale (with 0 being the most positive or most negative attribute, if applicable) was counterbalanced between attributes. Raters were asked if the person in the picture reminded them of a public person (from politics or media, open question). Lastly, we collected metadata about the rater's age, sex and gender, their preferred language, and the state they currently live in. To date, German does not have a straightforward wording that maps onto the concepts of *sex* and *gender* that exist in English. Therefore, the raters' sex and gender were assessed with a questionnaire close to the recommendation by Diethold, Watzlawik [17] for an inclusive assessment of sex and gender in German studies/questionnaires. Reflecting sex, we asked, "Was ist Dein biologisches Geschlecht?" (one of: *masculine*, *feminine*, *diverse*) and reflecting gender, we asked, "Welchem Geschlecht fühlst Du Dich zugehörig?" (one of: *masculine*, *feminine*, *non-binary/genderqueer, no gender, no indication/own indication*). Further, raters indicated their sexual orientation on the Kinsey Scale [18], and their self-ascribed masculinity and femininity on the traditional masculinity and femininity scale (TMF) [19]. The whole questionnaire took approximately 4 minutes.

## 2.3 Participant recruitment

Raters were recruited via clickworker (www.clickworker.de) and were, based on clickworkers' recommendation, paid 80 cents for participation. Every rater received only one MetaHuman/questionnaire. Per questionnaire, 15 male and 15 female individuals were recruited per age group, including individuals who identify as genderqueer or genderless. The three age groups contained individuals between 18 and 30, between 31 and 50, and between 51 and 75 years of age. Thus, each MetaHuman questionnaire was sent to a total amount of 90 individuals, leading to a total amount of 450 different raters recruited for this study. All participants were recruited as native speakers of German and based in Germany.

## 2.4 Data preparation

*Clickworker* allows for a selection of participants based on e.g. age, sex, language, etc. Although we recruited based on the criteria above, some noise in the data was found in the sense that the recruited age groups did not precisely match the input of our participants, presumably due to a mismatch between their data in *Clickworker* and the reality, or mishaps in age selection. Thus, of all the participants who completed a full questionnaire, we excluded two participants from the data who were below the age of 18, and three participants who clearly entered non-sense answers, such as rated the age above 70, and answered non-sense words in the open questions. Eventually, we achieved groups of 30±2 participants per age group/gender, and 90+2 for the total amount of participants per MetaHuman.

# 3 Results

## 3.1 Raters' demographics

Our 453 raters (219 male, 233, female, 1 diverse) were between the age of 18 and 75 (after data exclusion, see section 2.4) and self-identified as male (219), female (226), non-binary/gender-queer (6) and without gender (2). Two of them indicated to feel most comfortable in English, one in Turkish, and the rest in German. Our raters were proportionally located over the states of Germany measured on population density. Figure 2 shows our raters' mean self-ascribed femininity/masculinity on the TMF scale (a), and their sexual orientation on the Kinsey scale (b). A total of 30 participants did not indicate their sexual orientation.

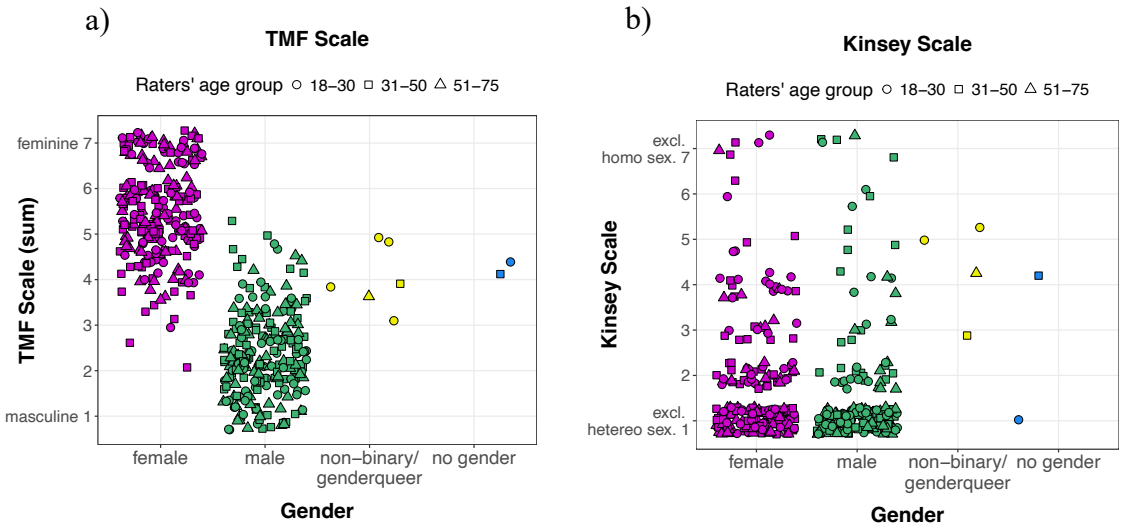



**Figure 2** – Raters' self-ascribed femininity/masculinity on the TMF scale (a) and sexual orientation (b).

## 3.2 MetaHuman ratings

The five created MetaHumans (MH) were intended to vary in their perceived gender but not in age, naturalness, or personality attributes. For perceived gender (*MH_Gender*), age (*MH_Age*), and naturalness (*MH_Naturalness*), we calculated one linear model each investigating the effect of *MetaHuman* (five levels) and *Rater's age group* (three levels) on the dependent variable. The models examined both main effects and their interaction. Analyses were conducted in *R-studio* (*R* version 4.4.1). Post-hoc testing with emmeans pairwise comparison was performed, if applicable. The following section provides a summary of the most relevant findings.

### 3.2.1 MetaHuman gender

The model for *MH_Gender* was overall significant ($F(14,438) = 75.57$, $p < .001$) explaining 69.8% of the variance in the data (adjusted $R^2$). As post-hoc tests revealed, all MetaHumans differed significantly from each other in perceived gender, thus positioning them along the intended gender continuum (see Figure 3a). Raters' age group was not significant, but some of the interactions between MetaHuman and rater's age group. Interactions were thus further examined with *emmeans* pairwise comparison assessing differences in *MH_Gender* scores across MetaHumans within each rater's age group (see Figure 3a, colored dots, squares, and triangles). Given the robust effect of MetaHuman on *MH_Gender*, the following section reports the non-significant effects of MetaHuman on *MH_Gender* for each age group.

For the youngest group of raters (18-30 years), the difference between MetaHumans A and B was not significant, thus those two MetaHumans were perceived equally feminine, while

for the other two raters' age groups (31-50 years, 51-75 years), MetaHuman B was perceived significantly less feminine than A (Figure 3a). MetaHumans C and D were rated similarly regarding their gender by the two age groups 31-50 years and 51-75 years, but significantly different by the youngest age group. These findings suggest firstly that for younger raters (group 18-30), there is a greater tendency towards more feminine ratings than for the older raters. Secondly, we assume that, for the younger age group, the different face features between MetaHumans B and C mainly contributed to the differences in gender ratings, while for the older groups of raters (31-50, 51-75) the similar short-hair hairstyle could have been a stronger indicator for rating MetaHumans C and D equally masculine. Strikingly, MetaHumans B, C, and D show a much higher rating variability than A and B, reflecting ambiguines in their gender perception.

Another comparison of differences in *MH_Gender* scores *across* raters' age group *within* MetaHumans revealed that within one MetaHuman, gender ratings differed significantly between all age groups for MetaHumans B and C, with the younger raters' tending towards more feminine ratings and the older raters towards more masculine ratings (see Figure 3a, similar colored dots, squares, and triangles).

### 3.2.2 MetaHuman age

MetaHumans were rated between a mean age of 39.9 and 48.4 years (Figure 3b, black dots). The model for *MH_Age* explained 19, 2% of the variance (adjusted $R^2$). Although modest, this effect was statistically significant ($F(14,438) = 8.657$, $p < .001$), suggesting that the combination of MetaHuman, raters' age group, and their interaction had a meaningful impact on *MH_Age*. The main effect of MetaHuman was significant for some of the MetaHumans, while the main effect of raters' age group was non-significant. Overall, the two most feminine MetaHumans A and B were rated older than the more masculine MetaHumans C, D and E.

Emmeans pairwise comparison examined differences in *MH_Age* scores *across* levels of raters' age group *within* each MetaHuman. The model revealed that age ratings of the youngest group (18-30 years) were significantly higher than for the two older rater groups for MetaHuman B (cf. Figure 3b, purple dots, triangles, and squares). Note that the two MetaHumans who differ exclusively in hairstyle but not in facial features differ the most (Figure 3b, MetaHuman B - 48.4 years and C - 39.9 years).

### 3.2.3 Naturalness and personality ratings

The model for MetaHuman naturalness was non-significant, indicating no differences in observed naturalness for any of the MetaHumans (Figure 3c). Mean values for all MetaHumans were below medium naturalness (Figure 3c, grey vertical line), however with high variability.

Personality ratings were observed visually by calculating the mean per MetaHuman for each of the attributes, independent of raters' age group. Figure 3d indicates that all MetaHumans show strong alignment for each of the attributes on the Likert scale (from 1 to 7). The highest variability is shown for the attribute "non-attractive – attractive" (*unattraktiv - attraktiv*), however, varying by less than 1 point.

## 4 Summary and outlook

This study examined the perception of 5 MetaHumans created for socio-phonetic experimental designs in virtual reality. Overall, the ratings validated the perception of the MetaHumans along a gender continuum. Younger raters showed a tendency for rating some of our MetaHumans more feminine. Importantly, MetaHumans did not differ in perceived naturalness, but in age.

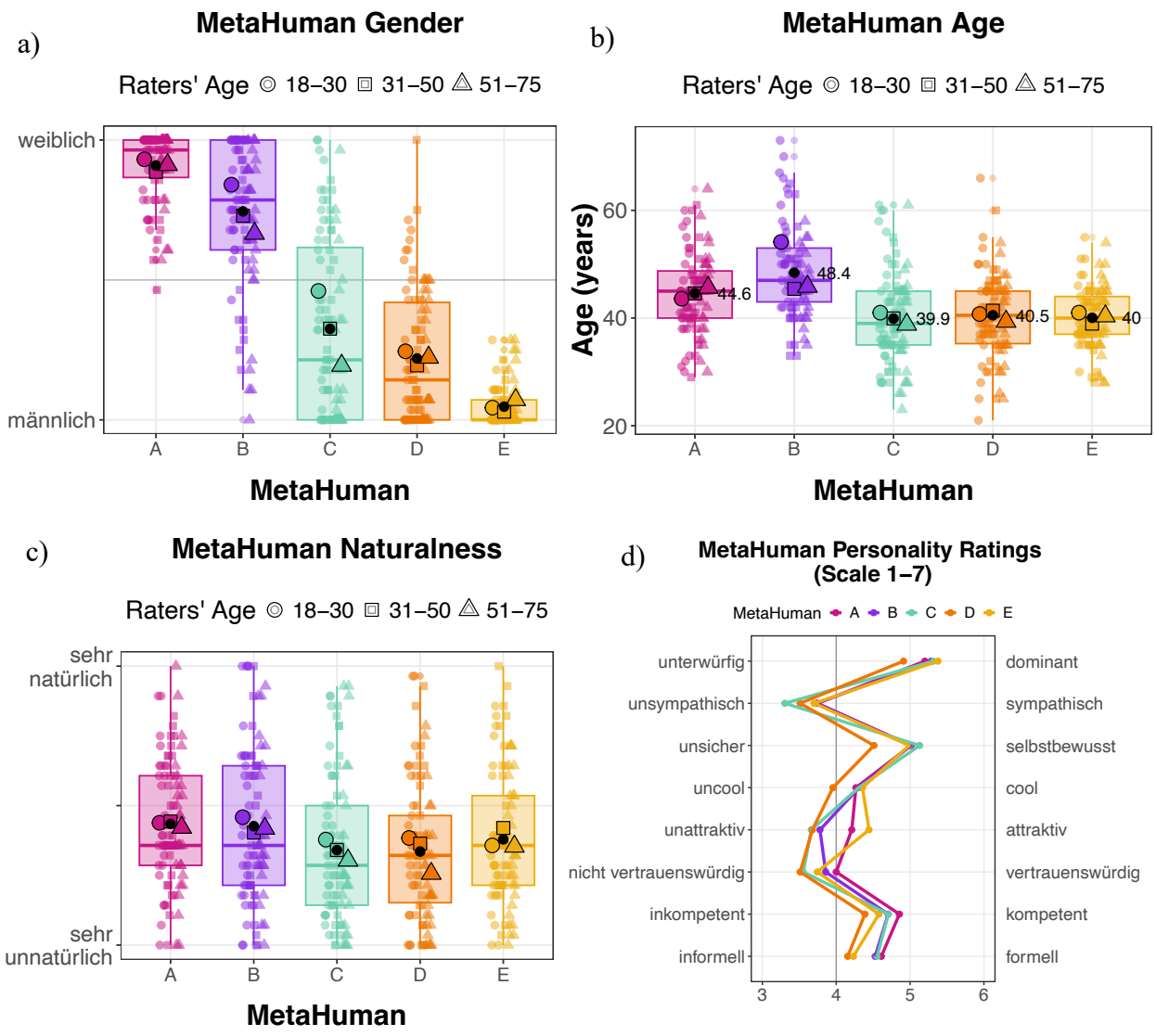



**Figure 3** – Boxplots of (a) gender, (b) age, and (c) naturalness per MetaHuman. Bars show the median, boxes mark 95*IQR, dots outside the whiskers are outliers. Light-colored dots, triangles, and squares present individual participant ratings, solid-colored dots the means per raters' age group. Black dots (and annotations) show the mean per MetaHuman for all raters. Plot d) presents mean personality ratings per MetaHuman. All plots display the original rating parameters in German.

However, variability in age between MetaHumans was within the range of the age group we aimed for in creation (31-50). The analyses of age differences between MetaHumans pointed to a possible interaction between perceived MetaHuman age and perceived MetaHuman gender, with the more feminine MetaHumans being perceived older than the more masculine ones. This possible interaction will be further examined in future analyses. Significant differences in ratings of gender and age pointed to a high importance of hairstyle for gender and age perception. Naturalness was overall lower than expected and rather on the unnatural side. However, keeping in mind that participants expected to see a natural person, these are acceptable.

The ratings presented in this study validate the implementation of these MetaHumans into our virtual reality scenarios as addressees with specific features. However, given the high variability in some of the parameters and the change from steady state pictures in 2D to animated characters in 3D in the final experimental setup, post-hoc ratings with the same questionnaire will be gathered after experimental testing in virtual reality. Those ratings of perceived MetaHuman *age*, *gender*, and *formality* will then be set in relation to the participants' speech to evaluate changes in fine phonetic detail associated with speech register.

## References


[1] Sumner, M., et al., The socially weighted encoding of spoken words: A dual-route approach to speech perception. Frontiers in psychology, 2014. **4**: p. 1015.

[2] Mortensen, L., A.S. Meyer, and G.W. Humphreys, Age-related effects on speech production: A review. Language and Cognitive Processes, 2006. **21**(1-3): p. 238-290.

[3] Bóna, J., Temporal characteristics of speech: The effect of age and speech style. The Journal of the Acoustical Society of America, 2014. **136**(2): p. EL116-EL121.

[4] Weirich, M., et al. *The phonetics of gender in Swedish and German*. in *Fonetik 2019, Stockholm, Sweden, 10-12 June, 2019*. 2019.

[5] Labov, W., Principles of Linguistic Change. Volume II: Social Factors, chapter 10. Social Networks, 2001: p. 325-365.

[6] Eckert, P. and W. Labov, Phonetics, phonology and social meaning. Journal of Sociolinguistics, 2017. **21**(4): p. 467-496.

[7] Bell, A., Language style as audience design. Language in society, 1984. **13**(2): p. 145-204.

[8] Kiesling, S.F., Style as stance. Stance: sociolinguistic perspectives, 2009. **171**.

[9] Koppen, K., M. Ernestus, and M.v. Mulken, The influence of social distance on speech behavior: Formality variation in casual speech. Corpus Linguistics and Linguistic Theory, 2019. **15**(1): p. 139-165.

[10] Hay, J. and P. Foulkes, The evolution of medial /t/ over real and remembered time. Language, 2016. **92**(2): p. 298-330.

[11] Pescuma, V.N., et al., Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods. Frontiers in psychology, 2023. **13**: p. 964658.

[12] Kutlu, E., et al., Does race impact speech perception? An account of accented speech in two different multilingual locales. Cognitive Research: Principles and Implications, 2022. **7**(1): p. 7.

[13] Jannedy, S. and M. Weirich, Sound change in an urban setting: Category instability of the palatal fricative in Berlin. Laboratory Phonology, 2014. **5**(1): p. 91-122.

[14] Duran, D., M. Weirich, and S. Jannedy. *Assessing Register Variation In Local Speech Rate*. in *Proceedings of the 20th international Congress of Phonetic Sciences*. 2023. Prague: Guarant International.

[15] Duran, D., M. Weirich, and S. Jannedy, Experimental assessment of phonetic variation in situated interaction. Register Studies, under review.

[16] D'Onofrio, A., Personae in sociolinguistic variation. Wiley interdisciplinary reviews: Cognitive science, 2020. **11**(6): p. e1543.

[17] Diethold, J.M., M. Watzlawik, and R.R. Hornstein, Die Erfassung von Geschlecht. Diagnostica, 2022.

[18] Kinsey, A.C., W.R. Pomeroy, and C.E. Martin, Sexual behavior in the human male. American journal of public health, 2003. **93**(6): p. 894-898.

[19] Kachel, S., M.C. Steffens, and C. Niedlich, Traditional Masculinity and Femininity: Validation of a New Scale Assessing Gender Roles. Frontiers in Psychology, 2016. **7**.

# INTONATION STYLE, MULTILINGUALISM AND CODE-SWITCHING IN CLASSROOM DISCOURSE: EVIDENCE FROM KENYAN ENGLISH AND SWAHILI

*Billian K. Otundo*[1] *& Simon Wehrle*[2]

[1]*Department of English Linguistics, University of Bayreuth, billian.otundo@uni-bayreuth.de,*

[2]*IfL-Phonetik, University of Cologne, simon.wehrle@uni-koeln.de*

**Abstract:** In this contribution, we analyse the intonation of classroom discourse in a rural Kenyan primary school. Our pilot corpus comprises 30 utterances from a grade 4 English language lesson recorded in Bungoma County. Although English is the prescribed medium of instruction from grade 4 onwards, the oscillation between English and Swahili is inevitable in this multilingual country, with speakers seamlessly code-switching to communicate effectively. Kenyan English and Swahili intonation exhibit unique patterns, influenced by the respective phonological systems of local languages (usually the L1), resulting in distinctive pitch and stress patterns distinct from most mainstream English varieties. For the current analysis, we quantified intonation style using the innovative measures of wiggliness (capturing pitch dynamics) and spaciousness (capturing pitch excursions) to account for the dynamic and flexible quality of intonation style, which is predicted to vary in English, Swahili, and code-switching. Our results reveal a clear tendency towards more melodic speech in English compared to Swahili (with higher values for both wiggliness and spaciousness) as well as more intonational variability in English, with code-switch utterances falling between the English and Swahili patterns. The current study not only provides new results on two language varieties that have been understudied in prosody research to date but does so using examples from real-world classroom discourse (as opposed to the monologic laboratory speech that typifies most related research). Our findings provide insight into the intricacies of speaking style and language use in multilingual communities and emphasise the importance of intonation in fostering learning and supporting linguistic diversity.

**Index terms:** classroom discourse; code-switching; intonation; Kenyan English; Swahili; wiggliness and spaciousness

## 1 Introduction

The study of intonation in multilingual settings provides insights into the intertwinement of language, prosody, and communication practices, particularly in the context of English and Swahili (e.g., [1]). Intonation styles using innovative measures of wiggliness (capturing pitch dynamics) and spaciousness (capturing pitch excursions) highlight their dynamic and flexible nature (see [2], chapter 3). In this contribution, we analyse the intonation patterns of classroom discourse in a rural Kenyan primary school. Kenya boasts approximately 68 living mother tongues [3] and uses both Kenyan English (English) and Kenyan Swahili (Swahili) as co-official languages and linguae francae, with a widely used urban mixed variety, Sheng, also in widespread use. Code-switching in this rich linguistic landscape is, thus, inevitable. Code-switching has been defined as the "juxtaposition of passages of speech belonging to two

different grammatical systems or subsystems within the same exchange" [4, pp. 1] or simply "the use of two or more languages in the same conversation" [5, pp. vii]. In this study, we specifically focus on a grade 4 English language lesson recorded in Bungoma County with the motivation that it is precisely at this level of education that the language of instruction changes to English from the language of the catchment area (usually the mother tongue) or the lingua franca (Swahili). As a result, code-switching between English and these languages is prevalent in classroom discourse, reflecting the complex multilingual nature of Kenyan society more generally. Code-switching, the alternation between languages within a single discourse, is a common phenomenon in this context, serving as a pragmatic tool to facilitate understanding and effective communication [5]. The main focus of this research is to capture the nuanced intonational patterns across three categories, English, Swahili, and instances of code-switching, by analysing wiggliness and spaciousness, which reflect the dynamics and pitch excursions of speech, respectively [2]. These measures allow for a comprehensive analysis of the variability in intonation style, as detailed in previous work on e.g. first-language acquisition in a rural German school [2], [6]. We focus on two related research questions:

- (RQ1) How is intonation style realised in multilingual classroom discourse in English and Swahili?
- (RQ2) How distinct are the intonation styles in English and Swahili and what happens in code-switching?

Our study fills an important gap in prosody research by focusing on a multilingual classroom recorded in a naturalistic, interactive setting, as opposed to the monologic laboratory speech typical for the field. The results underscore the importance of considering intonation as a key element of language use and pedagogical practice in multilingual contexts, where fostering linguistic diversity and effective communication is essential for educational success.

## 2 Prosodic Characteristics in Kenyan English, Kenyan Swahili, and Code-Switching

Kenyan English, like other post-colonial Englishes, is characterised by intonation patterns that diverge from those of standard British or American varieties. Research on the prosodic characteristics of the varieties of English, Swahili and local languages spoken in Kenya is scant [7], [8], [9], [10]. The available literature indicates that Kenyan speakers of English show less durational reduction of unstressed syllables compared to most mainstream English varieties [7]. Moreover, Kenyan English influences the first language (L1) pool in both reading and speech in interview sessions [8], where, for instance, statements predominantly have either a nuclear rise-fall (L*+H L%) or fall (H* L%) in both groups, depending on the presence or absence of prenuclear material. In [10], it is shown that the acrolectal black Kenyan English vowel system not only reveals L1 influence but also a trend towards a seven-vowel system with two front and two back mid vowels. Moreover, studies on Swahili have concentrated on the prosody of the Tanzanian variety rather than that of Kenya [11], [12], [13], [14], [15]. Ashton [12] addressed the intonation of different sentence types, and Maw and Kelly [13] transcribed natural dialogue, and gave a brief on intonation. In [14], Hyman discovered that Tanzanian Swahili, like other Bantu languages, had 'boundary narrowing', which occurs when a phonological phrase boundary follows the focused element. The other portions of the sentence are broken down into discrete phonological phrases, with no prosodic equivalent of accent reported on the main phrase. The limited literature on the prosody of Kenyan Swahili demonstrates that changes in word order alter sentence intonation for the Nairobi dialect rather than that spoken in Tanzania [15], but there were no examples of intonation alone being used to mark focus. Generally, stress in Swahili in wider East Africa is manifested by pitch alone [16]. But what happens in English–

Swahili code-switching? The complex prosodic adjustments in English–Swahili code-switching have only recently been explored (e.g., [1]). Otundo and Grice [1] found terminal falling intonation in advice-giving on radio phone-in programmes. However, whilst the global pitch contours in Kenyan English follow a marked downtrend for expressing advice in imperative, declarative and conditional forms (interpreted as a downstepping sequence of H* accents), those in Swahili have alternating rises and falls, suggesting a more elaborate intonational phonology [1]. In instances of code-switching in the same context, imperative forms of advice generally reveal alternating rises and falls, and a similar pattern is also found in declarative and conditional forms, although with an expanded pitch range [1].

# 3 Intonation Style, Multilingualism and Code-Switching in Classroom Discourse

In multilingual contexts, the way languages are used and alternated plays a crucial role in communication and learning. Kenya, with its rich linguistic diversity, provides an environment for examining how multiple languages interact within educational settings. English and Swahili are both official languages of instruction in Kenya, with English being the primary medium from grade 4 onwards, and Swahili often serving as a bridge in earlier grades and informal interactions. In grade 4 classrooms, where learners are not fully proficient in the language of instruction, intonation becomes an even more crucial tool for effective communication and comprehension. The prosodic features of a language - such as pitch range, rhythm, and stress - can either facilitate or hinder understanding, especially for students navigating multiple languages. Intonation, the variation of pitch in spoken language, is a key component of prosody and conveys meaning beyond the literal words spoken. It can indicate questions, emphasise certain points, and express emotions or attitudes [17]. In Kenyan classrooms, code-switching between English and Swahili has proven to be an effective instruction and learning approach used when learners fail to communicate through the medium of instruction [18] and to aid learners in understanding the concepts being taught (e.g., [19] [20]). Perceptions of and attitudes towards code-switching in classroom discourse are generally favourable (e.g., [21] [22]). While the linguistic aspects of code-switching in general have been extensively studied, its prosodic features in Kenyan schools remain unexplored. Understanding the intonational patterns associated with code-switching can shed light on how teachers and students navigate between languages and adapt their speech to meet learning needs.

# 4 Methodology

## 4.1 Data Collection

The data for this study were collected from a rural primary school in Bungoma County, Kenya. We recorded a grade 4 English language lesson, to capture naturalistic classroom interactions. The lesson, which lasted approximately 40 minutes, was conducted by a native Kenyan teacher and involved 30 students. The recording was made using a high-quality digital audio recorder placed unobtrusively in the classroom to minimise disruption and ensure the natural flow of discourse. Although the resulting single-channel recording includes some background noise, the quality was verified to be sufficient for intonational analysis (f0 is generally comparatively robust to noise and compression; cf. [23]). From this session, we extracted a pilot corpus of 30 utterances made by the teacher, focusing on segments that included instances of English, Swahili, and/or code-switching.

### 4.2 Annotation and Segmentation

The recorded lesson was transcribed and segmented into individual utterances using *Praat* [24]. Each utterance was annotated for language use (English, Swahili, or code-switching) and categorised based on its function within the lesson (e.g., explanation, statement). Code-switching utterances were identified as segments where a switch from one language to another occurred within a single turn or sentence (in this context, switches between English and Swahili in either direction). We independently made 3 language categories (English, Swahili, and code-switching) after extraction of the utterances, annotated (created textgrids and added metadata text to the corresponding sound) and verified the data to ensure reliability.

### 4.3 Analysis

To analyse the intonation patterns, we focused on two prosodic measures: wiggliness and spaciousness. Wiggliness is operationalised as the amounts of times a fundamental frequency (F0) contour changes direction in a given unit of speech [2]. It measures the time-varying dynamics of pitch in the form of slop changes per second [2]. Spaciousness captures pitch excursions in the form of the largest F0 rises and falls, measured in semitones [2]. These measures were calculated using a custom script in *Praat* that extracted pitch contours from each utterance and applied smoothing techniques to reduce micro-intonational fluctuations. The mean values for wiggliness and spaciousness were then computed for each language category. Then we applied descriptive statistics to compare the average values of wiggliness and spaciousness across the three language categories: English, Swahili, and code-switching. The results were visualised using scatter plots (see respective figures in the findings) with error bars representing the standard deviation of the mean values for each language category. The plots display the distribution of wiggliness and spaciousness, providing a clear visual comparison of intonation styles across English, Swahili, and code-switching utterances. All data were anonymised.

## 5 Findings

Analysing intonation patterns in the multilingual classroom revealed distinct prosodic characteristics for English, Swahili, and utterances involving code-switches. The findings are discussed in detail below, with reference to the visualisation in respective figures.

### 5.1 Intonational Characteristics of English Utterances

English utterances were characterised by the highest overall values for both wiggliness and spaciousness, indicating more dynamic and extended pitch contours compared to Swahili. The mean wiggliness value for English was 3.43 (SD = 1.11); mean spaciousness was 5.53 (SD = 1.65). These relatively high values suggest that English, as spoken in this context, exhibits a melodious quality, with frequent pitch movements that cover a broader range. This is consistent with previous studies that have noted the influence of English prosody, particularly its tendency towards larger pitch excursions in declarative and interrogative forms [17], even in the Kenyan English variety [1]. The higher wiggliness in English is particularly notable in segments where the teacher provided explanations or asked questions with an example in Figure 1 that shows a wiggliness value of 5.06 and spaciousness of 6.92.

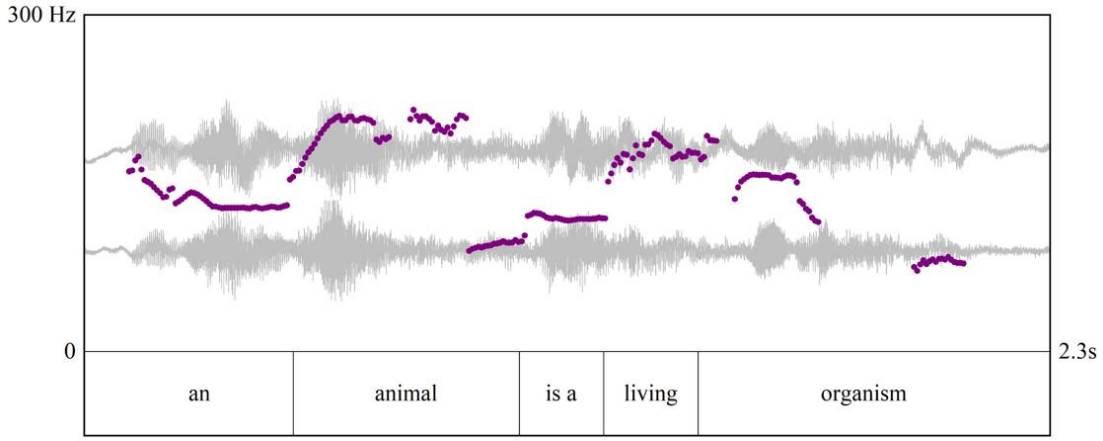



***Figure 1**: Intonation style in English, example sentence "An animal is a living organism" (wiggliness = 5.06), spaciousness = 6.92).*

These utterances often displayed a rise-fall intonation pattern, which may serve to engage students and maintain classroom dynamics. The use of expansive pitch movements in this context could also reflect an effort to accommodate the linguistic diversity of the students.

## 5.2 Intonational Characteristics of Swahili Utterances

Swahili utterances showed the lowest overall values for both prosodic measures, with a mean wiggliness value of 3.01 (SD = 0.91), and a mean spaciousness value of 4.78 (SD = 0.8). These values indicate a more constrained pitch range and fewer pitch movements (relative to utterance length) compared to English (see example in Figure 2).

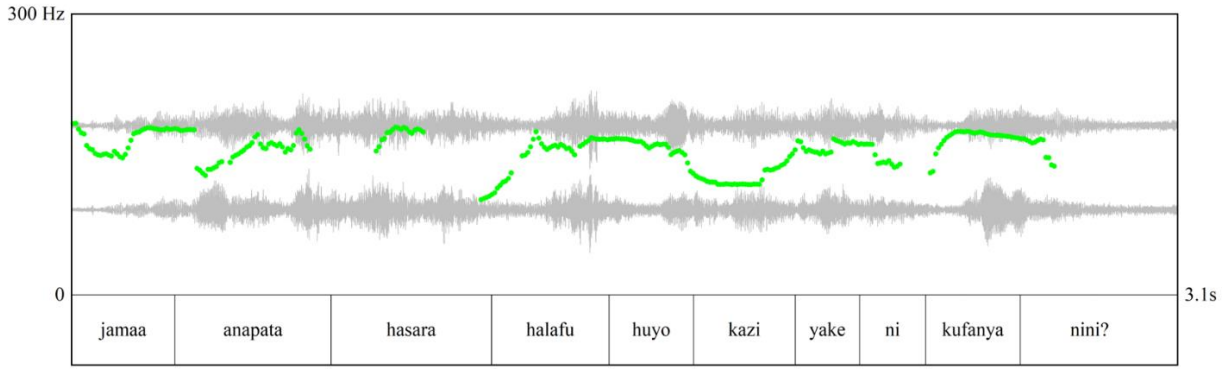



***Figure 2**: Intonation style in Swahili, "Jamaa anapata hasara halafu huyo kazi yake ni kufanya nini?" (The guy is incurring losses - then what is his job?) (wiggliness = 1.54, spaciousness = 4.13).*

The reduced pitch variation in Swahili may reflect the structural characteristics of the language, which typically utilises a narrower pitch range. It is also possible that the context of a formal classroom setting, where Swahili is often used for explanations and directives, contributes to a more conservative prosodic style.

## 5.3 Intonational Characteristics of Code-Switching Utterances

Code-switching utterances displayed intermediate values for both wiggliness and spaciousness, with mean values of 3.13 (SD = 0.95) and 5.21 (SD = 1.29), respectively. This suggests that

when speakers alternate between English and Swahili within a single utterance, the intonation patterns blend elements from both languages (see Figure 3).

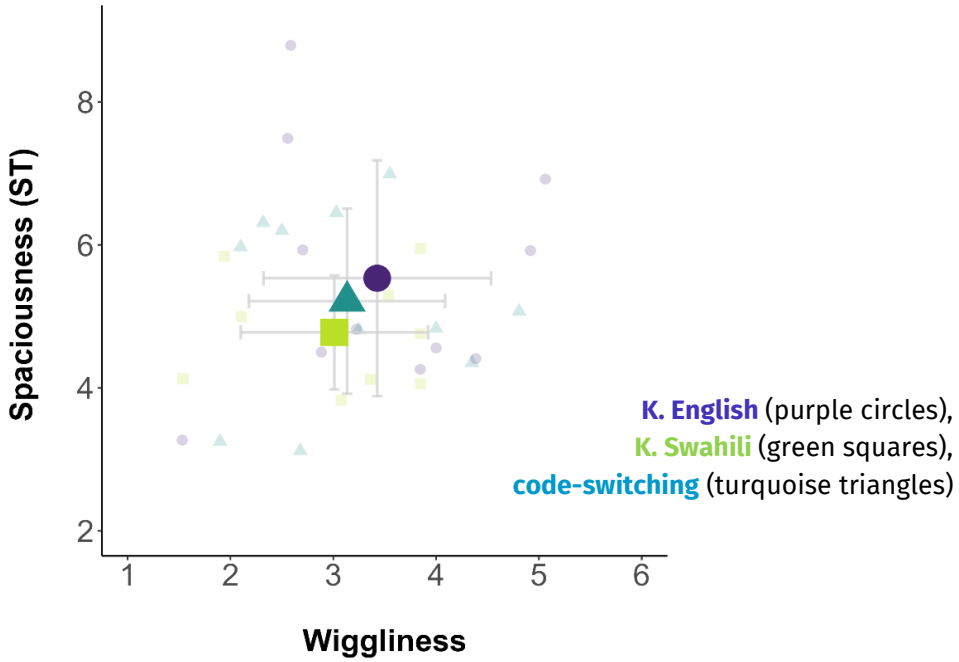



***Figure 3**: Intonation style in English (purple circles), Swahili (green squares) and code-switch utterances (turquoise triangles). Wiggliness on the x-axis, spaciousness (in semitones) on the y-axis. Large icons represent mean values; error bars represent standard deviations; small symbols represent individual utterances.*

The resulting prosody is neither as dynamic as English nor as restrained as Swahili but instead exhibits a distinct hybrid pattern. These results suggest that each language category is associated with a distinct intonational profile. English is characterised by high intonational variability and pitch excursions, Swahili by a more controlled and limited pitch range, and code-switching by a flexible, adaptive prosodic style that incorporates elements from both languages.

### 5.4 Pedagogical Implications

These results have significant implications for language instruction in multilingual schools. The prosodic diversity seen in English may aid comprehension and engagement, especially in situations where pupils are not entirely adept in the language of instruction. This is because English is introduced as the medium of instruction at the grade 4 level, while in previous years, it was taught solely as an additional subject. Before this, in grades 3 and below, the designated medium of instruction is typically the mother tongue or, occasionally, Swahili, rather than English. Dynamic intonation could help highlight important information, distinguish between different utterances, and keep students' attention. Conversely, Swahili's more regular intonation may be useful for giving clear and authoritative commands, decreasing any difficulties in understanding. Code-switching, with its adaptable intonation, appears as an effective strategy for closing the prosodic gap between English and Swahili. Teachers can easily transition between languages by modifying pitch and stress patterns, improving understanding and encouraging linguistic integration. This flexibility emphasises the pedagogical value of code-switching in multilingual educational contexts, where it can serve not only as a linguistic bridge but also as a prosodic one.

### 5.5 Limitations

Although this study is based on a small, pilot dataset of 30 utterances, which may not fully capture the range of intonational variation present in Kenyan classrooms, it breaks the ground for future research in this area. Thus, an expansion of the dataset to include a wider variety of classroom settings to enhance the generalisability of the findings is warranted. Additionally, while wiggliness and spaciousness provide useful measures of intonation, they do not capture all aspects of prosodic variation, such as rhythm or speech rate, which should be explored in subsequent studies.

## 6 Conclusion

The current study provides new insight into the prosodic aspects of English, Swahili, and code-switching in a Kenyan classroom setting. Our analysis quantifies intonation style using the measures of wiggliness (capturing pitch dynamics) and spaciousness (capturing pitch excursions). We found a clear tendency towards more melodic speech in English compared to Swahili (with higher values for both wiggliness and spaciousness) as well as more intonational variability in English, whilst code-switch utterances fall between English and Swahili patterns. In sum, Kenyan English is characterised by high intonational variability and pitch excursions, Kenyan Swahili by a more controlled and limited pitch range, and code-switching by a flexible, adaptive prosodic style that incorporates elements from both languages. The varied intonational patterns identified for each language category emphasise the adaptive use of prosody in multilingual speech, and the importance of intonation as a communication resource in education. Future research should build on these preliminary findings by including a larger number of speakers, utilising more extensive datasets and exploring a broader range of linguistic and educational situations in order to gain comprehensive knowledge of prosody in multilingual education.

## Acknowledgements

This research was conducted as part of a pilot study for the larger project, “Multilingual African Learning Spaces: Translanguaging in Kenyan Schools” of the Africa Multiple Cluster of Excellence at the University of Bayreuth, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2052/1 – 390713894., and by the DFG as part of the SFB 1252 “Prominence in Language” at the University of Cologne. Special thanks to all the participants of this study and Caleb Masinde for assistance with the annotation.

## References

[1] OTUNDO, B. K. and M. GRICE: Intonation in advice-giving in Kenyan English and Kiswahili. In: Proceedings of the 11th International Conference on Speech Prosody, S. 150–154. Lissabon, 2022. DOI: 10.21437/SpeechProsody.2022-31 [Accessed Oct. 12, 2024].

[2] WEHRLE, S.: Conversation and intonation in autism: a multi-dimensional analysis. Berlin: Language Science Press, 2023.

[3] EBERHARD, D. M., G. F. SIMONS, and C. D. FENNING, Eds.: Ethnologue: Languages

of the World. Twenty-Fourth Edition. Dallas, Texas: SIL International, 2021. [Online]. Available: https://www.ethnologue.com/country/KE/languages. [Accessed Nov. 12, 2021].
[4] GUMPERZ, J. J.: The sociolinguistic significance of conversational code-switching. RELC Journal, vol. 8, no. 2, S. 1–34, 1977.
[5] MYERS-SCOTTON, C.: Social Motivations for Codeswitching: Evidence from Africa. Oxford: Oxford University Press, 1993.
[6] WEHRLE, S. and C. SAPPOK: Evaluating prosodic aspects of oral reading proficiency in schoolchildren: Effects of gender, genre, and grade. Proceedings of the 20th International Congress of Phonetic Sciences, 2023.
[7] SCHMIED, J.: East African Englishes. In: Handbook of World Englishes, B. B. KACHRU, Y. KACHRU, and C. L. NELSON, Eds. London: Blackwell Publishing, 2006, S. 188–202.
[8] OTUNDO, B. K.: Exploring Ethnically-Marked Varieties of Kenyan English: Intonation and Associated Attitudes. Munster: LIT Verlag, 2018.
[9] GRICE, M., J. S. GERMAN, and P. WARREN: Intonation systems across varieties of English. In: The Oxford Handbook of Language Prosody, C. GUSSENHOVEN and A. CHEN, Eds. Oxford: University Press, 2020, S. 285–302.
[10] HOFFMANN, T.: The Black Kenyan English vowel system: An acoustic phonetic analysis. English World-Wide: A Journal of Varieties of English, vol. 32, no. 2, S. 147–173, 2011.
[11] MAGHWAY, J. B.: Aspects of Prosody in English and Kiswahili. Edinburg: Edinburgh University, 1988.
[12] ASHTON, E. O.: Swahili Grammar (Including Intonation), 2nd ed. London: Longmans, 1947.
[13] MAW, J. and J. KELLY: Intonation in Swahili. London: School of Oriental and African Studies, 1975.
[14] DOWNING, L. J.: The prosody of focus in Bantu languages and the primacy of phrasing. In: TIE Conference, Santorini, Greece, 2004, S. 9–10.
[15] AUGUSTIN, M.: Topic and focus in Swahili. Gialens: Electronic Notes [Online]. Vol. 1, no. 1. Nov. 14, 2007. Available: https://diu.edu/gialens/vol1num1/.
[16] ANYANWU, R.-J.: On the manifestation of stress in African languages. In: *Typology of African Prosodic Systems* (*TAPS) 2001 - Proceedings.* Bielefeld: Bielefeld Occasional Papers in Typology (BOPIT), pp. 149-157.
[17] WELLS, J. C.: English Intonation: An Introduction. Cambridge: Cambridge University Press, 2006.
[18] OGECHI, N. O.: Trilingual Code Switching in Kenya: evidence from Ekegusii, Kiswahili, English and Sheng. (Unpublished doctoral dissertation). Hamburg: Universitat Hamburg, 2002.
[19] MISATI, M. A. and D. W. LWANGALE: Influence of code switching on mastery of English-speaking skills in secondary schools in Kwanza Sub-County Trans-Nzoia County, Kenya. Journal of Educational Research, vol. 5, no. 2, S. 76–88, 2020.
[20] NTHINGA, M. P.: Patterns and Functions of Code Switching in Pre-Primary Classroom Discourse in Selected Schools at Kasarani Division. Nairobi: Unpublished M.A. Thesis, Kenyatta University, 2003.
[21] OTUNDO, B. K.: Student Teachers' Attitudes toward Translanguaging in Formal Learning Vis-à-Vis Informal Interaction: A Survey of a University in Kenya. In: MOHR, S. and L. FERRARA (eds.), Learning Languages, Being Social, S. 231–260. De Gruyter, 2024. DOI: 10.1515/9783110794670-009. Available: https://www.degruyter.com/document/doi/10.1515/9783110794670-009/html (18 October, 2024).

[22] OTUNDO, B. K.: Communicating in Multilingual Learning Ecologies: Students' Perceptions of Translanguaging in Lecture Rooms in Kenya. Journal of Linguistics, Literary and Communication Studies, vol. 2, no. 2, S. 24–34, 2023. DOI: 10.58721/jllcs.v2i2.378. Available: https://utafitionline.com/index.php/jltcs/article/view/378 (18 October, 2024).
[23] ZHANG, C., K. JEPSON, G. LOHFINK, and A. ARVANITI: Comparing acoustic analyses of speech data collected remotely. Acoustical Society of America, 146(6), 3910-3916, 2021.
[24] BOERSMA, P. and D. WEENINK: Praat: doing phonetics by computer (Version 6.0.40) [Online]. Available: http://www.praat.org.

# ZUR QUANTITATIVEN ERFASSUNG VON AUSSPRACHEABWEICHUNGEN DER VOKALE BEI CHINESISCHEN DEUTSCHLERNENDEN UNTER BERÜCKSICHTIGUNG VON LERNKONTEXTUELLEN FAKTOREN

*Zaixi Pan*[1], *Sven Grawunder*[1,2]

[1]*Martin-Luther-Universität Halle-Wittenberg,* [2]*Max-Planck-Institut für evolutionäre Anthropologie, Leipzig*
*zaixi.pan@sprechwiss.uni-halle.de*

**Kurzfassung:** In diesem Beitrag werden Möglichkeiten einer effektiven quantitativen Analyse der Ausspracheabweichungen chinesischer Deutschlernender mit Hinblick auf die Vokalquantität und Vokalqualität untersucht. 30 Probanden (L2-Deu), deren durchschnittliches Sprachniveau der oberen Mittelstufe (B2) entspricht, wurden rekrutiert. Alle sind männliche L1-Mandarinsprecher, die aus zwei Gruppen kommen: 15 Probanden sind mindestens ein Jahr in Deutschland („Deutschland-Gruppe"). Die übrigen 15 L2-Deu-Probanden waren noch nie in Deutschland und besuchen bislang nur Deutschkurse in China („China-Gruppe"). Darüber hinaus wurden 10 Deutsch-Muttersprachler als eine Referenzgruppe akquiriert. Alle Probanden lasen einen deutschsprachigen Lesetext wiederholt. Für die Analyse der Vokalquantität wurde die Vokaldauer unter Berücksichtigung der Vokalqualität gemessen. Ein Vergleich der beiden Versuchsgruppen jeweils mit der Referenzgruppe soll Aussagen dazu bringen, welche Gruppe der Referenzgruppe mehr ähnelt, d.h. welche Gruppe weniger ‚vokalische' Ausspracheabweichungen aufweist.

## 1 Ausgangspunkt - Analyse von Ausspracheabweichungen

### 1.1 Ausspracheabweichungen

Dieser Beitrag ist Teil des Promotionsprojekts „Untersuchung zu segmentalen, suprasegmentalen und phonotaktischen Ausspracheabweichungen chinesischer Deutschlernender unter der Berücksichtigung der Faktoren Alter, Unterrichtsmethode, Dauer und Art des Aufenthalts" des Erstautors. Ziel ist es, Möglichkeiten einer effektiven quantitativen Analyse der Ausspracheabweichungen chinesischer Deutschlernender mit Hinblick auf die Vokalquantität und Vokalqualität zu untersuchen. Die Forschungen zu Ausspracheabweichungen bei chinesischen Deutschlernenden lassen sich grob in zwei Hauptbereiche unterteilen. Der erste Bereich bezieht sich hauptsächlich auf qualitative Analysen, d.h. auf vergleichende Studien der phonologischen Systeme von Chinesisch und Deutsch. Zum Beispiel klassifizierte und fasste Liu [1] die in chinesischen Deutschkursen auftretenden Ausspracheabweichungen zusammen und stellte fest, dass die großen Unterschiede zwischen den phonetischen Systemen von Chinesisch und Deutsch die Ursache für diese Abweichungen sind. Hunold [2] führte eine qualitative empirische Forschung durch, bei der sie Muttersprachler als Kontrollhörer einsetzte, um ihre Hypothesen zu verschiedenen Arten von Ausspracheabweichungen zu überprüfen. Lay [3] berücksichtigte den Einfluss des Englischen als erste Fremdsprache (L2) der meisten Chinesen auf deren Deutschlernen (L3). Li [4] richtete ihren Fokus auf den wenig erforschten Bereich der Prosodie und entwickelte eine Lernvideoserie, die sich auf die prosodischen Merkmale des Deutschen und Chinesischen konzentriert, um chinesischen Deutschlernenden beim Erlernen der prosodischen Aspekte der

deutschen Sprache zu helfen. Der zweite Bereich der Literatur beschäftigt sich mit quantitativen Analysen, d.h. mit der Anwendung von akustischen/statistischen Methoden zur Verarbeitung und Analyse akustischer Daten, um die Abweichungsgrade und Details der Aussprachefehler von chinesischen Deutschlernenden präziser zu bestimmen. Beispielsweise entwarf Ding [5] mehrere Experimente, die sich auf Aspekte wie die Sprechdauer, die Vokaltrapez-Verteilung und den Intonationsverlauf konzentrierten und verband diese mit *MOS-Scores* (*Mean Opinion Score* = 'Gesamteindruck') und anderen Wahrnehmungsbewertungssystemen, um die Unterschiede zwischen den phonetischen Systemen des Chinesischen und Deutschen zu untersuchen. Viele der bisherigen Arbeiten, deren Ausgangs- und Zielsprache nicht unbedingt Chinesisch > Deutsch ist, versuchen durch verschiedene statistische Methoden, die Ähnlichkeit zwischen den phonetischen Systemen des Chinesischen und den L2-Varietäten anderer Sprachen zu untersuchen, um mögliche Ausspracheabweichungen der Lernenden zu evaluieren (s. auch [6]). So verglich Deng [7] mittels euklidischer Distanz die phonologischen Ähnlichkeiten zwischen Chinesisch und Englisch. Feng et al. [8] verwendeten Entropie, um die Verteilung von Aussprachefehlern zu quantifizieren. Schließlich setzten Xie und Jaeger [9] eine gemischte lineare Regression ein, um den Mittelpunkt von Vokal-Kategorien und den Abstand zwischen benachbarten Vokal-Kategorien zu untersuchen. Die letzteren beiden Arbeiten sollen hier aufgegriffen werden. Insbesondere soll das Konzept der Separabilität [9](s. Abschnitt 2.4) auf seine Eignung geprüft für eine Evaluierung von L2-Sprache geprüft werden.

### 1.2 Einflussfaktoren

Es ist offensichtlich, dass viele Aspekte des L2-Sprachlernens in Bezug auf phonologische und phonetische Faktoren untersucht werden können, die mit den Ähnlichkeiten und Unterschieden zwischen den Lautsystemen der Erstsprache (L1) und der Zweitsprache (L2) zusammenhängen. Gleichzeitig wird das phonetische Lernen durch ein komplexes Zusammenspiel von Lernervariablen beeinflusst. Dazu gehören altersbedingte Entwicklungsfaktoren, Quantität und Qualität des Kontakts mit der Zweitsprache, der Gebrauch der Erst- und Zweitsprache im Laufe der Zeit, Zweitsprachunterricht sowie individuelle Unterschiede in Motivation, Begabung und emotionaler Einstellung [vgl. 10, 3]. Das *Speech Learning Model (SLM)* von Flege [11] zählt zu einer der bekanntesten Theorien dazu. Es beschreibt empirisch fundiert den Erwerb von Aussprache und Sprachwahrnehmung in einer Zweitsprache (L2) sowie den Einfluss des Alters auf diesen Prozess. Dieses Modell erklärt Alterseffekte und phonetische Kategorien bei bilingualen Sprechern, wobei die Aussprachefähigkeit der Lernenden in der Fremdsprache in engem Zusammenhang mit ihrer Lebensumgebung steht, ebenso mit der Häufigkeit der Nutzung der L2 sowie dem Alter, in dem sie mit der L2 in Kontakt gekommen sind. Im Vergleich dazu fokussiert das *Perceptual Assimilation Model (PAM)* von Best (1993,1995) die anfänglichen Wahrnehmungsstadien (erste Begegnung mit neuen Lauten) und betont, wie die Lautkategorien von L1 die Wahrnehmung von L2 beeinflussen. Sowohl das SLM als auch das PAM betrachten phonetische Ähnlichkeit als entscheidend, um den Erfolg von L2-Lernenden bei der Wahrnehmung und Unterscheidung segmentaler Kontraste vorherzusagen [12, vgl. 2-3]. Aus diesem Grund liegt der Schwerpunkt des Datenanalyseabschnitts dieser Arbeit auch auf der Berechnung der phonetischen Ähnlichkeit zwischen verschiedenen Versuchspersonengruppen. Das Promotionsprojekt berücksichtigt Einflussfaktoren wie Lernumgebung und Dauer des Aufenthalts (sowie indirekt auch Alter), die zum Lernerfolg beitragen [vgl. 13, 27-57] und anhand der beiden Gruppen (s. Abschnitt 2.1) untersucht werden.

# 2 Methoden

## 2.1 Probanden

Es wurden für das Projekt insgesamt 60 Probanden, deren durchschnittliches Sprachniveau der oberen Mittelstufe (B2) entspricht, rekrutiert. Alle sind männliche Mandarin-Muttersprachler, die aus zwei Gruppen kommen: Die Hälfte der Probanden ist mindestens ein Jahr in Deutschland und hat regelmäßige Kontakte mit Einheimischen („Deutschland-Gruppe“). Die andere Hälfte war noch nie in Deutschland und besuchte bislang nur Deutschkurse in China („China-Gruppe“). Für diese Pilot-Studie wurden die Daten von 30 Probanden der genannten 60 einbezogen, wobei 15 Probanden aus der China-Gruppe, die an der Dongbei-Universität China Germanistik studierten und bislang noch nie in Deutschland waren. Die übrigen 15 aus der „Deutschland-Gruppe“, studierten zum Zeitpunkt der Aufnahmen an der TU Dresden. Im Rahmen dieser Pilot-Studie wurden insgesamt 11 Deutsch-Muttersprachler. rekrutiert. Fünf davon arbeiteten bzw. studierten im Bereich Bildung/Ausbildung sowie in der Medienbranche in Frankfurt am Main (Referenzgruppe). Die anderen sechs waren Studierende im Fach Sprechwissenschaft an der Martin-Luther-Universität Halle-Wittenberg und fungierten als Kontrollhörer für die perzeptive/subjektive Analyse.

## 2.2 Material & Aufnahmen

Alle 30 Probanden sowie die 5 Deutsch-Muttersprachler von der Referenzgruppe wurden gebeten, einen Lesetext (Märchentext „Die Söhne“ aus Hunold [2] nach Leo Tolstoi „Die drei Söhne“) vorzulesen, der in relativ konzentrierter Form alle wichtigen segmentalen, suprasegmentalen und phonotaktischen Phänomene der deutschen Sprache beinhaltet. Die Vorlesephase wurde dann mittels ZOOM H5 HANDY RECORDER in 48 kHz, 24 bit, Mono aufgenommen. Die Aufnahmen fanden im Sprachlabor an der Dongbei-Universität China und in der Aufnahmekabine an der Goethe-Universität Frankfurt am Main statt. Jede Versuchsperson wurde zweimal aufgenommen. Im Allgemeinen wurde die jeweils zweite Aufnahme ausgewertet und nur dann auf die erste Aufnahme zurückgegriffen, wenn der entsprechende Ausschnitt der zweiten Aufnahme aufgrund eines Störgeräusches oder aber zu geringer Lautstärke, zu starker Laryngalisierung o.ä. unbrauchbar war.

## 2.3 Segmentierung, Dauer- und Formant-Messung

Unter Verwendung von PRAAT [14] und entsprechendem PRAAT-Scripting wurde die Dauer aller ausgewählten Monophthonge (/iː, ɪ, eː, ɛ, aː, a, oː, ɔ, uː, ʊ/) auf Basis einer Segmentierung in Textgrids gewonnen. Was die Formantanalyse anbelangt, wurden für jeden Probanden und für jedes einzelne Vokalphonems dieselben 5 Laute, die immer an der gleichen Position in dem Lesetext liegen, mit Hilfe von PRAAT segmentiert. Dabei wurde versucht, die ersten zwei Formanten im Mittelpunkt jedes Intervalls zu berechnen. Danach wurden die angezeigten Werte auf Plausibilität geprüft, indem die Formanten mithilfe des Spektrums visuell überprüft wurden. Die Normalisierung der Formantwerte erfolgte nach Lobanov ([15] zuletzt Adank et al. [16]) gewählt:

$$F_{n[V]}{}^{N} = (F_{n[V]} - MEAN_n)/s_n \quad (1)$$

Hierbei ist $F_{n[V]}{}^{N}$ der normalisierte Wert für den Formant $n$ des Vokals $V$. Und $MEAN_n$ ist der Mittelwert von Formant $n$ des jeweiligen Sprechers mit $s_n$, der dazugehörigen Standardabweichung. Die Verarbeitung der Messwerte erfolge neben der Aufbereitung in EXCEL weiter

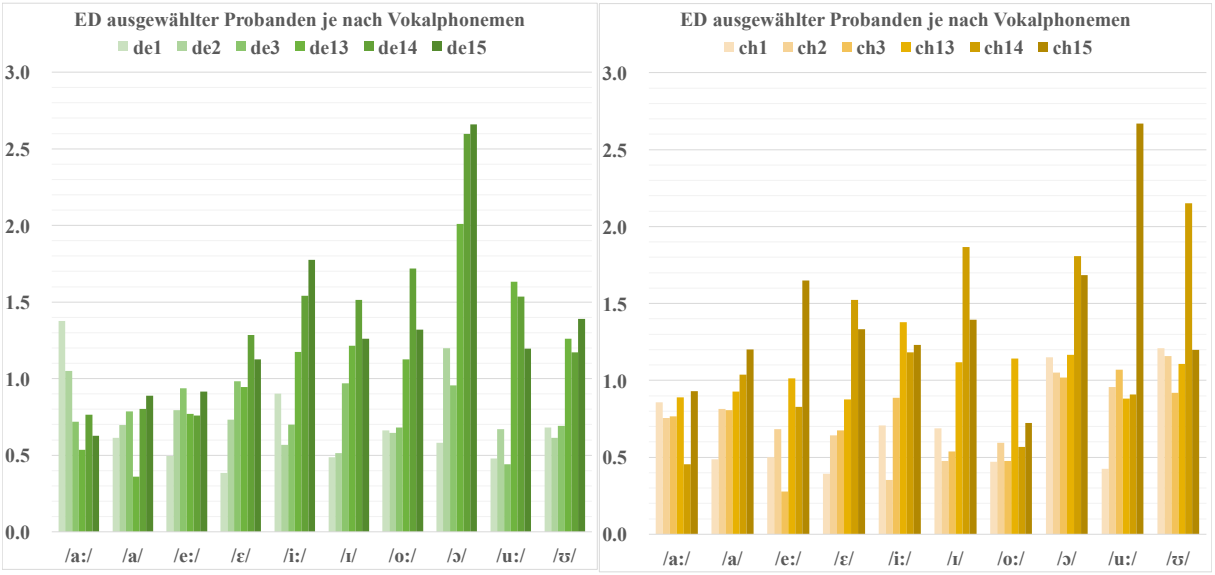



**Abbildung 1** – ED ausgewählter Probanden je nach Vokalphonemen

mit PYTHON unter Verwendung der packages *numpy*, *openpyxl*, *pandas*, *sklearn.metrics. pairwise_distances* sowie *statsmodels.formula.api*. Die Normalisierung der Dauer erfolgt hier über die Medianwerte der Gesamtsumme der gemessenen Dauern pro Sprecher.

### 2.4 Separabilität

In Anlehnung an Xie und Jaeger [9] wird die Erfassung der Variabilität einer Merkmalsausprägung, wie etwa die Vokalqualität in Form von F1/F2-Formantpositionen, als eine Möglichkeit betrachtet, um die Variabilität der Realisierung einer Kategorie mit Rücksicht auf deren phonetischen Nachbarn untersucht. Von den Autoren [9] wird dies als Separabilität (*separability*) bezeichnet, hier z.B. für die Werte von /iː/ und /ɪ/, die man sich als eine Art gemittelter Euklidischer Distanz vorstellen kann.

$$Separabilität = \frac{\sum_{k=1}^{n} \sqrt{(F1_{token\,k\,of\,/i/} - F1_{center\,of\,/I/})^2 + (F2_{token\,k\,of\,/i/} - F2_{center\,of\,/I/})^2}}{n} \tag{2}$$

## 3 Ergebnisse

### 3.1 Dauer

Im Rahmen der objektiven Evaluierung wurden für jeden Vokal 5 Tokens an der gleichen Position (Wort im Text) gemessen. Die euklidische Distanz (ED) der Vokaldauer wurde für die beiden Probanden-Gruppen jeweils zur Referenz-Gruppe berechnet. Dabei wurde auf Gruppenebene pro Gruppe der Medianvektor mit dem Medianvektor der Referenzgruppe verglichen. Dies passiert ebenso auf Ebene der Sprecher und Vokale. Für die Darstellung der individuellen ED-Tendenz wurden 6 Probanden von jeder Gruppe ausgewählt, wobei 3 die wenigsten und 3 die meisten ED-Punkte aufwiesen (s. Abb. 1).

Die Ergebnisse zeigten insgesamt eine leicht geringere euklidische Distanz bei der China-Gruppe (2.143) im Vergleich zur Deutschland-Gruppe (2.148). Besonders bei den Vokalpaaren /iː/ – /ɪ/, /oː/-/ɔ/ sowie /uː/-/ʊ/ lassen sich sowohl bei der China-Gruppe als auch bei der Deutschland-Gruppe relativ größere Abweichungen in der Dauer beobachten. Dies deutet dar-

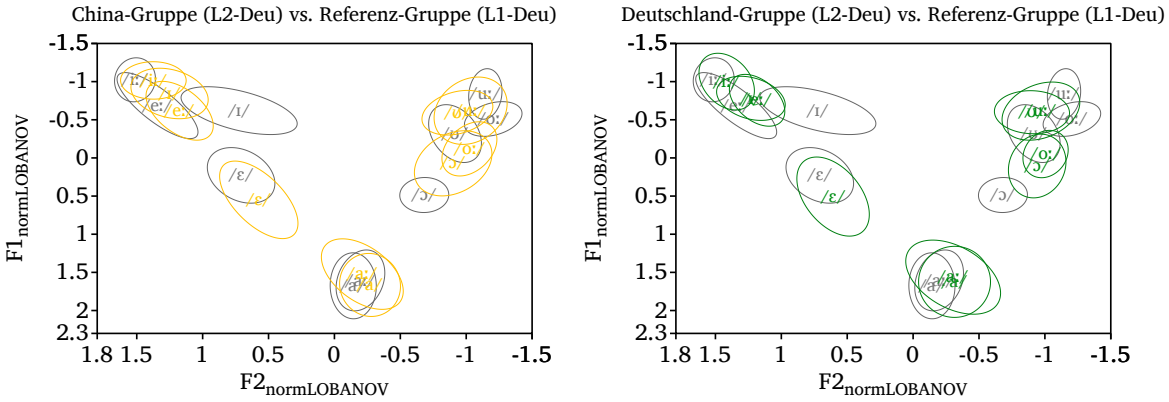



**Abbildung 2** – Vergleich der gemessenen Formantwerte für je für China-Gruppe vs. Deu-L1-Referenzgruppe (links) und Deutschland-Gruppe vs. Deu-L1-Referenz-Gruppe (rechts)

auf hin, dass einerseits die China-Gruppe in Bezug auf die Vokaldauer näher an der Referenz-Gruppe liegt als die Deutschland-Gruppe, wobei andererseits bei bestimmten Vokalpaaren in beiden Gruppen Abweichungen von der Referenz bemerkbar sind. Insbesondere in der D-Gruppe wurden den Vokalpaaren /iː/-/ɪ/, /oː/-/ɔ/ sowie in der C-Gruppe /uː/-/ʊ/ größere ED verzeichnet (s. Abb. 1).

## 3.2 Formantanalyse

Bezüglich der Formantanalyse werden in dieser Pilot-Studie hauptsächlich zwei Aspekte mit Bezug auf die Studie von Xie und Jaeger [9] untersucht, nämlich die Kategoriemittelwerte und die Separabilität benachbarter Vokalkategorien.

### 3.2.1 Kategoriemittelwerte

Um den Vergleich der Kategoriemittelwerte von L1- und L2-Sprechern (s. Abb. 2) vorzunehmen, wurde eine gemischte lineare Regression angewandt, basierend auf 630 Datenpunkten, die aus 35 Sprechern mit zwei verschiedenen Akzenten und neun Vokalen resultieren (35 Sprecher x 2 Akzente x 9 Vokale) (s. Abb. 3). Als feste Faktoren wurden VOKALE, AKZENT und deren Wechselwirkung sowie als zufälliger Faktor SPRECHER berücksichtigt. Die Ergebnisse zeigten Unterschiede in den Mittelwerten fast aller Vokalkategorien in mindestens einer Formantendimension zwischen der Referenzgruppe und den beiden Probanden-Gruppen. Lediglich bei den Vokalen /aː/, /eː/ und /iː/ waren die Unterschiede nicht signifikant (p-Wert $>$ 0,05). Im Allgemeinen zeigten beide Probanden-Gruppen höhere F1- sowie F2-Werte im Vergleich zur Referenzgruppe, mit Ausnahme von /ɔ/ und /ɛ/.

Unterschiede der Mittelwerte zwischen der Referenzgruppen und den beiden Probanden-Gruppen bestehen in fast allen Vokalkategorien in mindestens einer Formantendimension, ausgenommen /aː/, /eː/ und /iː/. Generell gibt es in beiden Probanden-Gruppen höhere F1 sowie F2-Werte als in der Referenzgruppe (abgesehen von /ɔ/ und /ʊ/). Somit verbleiben Unterschiede zwischen den beiden Probanden-Gruppen.

### 3.2.2 Separabilität benachbarter Vokalkategorien

In Anlehnung an Xie und Jaeger [9] sowie Wedel et al. [17] wird die Separabilität benachbarter Vokalkategorien in dieser Studie untersucht. Hierbei werden die durchschnittlichen F1- und F2-Werte jedes Vokalpaars, abgesehen vom Vokalpaar /aː/ und /a/, als Kategoriezentren herangezogen.

**Mixed-effects models: Category means (NN - Deutschland-Gruppe; N - Referenz-Gruppe)**

| | Formanten | Coefficient | SE | t_value | p_value | △μ (NN-N) |
|---|---|---|---|---|---|---|
| **a**: | **F1** | 0.00 | 0.06 | 0.07 | 0.94 | |
| | **F2** | 0.03 | 0.04 | 0.78 | 0.44 | |
| **e**: | **F1** | -0.01 | 0.06 | -0.14 | 0.89 | |
| | **F2** | 0.01 | 0.04 | 0.16 | 0.87 | |
| **ɛ** | **F1** | 0.16 | 0.06 | 2.63 | 0.01 | Higher F1 |
| | **F2** | 0.04 | 0.04 | 0.85 | 0.40 | |
| **i**: | **F1** | 0.04 | 0.06 | 0.60 | 0.55 | |
| | **F2** | 0.04 | 0.04 | 0.87 | 0.38 | |
| **ɪ** | **F1** | -0.05 | 0.06 | -0.87 | 0.38 | |
| | **F2** | 0.34 | 0.04 | 8.12 | 0.00 | Higher F2 |
| **o**: | **F1** | 0.26 | 0.06 | 4.08 | 0.00 | Higher F1 |
| | **F2** | 0.19 | 0.04 | 4.44 | 0.00 | Higher F2 |
| **ɔ** | **F1** | -0.16 | 0.06 | -2.62 | 0.01 | Lower F1 |
| | **F2** | -0.04 | 0.04 | -1.05 | 0.29 | |
| **u**: | **F1** | 0.14 | 0.06 | 2.21 | 0.03 | Higher F1 |
| | **F2** | 0.17 | 0.04 | 4.15 | 0.00 | Higher F2 |
| **ʊ** | **F1** | -0.11 | 0.06 | -1.82 | 0.07 | |
| | **F2** | 0.09 | 0.04 | 2.21 | 0.03 | Higher F2 |

**Mixed-effects models: Category means (NN - China-Gruppe; N - Referenz-Gruppe)**

| | Formanten | Coefficient | SE | t_value | p_value | △μ (NN-N) |
|---|---|---|---|---|---|---|
| **a**: | **F1** | -0.04 | 0.06 | -0.57 | 0.57 | |
| | **F2** | 0.06 | 0.04 | 1.32 | 0.19 | |
| **e**: | **F1** | 0.04 | 0.06 | 0.63 | 0.53 | |
| | **F2** | -0.02 | 0.04 | -0.47 | 0.64 | |
| **ɛ** | **F1** | 0.17 | 0.06 | 2.65 | 0.01 | Higher F1 |
| | **F2** | 0.00 | 0.04 | -0.08 | 0.94 | |
| **i**: | **F1** | 0.01 | 0.06 | 0.16 | 0.88 | |
| | **F2** | -0.01 | 0.04 | -0.14 | 0.89 | |
| **ɪ** | **F1** | -0.11 | 0.06 | -1.71 | 0.09 | |
| | **F2** | 0.33 | 0.04 | 7.62 | 0.00 | Higher F2 |
| **o**: | **F1** | 0.20 | 0.06 | 3.22 | 0.00 | Higher F1 |
| | **F2** | 0.15 | 0.04 | 3.58 | 0.00 | Higher F2 |
| **ɔ** | **F1** | -0.20 | 0.06 | -3.23 | 0.00 | Lower F1 |
| | **F2** | -0.05 | 0.04 | -1.06 | 0.29 | |
| **u**: | **F1** | 0.12 | 0.06 | 1.85 | 0.06 | |
| | **F2** | 0.12 | 0.04 | 2.78 | 0.01 | Higher F2 |
| **ʊ** | **F1** | -0.13 | 0.06 | -2.10 | 0.04 | Lower F1 |
| | **F2** | 0.05 | 0.04 | 1.27 | 0.20 | |

**Abbildung 3** – *Mixed-effects models*: Kategorienmittelwerte

Es wird der durchschnittliche Abstand der Vokale eines Sprechers zur mittleren Position („Vokalzentrum“) der benachbarten Vokalkategorie berechnet (s. Formel 2). Die Anwendung einer gemischten linearen Regressionen, wobei als *fixed factors* AKZENT (L1 vs. L2 bzw. RefGr vs. DGr/CGr), VOKALkategorie und als *random factor* SPRECHER dienen.

Die Ergebnisse zeigten, dass bei den Vokalpaaren /iː/-/ɪ/ sowie /oː/-/ɔ/ eine höhere Separabilität in der Referenzgruppe im Vergleich zu den beiden Probanden-Gruppen beobachtet wurde (s. Abb. 4). Für die Vokalpaare /eː/-/ɛ/ und /uː/-/ʊ/ wurden keine signifikanten Unterschiede in der Separabilität zwischen der (nativen, N) Referenzgruppe und den (nicht-nativen, NN) Probanden-Gruppen festgestellt. Es konnten jedoch keine eindeutigen Unterschiede in der Separabilität innerhalb der beiden Probanden-Gruppen durch das verwendete Modell ermittelt werden. Dies offenbart eine Forschungslücke, die in zukünftigen Studien speziell adressiert und untersucht werden sollte.

**Mixed-effects models: Category Separability (NN- Deutschland-Gruppe; N- Referenz-Gruppe)**

| Vowel | Coefficient | SE | t_value | p_value | Comparison | △separability(NN-N) |
|---|---|---|---|---|---|---|
| **e**: | 0.13 | 0.09 | 1.47 | 0.14 | **e**: → **ɛ center** | |
| **ɛ** | 0.12 | 0.09 | 1.42 | 0.16 | **ɛ** → **e**: **center** | |
| **i**: | -0.24 | 0.09 | -2.79 | 0.01 | **i**: → **ɪ center** | N 〉 NN |
| **ɪ** | -0.25 | 0.09 | -2.92 | 0.00 | **ɪ** → **i**: **center** | N 〉 NN |
| **o**: | -0.39 | 0.09 | -4.54 | 0.00 | **o**: → **ɔ center** | N 〉 NN |
| **ɔ** | -0.36 | 0.09 | -4.13 | 0.00 | **ɔ** → **o**: **center** | N 〉 NN |
| **u**: | -0.11 | 0.09 | -1.32 | 0.19 | **u**: → **ʊ center** | |
| **ʊ** | -0.12 | 0.09 | -1.43 | 0.15 | **ʊ** → **u**: **center** | |

**Mixed-effects models: Category Separability (NN- China-Gruppe; N- Referenz-Gruppe)**

| Vowel | Coefficient | SE | t_value | p_value | Comparison | △separability(NN-N) |
|---|---|---|---|---|---|---|
| **e**: | 0.09 | 0.08 | 1.14 | 0.25 | **e**: → **ɛ center** | |
| **ɛ** | 0.09 | 0.08 | 1.11 | 0.27 | **ɛ** → **e**: **center** | |
| **i**: | -0.29 | 0.08 | -3.55 | 0.00 | **i**: → **ɪ center** | N 〉 NN |
| **ɪ** | -0.29 | 0.08 | -3.51 | 0.00 | **ɪ** → **i**: **center** | N 〉 NN |
| **o**: | -0.39 | 0.08 | -4.81 | 0.00 | **o**: → **ɔ center** | N 〉 NN |
| **ɔ** | -0.36 | 0.08 | -4.39 | 0.00 | **ɔ** → **o**: **center** | N 〉 NN |
| **u**: | -0.11 | 0.08 | -1.38 | 0.17 | **u**: → **ʊ center** | |
| **ʊ** | -0.15 | 0.08 | -1.80 | 0.07 | **ʊ** → **u**: **center** | |

**Abbildung 4** – *Mixed-effects models*: Separabilität der Vokalnachbarpaare; N - native; NN - nicht-native

Bei den Vokalpaaren /iː/-/ɪ/ sowie /oː/-/ɔ/ wird mehr Separabilität bei der Referenzgruppe beobachtet als bei den beiden Probanden-Gruppen. Es gibt keine wesentlichen Unterschiede der Separabilität für die Vokalpaare /eː/-/ɛ/ und /uː/-/ʊ/ zwischen der Referenz- und den Probanden-Gruppen. Unterschiede zwischen den Probanden-Gruppen verbleiben in der Ausprägung der

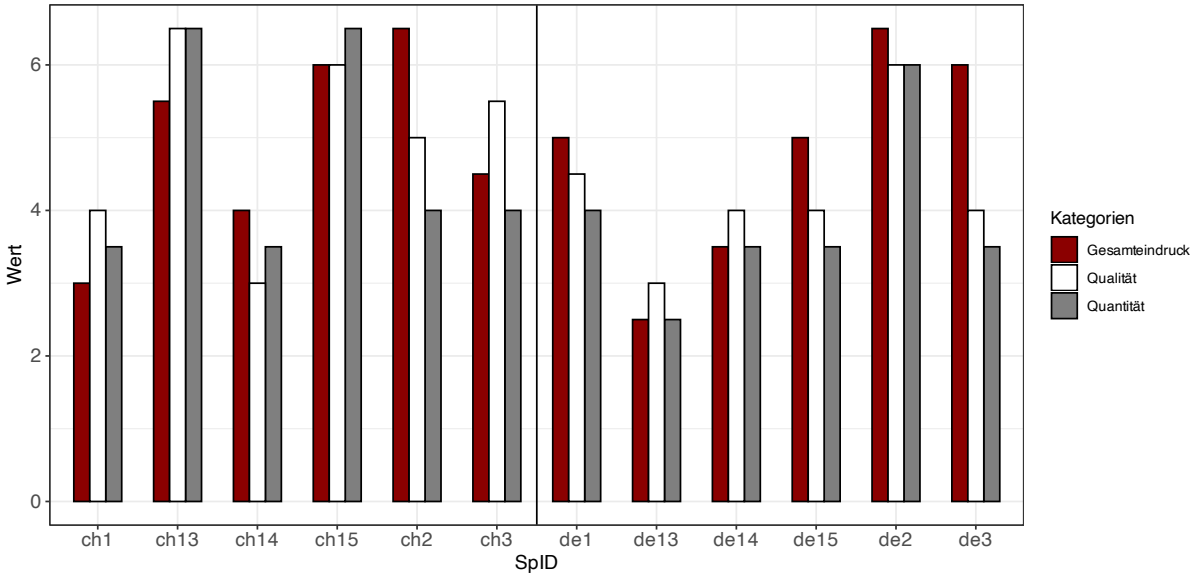



**Abbildung 5** – subjektive Bewertungen (Median) durch De-L1-Hörer:innen

Distanz.

### 3.3 Evaluierung durch L1-Hörer:innen

Aus den Ergebnissen ED der Vokaldauer (s. 3.1) wurde von jeweiliger Probanden-Gruppe 6 Probanden nach ED ausgewählt (3 mit den wenigsten ED-Punkten und 3 mit den höchsten). Diese 12 Probanden wurden dann von fünf Deutsch-Muttersprachlern, die Sprechwissenschaft studieren, auf der Grundlage eines Bewertungsbogens subjektiv bewertet. Diese Muttersprachler:innen bewerteten die Aufnahmen der Probanden subjektiv in drei Kategorien: Gesamteindruck, Vokal-Quantität und Vokal-Qualität, wobei jede Kategorie eine 10-stufige Skala aufwies, von 1 (= minimale Abweichung) bis 10 (= größere/häufige Abweichungen). D.h., jede hohe Punktzahl deutet auf eine größere Abweichung der Aussprache der Probanden von der Muttersprachler-Norm hin und eine niedrigere Verständlichkeit. Auf diese Weise wurden Durchschnittswerte und Medianwerte der subjektiven Bewertungen der Muttersprachler für die Aussprache der 12 Probanden ermittelt (siehe Abb. 5 für Medianwerte).

## 4 Diskussion

Die obigen Ergebnisse der Bewertungen können mit den durchschnittlichen ED-Werten der obigen 12 Probanden (s. Abb. 1) korreliert werden. Allerdings belaufen sich die (Pearson) Korrelationskoeffizienten (für ED~Gesamteindruck -0.05, ED~V.Quantität 0.36, ED~V.Qualität 0.46) auf niedrigem Niveau und sind zudem nicht signifikant. Für die China-Gruppe scheinen die subjektiven Bewertungen der Muttersprachler im Großen und Ganzen mit den objektiven Daten übereinzustimmen, aber für die Deutschland-Gruppe sind die Bewertungen eher unerwartet. Die Ursachen für diese Ergebnisse könnten sowohl in der Anzahl der L1-Hörer als in der Anzahl der Probanden liegen, d.h. dass die Variabilität der Lerner insgesamt sehr groß ist. Darüber hinaus könnte die Bewertungsoption des Bewertungsbogens zu einseitig sein, da sie die Auswirkungen von Konsonanten, Silben, Akzent und suprasegmentalen Merkmalen auf die Perzeption der Muttersprachler nicht weiter berücksichtigt.

## 5 Schlussfolgerungen & Ausblick

Die vorliegende Methode zur Darstellung der Unterschiede zwischen Probandengruppen mit unterschiedlicher Lernbiographie und Lernumgebung zeigt erste Ansätze zur Differenzierung der Gruppen, bleibt jedoch in ihrer Aussagekraft begrenzt. Zwar lassen sich signifikante Unterschiede der Kategorienmittelwerte zwischen den beiden Probandengruppen und der Referenzgruppe feststellen, doch innerhalb der beiden L2-Probandengruppen sind diese Unterschiede mit dem verwendeten Modell bislang nicht eindeutig nachweisbar. Auch die gruppenbasierte Analyse der Vokaldauerdistanzen liefert bislang keine konsistente Richtung der Abweichungen. Ebenso bleiben die F1/F2-Lageunterschiede zwischen den beiden L2-Gruppen ohne klar interpretierbares Muster. Für eine belastbarere Interpretation der Ergebnisse erscheinen daher folgende Erweiterungen des Untersuchungsdesigns sinnvoll: (a) Vergrößerung der Referenzgruppe zur Stabilisierung der Vergleichsbasis, (b) Erweiterung der Evaluierungsgruppe zur Verbesserung der statistischen Aussagekraft, (c) Integration weiterer prosodischer Parameter sowie der Abweichungen von Konsonanten, Silben, Wortakzent um differenziertere Akzent- und Rhythmusunterschiede erfassen zu können. Diese Anpassungen könnten dazu beitragen, die Sensitivität der Methode zu erhöhen und ein klareres Bild der sprecherbezogenen und lernbiographischen Einflüsse auf die akustischen Parameter zu gewinnen.

## Dank

Wir danken allen Teilnehmer:innen sowohl des Produktionsexperimentes in Deutschland und China als auch denen der Hörevaluation. Ebenfalls danken wir für die Möglichkeiten der Audioaufnahmen im Phonetiklabor an der Goethe-Universität Frankfurt am Main (Institut für empirische Sprachwissenschaft, FB9).

## Literatur

[1] LIU, C.: *Einige Überlegungen zum gegenwärtigen Deutschunterricht in China. Zielsprache Deutsch*, 4(1982), S. 29–39, 1982.

[2] HUNOLD, C.: *Untersuchungen zu segmentalen und suprasegmentalen Ausspracheabweichungen chinesischer Deutschlernender*. 28. Peter Lang, Frankfurt am Main, 2009.

[3] LAY, T.: *Lernschwierigkeiten chinesischer deutschlerner im anfängerunterricht unter berücksichtigung des chinesischen als erstsprache und des englischen als erster fremdsprache. Zielsprache Deutsch*, 44(2), S. 19–38, 2017.

[4] LI, X.: *Aussprachetraining im Bereich der Prosodie für chinesische DaF-Lernende*. Nr. 28 in Schriften zur Sprechwissenschaft und Phonetik. Frank & Timme GmbH, Berlin, 1st edn., 2023. doi:10.57088/978-3-7329-8998-0.

[5] DING, H.: *An Acoustic-Phonetic Analysis of Chinese in Comparison with German in Text-to-Speech Systems and Foreign Language Speech Learning*. Nr. 67 in Studientexte Zur Sprachkommunikation. TUDpress, Verl. der Wiss, Dresden, 2013.

[6] CHEN, M.: *Phoneme Similarity in English and Mandarin Chinese Language Production*. Ph.D. thesis, Lehigh University, 2023. URL `https://preserve.lehigh.edu/_flysystem/fedora/2023-11/preserve30918.pdf`.

[7] DENG, D.: *[Effect of Cross-language Phonetic Similarity on American Speakers Acquire Chinese Vowels]. [Chinese Language Learning]*, 3, S. 71–84, 2017. URL http://111.205.231.68/docs/2019-03/20190309230405980914.pdf.

[8] FENG, X., Y. GAO, B. LIN, und J. ZHANG: *[An Entropy-based Evaluation of L2 Speech Acquisition: The Preliminary Report on Chinese Initials Produced by Japanese Learners)]*. In M. SUN, Y. LIU, W. CHE, Y. FENG, X. QIU, G. RAO, und Y. CHEN (Hrsg.), *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, S. 79–87. Chinese Information Processing Society of China, Nanchang, China, 2022. URL https://aclanthology.org/2022.ccl-1.8/.

[9] XIE, X. und T. F. JAEGER: *Comparing non-native and native speech: Are L2 productions more variable? The Journal of the Acoustical Society of America*, 147(5), S. 3322–3347, 2020. doi:10.1121/10.0001141.

[10] MUNRO, M. J.: *Variability in L2 Vowel Production: Different Elicitation Methods Affect Individual Speakers Differently*. *Frontiers in Psychology*, 13, S. 916736, 2022. doi:10.3389/fpsyg.2022.916736.

[11] FLEGE, J. E.: *Second language speech learning: Theory, findings, and problems*. In W. STRANGE (Hrsg.), *Speech perception and linguistic experience: Issues in cross-language research*, Bd. 92, S. 233–277. York Press, Baltimore, 1995.

[12] STRANGE, W.: *Cross-language phonetic similarity of vowels: Theoretical and methodological issues*. In O.-S. BOHN und M. J. MUNRO (Hrsg.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, S. 35–55. John Benjamins Publishing Company, 2008.

[13] HIRSCHFELD, U. und K. REINKE: *Phonetik im Fach Deutsch als Fremd- und Zweitsprache: unter Berücksichtigung des Verhältnisses von Orthografie und Phonetik*. Nr. 1 in Grundlagen Deutsch als Fremd- und Zweitsprache. Erich Schmidt Verlag, Berlin, 2., neu bearbeitete Auflage edn., 2018.

[14] BOERSMA, P. und D. WEENINK: *Praat: Doing phonetics by computer [Computer Program]*. 2024-09-18. URL https://www.fon.hum.uva.nl/praat/.

[15] NEAREY, T. M.: *Phonetic feature systems for vowels.* Ph.D. thesis, University of Alberta. Reprinted 1978 by the Indiana University Linguistics Club, 1977. URL https://sites.ualberta.ca/~tnearey/Nearey1978_compressed.pdf.

[16] ADANK, P., R. SMITS, und R. VAN HOUT: *A comparison of vowel normalization procedures for language variation research. The Journal of the Acoustical Society of America*, 116(5), S. 3099–3107, 2004. doi:10.1121/1.1795335.

[17] WEDEL, A., N. NELSON, und R. SHARP: *The phonetic specificity of contrastive hyperarticulation in natural speech. Journal of Memory and Language*, 100, S. 61–88, 2018. doi:10.1016/j.jml.2018.01.001.

**THE RECOGNITION OF ENGLISH STRESS IN BACKGROUND BABBLE NOISE**

*Marcel Schlechtweg[12], Leah Klußmann[1], Stephanie Kaucke[12]*

*[1]Institute for English and American Studies, Carl von Ossietzky Universität Oldenburg,*
*[2] Cluster of Excellence Hearing4all*
*marcelschlechtweg@gmail.com*

**Abstract:** Individuals are frequently exposed to spoken language in a variety of adverse listening conditions, such as noise. Using a forced-choice experiment and analyzing response accuracy with binomial logistic regression, the present paper investigates whether and how native speakers of English recognize the stress difference in items like *incline* (noun, initial stress) and *incline* (verb, non-initial stress) if these words are embedded in background noise. While all items were examined with both initial and non-initial stress, the pairs differed with respect to the degree of vocalic variation in the first syllable. For instance, the quality of the first vowel in *incline* (initial stress) and *incline* (non-initial stress) was nearly identical, as expressed in the F1 and F2 values. In *conflict*, however, we found substantial vocalic variation between the item with initial and the one with non-initial stress. The analysis showed that an increasing F1 (but not F2) distance between the two variants of a pair supported the listener in keeping the two variants of a pair apart in adverse listening conditions. Crucially, an increasing F1 distance cancelled out the negative effect of the most adverse condition. We used background babble noise in the study, with the voices of two females, two males, or two females plus two males competing with the target signal (female).

## 1 Introduction

Listening to spoken language in adverse conditions can represent an obstacle for individuals. Background babble noise, the presence of background talkers whose voices compete with the voice of the speaker talking to the listener, is one example of an adverse listening condition. Generally speaking, noise can be a challenge for the listener on various levels (see, e.g., [1]; [2]; [3]). For one, noise stands in acoustic competition with the target signal and a physical overlap between noise and target signal can possibly prevent the listener from completely recognizing the target signal. Beyond the pure acoustic level, listeners can face an additional burden during language processing when being exposed to spoken language in noise, potentially leading to a reduction of processing resources.

Research has shown that several factors come into play when one discusses how strongly noise affects recognition and perception (see also, e.g., [4]). Some of these factors refer to the listener. Native speakers seem to be less (negatively) affected by the presence of noise than non-native speakers (see, e.g., [5]; [6]). Equally, monolinguals, compared to bilinguals, adults, in comparison to children, and individuals without hearing loss, compared to individuals with hearing loss, can handle adverse listening conditions more easily (see, e.g., [7]; [8]; [9]; [10]). Moreover, some factors refer to the noise itself. A listening condition is generally more optimal if the target signal is more dominant, or more intense, than the noise (see, e.g., [11]). Also, when background babble serves as a masker, the number of background talkers and the language of the noise can increase or decrease the challenges on the listener's side (see, e.g., [12]). For instance, it is typically more difficult to block out noise produced in one's native language, in comparison to noise in a different language. Finally, some factors refer to the linguistic materials. Specific accents, such as a non-native or an unfamiliar native accent, and sequences containing code switching present in the target signal have been found to increase the degree of difficulty in perceiving spoken language in noise (see, e.g., [7]; [13]; [14]). Furthermore,

dealing with certain vowels, consonants, morphological paradigms, or syntactic structures has been reported to be less challenging than dealing with others (see, e.g., [1]; [15]; [16]). Also, we observe that the frequency of occurrence of items, the context, and the presence of redundancy, such as morphosyntactic agreement, can facilitate spoken language processing in noise (see, e.g., [4]; [9]; [10]).

In the present work, we examine variation on the level of the linguistic materials and on the level of the noise. In the category of the linguistic materials, we focus on English items that occur with either initial (as a noun) or non-initial stress (as a verb). While for some of these pairs the vowel quality is similar across the two variants, it is more or less different for other pairs due to vowel reduction. We look at whether this variation facilitates or inhibits the listener's processing in noise. In the literature, we find a few contributions examining the role of suprasegmental aspects during the recognition and perception of spoken language in noise ([17]; [18]; [19]; [20]; [21]; [22]). However, all of the aforementioned studies investigated suprasegmental features in linguistic materials larger than individual words, such as sentences. In the present article, we analyze the recognition of word stress in noise, more precisely, the listener's ability to recognize variation in word stress between two items.

In the noise category, we investigate the effects of female-only (two speakers), male-only (two speakers), and mixed background babble (two female plus two male speakers) on the recognition of items produced by a female speaker. Research has shown that a lower degree of similarity between the linguistic materials and the noise typically facilitates the listener's task. If the voice speaking the linguistic materials and the voice producing the noise originate from individuals with a different gender, the masking effect is less severe (see, e.g., [23]; [24]; [25]). Fundamental frequency (F0) represents one potential reason for this trend, since female and male voices often differ with respect to this parameter, and since increasing F0 differences between noise and linguistic materials has been found to be beneficial for the listener (see, e.g., [25]; [26]).

## 2 Methodology

Relying on a forced-choice design, we tested whether and how native speakers of English recognized the stress difference in pairs like *incline* (stressed on the first syllable if it is a noun) and *incline* (stressed on the second syllable if it is a verb) in adverse listening conditions (see also [27] for a small pilot study in a similar direction).

### 2.1 Participants

60 native speakers of US American English without any other native language, with normal hearing, and without a speech or language disorder completed the whole study (35 male, 25 female; mean age: 28.33 years; SD age: 4.70 years; age range: 19–35 years).[1]

### 2.2 Materials

24 pairs of English words served as the materials. Each pair consisted of a disyllabic noun, which was spoken with initial stress (e.g., *incline*), and the verbal counterpart, which was produced with final stress (e.g., *incline*).

*combat, conduct, conflict, contest, contract, contrast, desert, discount, extract, incline, insult, invite, object, permit, pervert, present, progress, project, protest, rebel, record, refund, remake, subject*

[1] The data from four individuals was not included in the analysis due to missing information. That is, the data from 60 rather 64 individuals was examined.

For each of the 48 words (24 pairs), we measured F1 and F2 at the midpoint of the vowel of the first syllable. Then, we calculated the difference between F1 of the vowel of the first syllable of the initially stressed item (e.g., *incline*) and F1 of the vowel of the first syllable of the non-initially stressed item (e.g., *incline*). We also created the F2 difference.

The 24 pairs were recorded by a 33-year-old female native speaker of US American English from Philadelphia in a sound-proof booth, using a large-diaphragm condenser microphone (Røde NT USB, transmission range: 20 Hz to 20 kHz; limit sound pressure level: 110 dB SPL), and Praat ([28]; 44.1 kHz, mono). The speaker was asked to realize the respective stress pattern (initial or non-initial stress) by speaking as naturally as possible. Three recordings were obtained for each word and the best sound file was selected for the experiment. Relying on a perceptual judgment and visual information obtained from Praat, the stress pattern was checked by the first and second author independently. All sound files were normalized to 60 dB.

All items were embedded in three types of multi-speaker background babble: (a) background babble with two female voices (the two were 25 years old), (b) background babble with two male voices (the two were 34 and 35 years old), and (c) background babble with the same two female and male voices. All speakers were native speakers of German from Northern Germany. The background babble was produced in a language (German) that none of the 60 participants of the experiment (native speakers of US American English) was familiar with. The four native speakers of German read out a passage and were recorded with the same equipment and in the same booth as the speaker of the materials mentioned above. The recordings were normalized to 60 dB. Portions of what was read out were then used to create the background babble, by combining portions from the two females, the two males, or both in Praat. This resulted in three babble files (female, male, mixed) for each of the test pairs.

### 2.3 Procedure

All participants of the experiment were recruited and paid via prolific.com, an online platform that helps researchers to anonymously screen, recruit, and pay individuals. The experiment was run online using PsychoPy3 ([29]) via the platform Pavlovia (www.pavlovia.org).

In each trial, participants were exposed to two auditorily presented items, either to the same item with the same stress pattern twice (e.g., *incline* with initial stress and *incline* with initial stress) or to the same item but with a different stress pattern (e.g., *incline* with initial stress and *incline* with non-initial stress). Each item was presented at 60 dB and in each of the three background babble types, which were also given at 60 dB, creating a signal-to-noise-ratio (SNR) of 0 dB. Participants were asked to indicate via button press whether the two items they heard in a trial (e.g., *incline* with initial stress and *incline* with non-initial stress) were stressed in the same or in a different way.

All participants were tested on 144 trials (24 pairs x 3 background babble types x 2 (same/different)), which were presented at random. That is, everyone was exposed to each pair in each of the three background babble types once with the two items sharing the same stress pattern and once with the two items differing in their stress pattern. Two experiment versions were built. The buttons to be pressed, the order of the stress patterns in different trials, and the distribution of the stress patterns in same trials were counterbalanced. All participants completed a trial run before starting the experiment.

### 2.4 Data analysis

During the analysis, we did not consider the data from nine participants since their overall response accuracy was lower than two third. The average response accuracy of the remaining 51 participants was 90.90 %. A total of 7,344 data points entered the final analysis (51 participants x 144 trials per participant). We examined the response variable Accuracy in two analyses. In the first, only the four main effects (fixed effects) were considered: Log10F1_z,

Log10F2_z, BabbleType (female, male, mixed), and SameDifferent (same, different). "Log10F1_z" refers to the difference of F1 of the vowel in the first syllable of the initially stressed item and F1 of the vowel in the first syllable of the non-initially stressed item. This difference was log transformed (to the base 10), centered, and standardized. We proceeded accordingly for "Log10F2_z". In the second analysis, we also added the two-way interactions (fixed effects). Using R ([30]) and relying on the packages lme4 ([31]) and lmerTest ([32]) (see, e.g., [33]; fit by maximum likelihood (Laplace approximation), see, e.g., [34]), a binomial logistic regression was run. On the level of random effects, we took the intercepts by Participant and Item into consideration. Relying on the emmeans package ([35]), Tukey tests were performed to investigate comparisons not given in the model output. In case of convergence issues, the optimizer of the model was changed to "bobyqa" (see, e.g., [33]). Non-significant terms were removed from the models step-by-step, relying on the *p* values.

## 3 Results

In the analysis with main effects only, we found three main effects. Accuracy was significantly higher (a) with an increasing F1 difference, (b) for the male compared to the female and for the mixed compared to the female babble, and (c) for the "same" trials, in comparison to the "different" trials. The model output is presented in Tables 1 and 2. A Tukey test did not reveal a significant difference for the comparison not specified in the model output (male versus mixed: estimate = 0.205; SE = 0.114; df = Inf; z ratio = 1.806; $p > 0.05$). No main effect of F2 was detected.

**Table 1** - Random effects statistics of the model of Accuracy (without interaction)

| | Variance | Std. Dev. |
|---|---|---|
| Participant (Intercept) | 1.182 | 1.087 |
| Item (Intercept) | 0.160 | 0.400 |

**Table 2** - Fixed effects statistics of the model of Accuracy (without interaction)

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.995 | 0.192 | 10.396 | < 0.001 *** |
| Log10F1_z | 0.411 | 0.093 | 4.436 | < 0.001 *** |
| BabbleType_Male | 0.670 | 0.107 | 6.249 | < 0.001 *** |
| BabbleType_Mixed | 0.465 | 0.103 | 4.518 | < 0.001 *** |
| SameDifferent_Same | 1.240 | 0.095 | 13.025 | < 0.001 *** |

*** < 0.001

The results for the analysis with the two-way interactions are given in Tables 3 and 4 (see also Figure 1). Two interactions were significant (Log10F1_z*BabbleType; Log10F1_z*SameDifferent). While response accuracy was lower for the different than for the same trials for pairs with a smaller F1 difference, there was no difference for the pairs with a greater F1 difference. Also, while response accuracy was lower for the female babble than for the male or mixed babble for pairs with a smaller F1 difference, there was no difference for the pairs with a greater F1 difference.

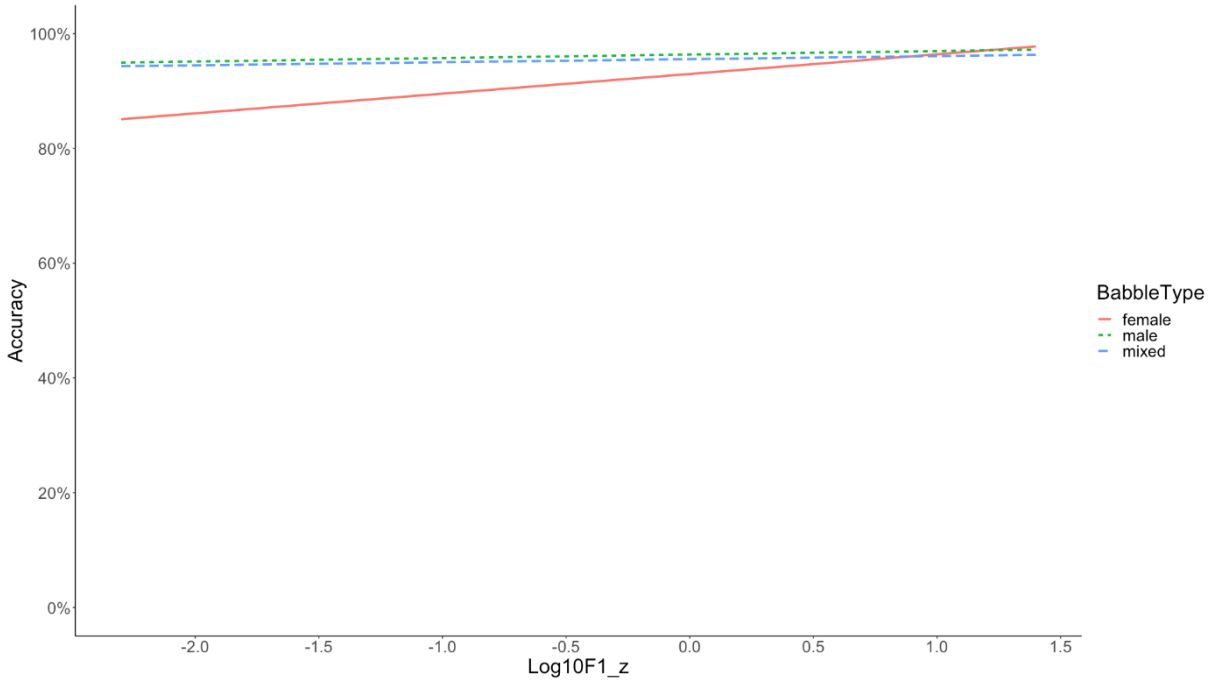



**Figure 1** – Interaction Log10F1_z*BabbleType.

**Table 3** - Random effects statistics of the model of Accuracy (with interaction)

| | Variance | Std. Dev. |
|---|---|---|
| Participant (Intercept) | 1.251 | 1.118 |
| Item (Intercept) | 0.179 | 0.423 |

**Table 4** - Fixed effects statistics of the model of Accuracy (with interaction)

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 2.120 | 0.200 | 10.612 | < 0.001 *** |
| Log10F1_z | 0.821 | 0.116 | 7.095 | < 0.001 *** |
| BabbleType_Male | 0.614 | 0.112 | 5.505 | < 0.001 *** |
| BabbleType_Mixed | 0.400 | 0.107 | 3.749 | < 0.001 *** |
| SameDifferent_Same | 1.114 | 0.097 | 11.482 | < 0.001 *** |
| Log10F1_z:BabbleType_Male | -0.295 | 0.107 | -2.756 | < 0.01 ** |
| Log10F1_z:BabbleType_Mixed | -0.338 | 0.103 | -3.280 | < 0.01 ** |
| Log10F1_z: SameDifferent_Same | -0.726 | 0.098 | -7.435 | < 0.001 *** |

*** < 0.001; ** < 0.01

## 4 Discussion

Relying on data from a forced-choice experiment, the present paper investigated the effect of adverse listening conditions on whether and how native speakers of US American English recognized the stress difference in items like *incline* and *conflict* (both carry initial stress as a noun but non-initial stress as a verb), which were produced by a female native speaker. In the first analysis, focusing on main effects only, three main effects emerged: Response accuracy was higher when participants were exposed to (a) same trials (items with the same stress pattern), compared to different trials, (b) male and mixed babble, in comparison to female babble, and (c) pairs with a greater F1 difference between the vowels of the first syllable of the initially and non-initially stressed item (e.g., *conflict*), compared to items with a lower F1 difference (e.g., *incline*). In the second analysis, including two-way interactions, we found two significant interactions, Log10F1_z*SameDifferent and Log10F1_z*BabbleType: While for

pairs with a smaller F1 difference we observed a higher response accuracy for same than for different trials and for male as well as mixed babble than for female babble, no differences were evident for pairs with a greater F1 difference.

Previous work has shown that a clearer dissimilarity between the target signal and the background babble is often beneficial for the listener (see, e.g., [23]; [24]; [25]; [26]). It is therefore unsurprising that participants in our experiment responded more accurately to target stimuli produced by a female when the background babble was produced by two males. However, two additional observations complete the picture. First, subjects performed in mixed babble (two females plus two males) as well as in male babble (two males). Put differently, adding male babble as a competitor to the target signal produced by a female increased response accuracy. Hoen et al. [2] do not report a difference between male and mixed babble either; however, in their experiment, the target stimuli were produced by a male. Hence, in their study, the presence of a background voice from a person with the same gender as the voice of the target stimuli seemed to interfere with the target independently from additional background voices with a different gender.

Second, segmental variation across our word pairs significantly affected the participants' reactions. That is, a more prominent vocalic difference on the F1 dimension, hence in terms of vowel height, between the members of a pair increased response accuracy. This is in line with previous findings from the literature, such as Kaucke and Schlechtweg's [6], whose subjects were more successful in discriminating the low vowel /a/ and a high or mid vowel, in comparison to the discrimination of vowels of the same height. Crucially, a more prominent vocalic difference between the members of a pair also resulted in a similar performance across the three babble types in the present experiment. Again, we see parallels to recent work, for example, to Schlechtweg's [16] data. He tested, among other things, whether individuals' ability to keep English singulars and plurals apart (in quiet and white noise) depended on the nature of the nominal paradigm. It was found that for the item *goose/geese*, with vowel fronting signaling the number contrast, the ability to detect a singular and plural drastically decreased across different adverse listening conditions, specifically from a SNR of -6 dB (84 % response accuracy) to a SNR of -12 dB (68 % response accuracy). For *foot/feet*, however, where vowel fronting is accompanied by vowel lengthening to indicate the number distinction, response accuracy remained comparable (-6 dB: 95 %; -12 dB: 93 %). Hence, while vowel lengthening compensated the effect of a more adverse listening condition (namely -12 dB) in Schlechtweg [16], a greater F1 difference did so in our study, with the more adverse condition being the one in which the speaker of the stimulus and all speakers of the babble shared the same gender.

## 5 Conclusion

We showed that increasing variation along the F1 dimension can support the listener in the presence of background speakers competing with the target speaker. Our data suggests that this variation can even cancel out the negative effect of the most adverse listening condition.

## 6 References


[1] CARROLL, R., and E. RUIGENDIJK: *The effects of syntactic complexity on processing sentences in noise*. In *Journal of Psycholinguistic Research*, 42(2), pp. 139–159. 2013.

[2] HOEN, M., F. MEUNIER, C.-L. GRATALOUP, F. PELLEGRINO, N. GRIMAULT, F. PERRIN, X. PERROT, and L. COLLET: *Phonetic and lexical interferences in informational masking during speech-in-speech comprehension*. In *Speech Communication*, 49, pp. 905–916. 2007.

[3] MATTYS, S.L., M.H. DAVIS, A.R. BRADLOW, and S.K. SCOTT: *Speech recognition in adverse conditions: A review*. In *Language and Cognitive Processes*, 27(7–8), pp. 953–978. 2012.

[4] SCHLECHTWEG, M.: *Morphosyntactic agreement in English: Does it help the listener in noise?* In *English Language and Linguistics*, FirstView. 2024.

[5] GARCIA LECUMBERRI, M.L., and M. COOKE: *Effect of masker type on native and non-native consonant perception in noise*. In *The Journal of the Acoustical Society of America*, 119(4), pp. 2445–2454. 2006.

[6] KAUCKE, S., and M. SCHLECHTWEG: *English speakers' perception of non-native vowel contrasts in adverse listening conditions: A discrimination study on the German front rounded vowels /y/ and /ø/.* In *Language and Speech*, OnlineFirst. 2024.

[7] BENT, T., and E. ATAGI: *Children's perception of nonnative-accented sentences in noise and quiet*. In *The Journal of the Acoustical Society of America*, 138(6), pp. 3985–3993. 2015.

[8] SCHLECHTWEG, M., M. GIBSON, J. AYALA ALCALDE, X. WANG, and L. XU: *Vowel recognition in noise: A comparison of children with cochlear implants and children with typical hearing*, Submitted. 2024.

[9] SCHMIDTKE, J: *The bilingual disadvantage in speech understanding in noise is likely a frequency effect related to reduced language exposure*. In *Frontiers in Psychology*, 7(Article 678), pp. 1–15. 2016.

[10] SMILJANIC, R., and D. SLADEN: *Acoustic and semantic enhancements for children with cochlear implants*. In *Journal of Speech, Language, and Hearing Research*, 56(4), pp. 1085–1096. 2013.

[11] BENT, T., D. KEWLEY-PORT, and S.H. FERGUSON: *Across-talker effects on non-native listeners' vowel perception in noise*. In *The Journal of the Acoustical Society of America*, 128(5), pp. 3142–3151. 2010.

[12] VAN ENGEN, K.J., and A.R. BRADLOW: *Sentence recognition in native- and foreign-language multi-talker background noise*. In *The Journal of the Acoustical Society of America*, 121(1), pp. 519–526. 2007.

[13] ADANK, P., B.G. EVANS, J. STUART-SMITH, and S.K. SCOTT: *Comprehension of familiar and unfamiliar native accents under adverse listening conditions*. In *Journal of Experimental Psychology: Human Perception and Performance*, 35(2), pp. 520–529. 2009.

[14] GROSS, M.C., H. PATEL, and M. KAUSHANSKAYA: *Processing of code-switched sentences in noise by bilingual children*. In *Journal of Speech, Language, and Hearing Research*, 64(4), pp. 1283–1302. 2021.

[15] CUTLER, A., A. WEBER, R. SMITS, and N. COOPER: *Patterns of English phoneme confusions by native and non-native listeners*. In *The Journal of the Acoustical Society of America*, 116(6), pp. 3668–3678. 2004.

[16] SCHLECHTWEG, M: *The interaction of paradigmatic and syntagmatic properties during the perception of spoken English in noise*, Submitted. 2024.

[17] SHEN, J.: *Older listeners' perception of speech with strengthened and weakened dynamic pitch cues in background noise*. In *Journal of Speech, Language, and Hearing Research*, 64, pp. 348–358. 2021.

[18] SMITH, M.R., A. CUTLER, S. BUTTERFIELD, and I. NIMMO-SMITH: *The perception of rhythm and word boundaries in noise-masked speech*. In *Journal of Speech and Hearing Research*, 32, 912–920. 1989.

[19] VAN ZYL, M., and J.J. HANEKOM: *Speech perception in noise: A comparison between sentence and prosody recognition*. In *Journal of Hearing Science*, 1(2), 54–56. 2011.

[20] WATSON, P.J., and R.S. SCHLAUCH: *The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours*. In *American Journal of Speech-Language Pathology*, 17, 348–355. 2008.

[21] PINET, M., and P. IVERSON: *Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience*. In *The Journal of the Acoustical Society of America*, 128(3), pp. 1357–1365. 2010.

[22] SCHARENBORG, O., E. KOLKMAN, S. KAKOUROS, and B. POST: *The effect of sentence accent on non-native speech perception in noise*. In *Proceedings of Interspeech 2016*, pp. 863–867. 2016.

[23] BRUNGART, D.S.: *Informational and energetic masking effects in the perception of two simultaneous talkers*. In *The Journal of the Acoustical Society of America*, 109(3), pp. 1101–1109. 2001.

[24] BRUNGART, D.S., B.D. Simpson, M.A. Ericson, and K.R. Scott: *Informational and energetic masking effects in the perception of multiple simultaneous talkers*. In *The Journal of the Acoustical Society of America*, 110(5), pp. 2527–2538. 2001.

[25] DARWIN, C.J., D.S. BRUNGART, and B.D. SIMPSON: *Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers*. In *The Journal of the Acoustical Society of America*, 114(5), pp. 2913–2922. 2003.

[26] COOKE, M., M.L. GARCIA LECUMBERRI, and J. BARKER: *The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception*. In *The Journal of the Acoustical Society of America*, 123(1), pp. 414–427. 2008.

[27] KLUßMANN, L: *The perception of English prosody by native speakers of German in adverse listening conditions*. MA thesis, Carl von Ossietzky Universität Oldenburg, Germany. 2023.

[28] BOERSMA, P., and D. WEENINK: *Praat: Doing phonetics by computer* [Computer program]. Version 6.3.10. http://www.praat.org/. 2023.

[29] PEIRCE, J., J.R. GRAY, S. SIMPSON, M. MACASKILL, R. HÖCHENBERGER, H. SOGO, E. KASTMAN, and J.K. LINDELØV: *Psychopy2: Experiments in behavior made easy*. In *Behavior Research Methods*, 51(1), pp. 195–203. 2019.

[30] R CORE TEAM. *R: A language and environment for statistical computing*. R version 4.0.5. R Foundation for Statistical Computing, Vienna, Austria. 2021. https://www.R-project.org.

[31] BATES, D., M. MÄCHLER, B. BOLKER, and S. WALKER: *Fitting linear mixed-effects models using lme4*. Version 1.1.29. In *Journal of Statistical Software*, 67(1), pp. 1−48. 2015.

[32] KUZNETSOVA, A., P.B. BROCKHOFF, and R.H.B. CHRISTENSEN: *lmerTest package: Tests in linear mixed effects models*. Version 3.1.3. In *Journal of Statistical Software*, 82(13), pp. 1–26. 2017.

[33] WINTER, B: *Statistics for linguists: An introduction using R*. Routledge, New York. 2020.

[34] FIELD, A., J. MILES, and Z. FIELD: *Discovering statistics using R*. Sage, Los Angeles. 2012.

[35] LENTH, R.V: *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.2.1. 2020. https://cran.r-project.org/package=emmeans.

## 7 Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 – Project ID 390895286.

# WHAT NOISE DOES A DRAGON MAKE? ICONICITY IN THE PRODUCTION OF NON-WORD VOCALISATIONS

*Alexandra Schmidt*[1] *and Farhat Jabeen*[1]
[1]*Bielefeld University, Germany*
*firstname.lastname@uni-bielefeld.de*

**Abstract:** This study investigates the relation between iconic word forms and the properties of their referents. Using spontaneous speech productions, we analyze the role of animacy, size, and animal group in the selection of vocalic and consonantal features used in the associated word forms. Our analysis showed a significant effect of animacy on all the consonantal features and on vowel openness. Moreover, size played a role in the selection of manner of articulation and vowel openness. Animal group also affected the use of place and manner of articulation in the production of word forms. Our results replicate the previous findings on the relation between referent size and iconic word forms. We also bring new evidence regarding the role of animacy and animal group in the selection of associated word forms. Our findings indicate the importance of using spontaneous speech to investigate the sound-symbol relation.

## 1 Introduction

### 1.1 Definition and Examples of Iconicity

Iconicity is the non-arbitrary relationship between a word's form and its referent's properties, which can be found across languages [1] and writing systems [2]. Köhler, (as cited in [3]), described participants matching a shape consisting of straight lines (i.e., a sharp shape) to the word "takete" and a round shape to the word "maluma". With their "bouba-kiki" paradigm, the most famous example of iconicity, [4] replicated Köhler's findings. Furthermore, [1] found the place of articulation to be a meaningful factor in sound-shape correspondence and showed that the degree of vowel openness can be related to real-world objects as well as abstract shapes.

Recently, many more properties of a referent have been demonstrated to influence the associated sound, including friendliness [5], strength [6], and taste [2]. For the current study, we are interested in investigating the relation between size and word form. This connection can partly be attributed to Ohala's Frequency Code [7]. The code predicts that low F0 and low formants are employed by many mammals, including humans, to appear dominant, whereas higher F0 and formants signal submission or helplessness. Moreover, in recent studies on iconicity, small entities have been shown to be associated with closed [8] and front vowels as well as voiceless consonants [9]. Big entities are associated with back vowels [8] and voiced consonants [9].

### 1.2 Motivation

While most of the aforementioned works are based on the perception data or controlled production experiments, this study seeks to investigate the sound-shape correspondence in spontaneous productions. We aim to replicate findings on the associations between referent size (big vs. small) and their word forms. Furthermore, we analyze the role of animacy (animate vs. inanimate) and animal group (mammals vs. reptiles vs. birds) in the sound-shape relationship.

### 1.3 Hypotheses

We expect the sound-shape correspondence to behave similarly in production as well as in perception. Hence we hypothesize that the word forms for small animals contain close front vowels and voiceless consonants (H1). Correspondingly, the word forms for big animals are predicted to consist of open vowels and voiced consonants. We also hypothesize an effect of animacy (H2) and animal group (H3) on the sound-shape correspondence. The details of our dataset and analysis are as follows.

## 2 Methods

### 2.1 Data Set

Our data stems from the quiz section of the BBC Radio1 show "Breakfast with Greg James". In the quiz, participants were asked to produce sounds they associated with different animals and inanimate objects. From that quiz show, we recorded 73 sound produced by 49 participants (27 m, 22 f) and manually annotated the sounds for consonant and vowel types using Praat [10].

### 2.2 Analysis

The referents in the audio recordings were categorized as animate (n = 236) and inanimate (n = 40). Furthermore, the animate category was sorted by size and animal group. Table 1 provides an overview of the frequencies for each category used in our analysis. Due to the scarcity of data points, amphibians and insects were omitted from the analysis (n = 14).

**Table 1** – Frequency of different categories in the animate class.

| **Group** | | | **Size** | |
|---|---|---|---|---|
| Birds | Mammals | Reptiles | Big | Small |
| 30 | 180 | 12 | 166 | 70 |

In accordance with the IPA standards [11], consonants were grouped by place of articulation (POA), manner of articulation (MOA), and voicing (VOI). Vowels were categorized by the degree of openness (OPN), tongue position (POS), and roundedness (ROU). The openness category was reduced to three levels i.e., close, mid, and open. Velar and glottal sounds were labeled as "Cback" place of articulation for the consonants and the bilabial and labiodentals were jointly categorized as "labial".

The statistical analysis was carried out in R [12]. Multinomial regression, using "nnet" package [13], was run to analyze the role of consonantal features (POA, MOA, VOI) and vocalic features (OPN, POS, and ROU) to predict animacy, size, and animal group. Likelihood Ratio Tests were conducted using the "lmTest" package [14].

## 3 Results

Our analysis yielded 493 data points: 204 vowels and 289 consonants. The number of vowels per feature can be seen in Table 2. It indicates a general tendency to use unrounded close/mid vowels produced with a fronted tongue position. The frequency of consonants for each vocalic feature is shown in Table 3. The table illustrates our participants' overall preference for voiced plosives and approximants at alveolar and back place of articulation.

**Table 2** – Frequency of different vowel types.

| OPN | | | POS | | | ROU | |
|---|---|---|---|---|---|---|---|
| Close | Mid | Open | Front | Central | Back | Round | Unrouded |
| 77 | 80 | 47 | 94 | 50 | 60 | 62 | 142 |

**Table 3** – Frequency of different consonant types.

| POA | | | MOA | | | | | | VOI | |
|---|---|---|---|---|---|---|---|---|---|---|
| Lab. | Alv. | Back | Plo. | Nas. | Tri. | Fri. | App. | Cli. | Voiced | Voiceless |
| 84 | 101 | 104 | 82 | 37 | 13 | 60 | 91 | 6 | 172 | 117 |

### 3.1 Animacy

The statistical analysis showed a significant effect of animacy on all the consonantal features analyzed here i.e., (POA: $\chi^2$ = 6.34, p = .04 ; MOA: $\chi^2$ = 53.14; p < .001, VOI: $\chi^2$ = 21, p < .001). Moreover, animacy significantly affected vowel OPN ($\chi^2$ =21.37, p < .001). Figure 1 illustrates that both open and mid vowels are almost exclusively used for animates. Closed vowels sometimes represent word forms for the inanimate objects (25%).

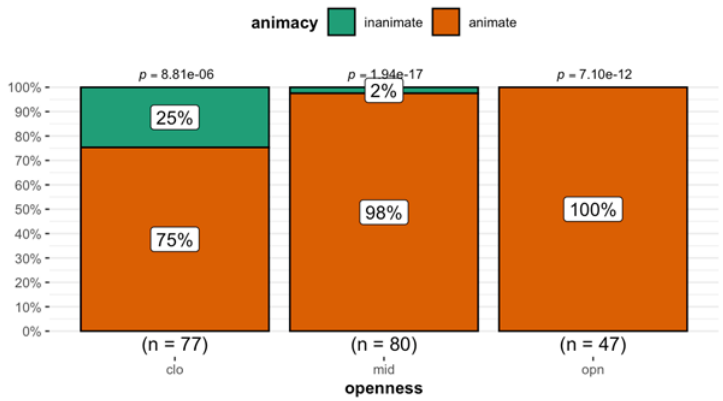



**Figure 1** – Animacy and vowel openness.

Figure 2 (top) shows that the most common POAs for inanimate objects are labial (25%) and alveolar (17%). Back consonants are mostly used for animate objects (95%). As for the manner of articulation (see middle panel in Figure 2), inanimate objects are frequently represented by using trills (62%). Fricatives (18%), nasals (16%), and plosives (16%) also occur but are more often associated with animates. Similarly, approximats (95%) and clicks (100%) are more strongly associated with the animate category. Finally, the bottom panel of Figure 2 shows that the voiceless consonants are strongly associated with animates (95%) compared to the voiced consonants. This data confirms our H2 as animacy plays a role in the use of consonantal and vocalic features produced in the word forms analyzed here.

The analysis in the following subsections is based on the animate category only.

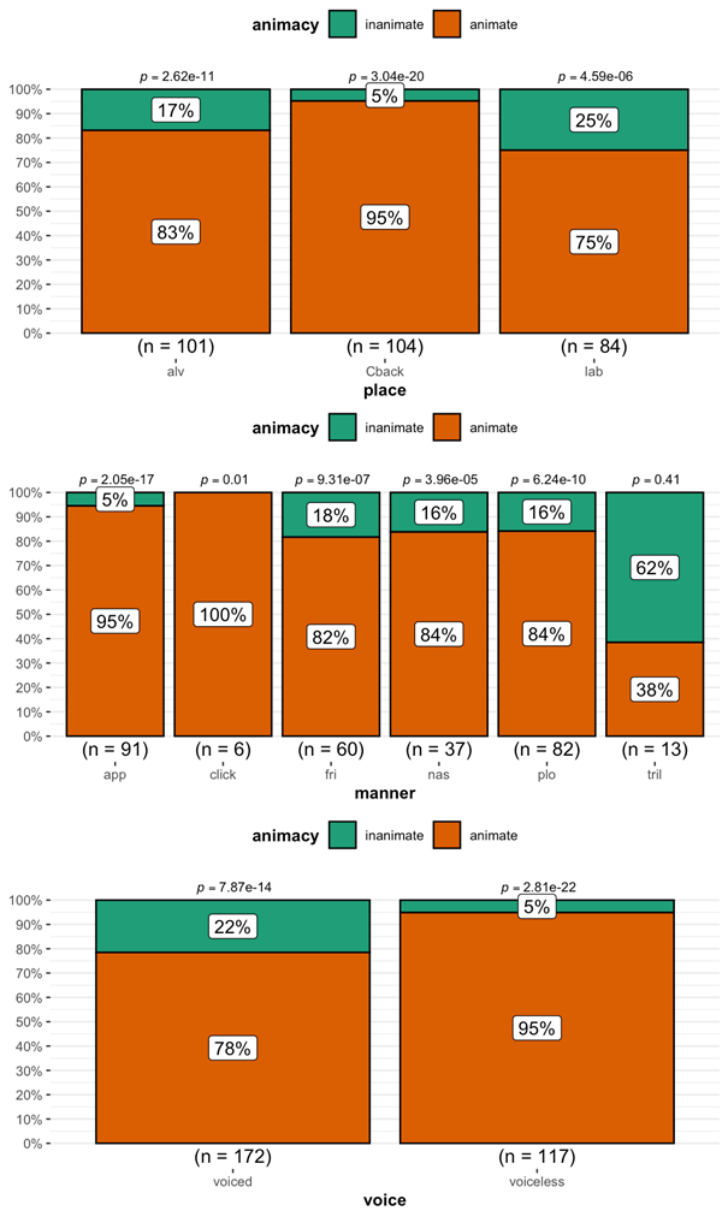



**Figure 2** – Animacy and voicing, place, and manner of articulation.

### 3.2 Size

The analysis showed significant associations between size and OPN ($\chi^2$ = 21.377, p < .001). As illustrated by Figure 3 (top), the word forms produced for small animals frequently consist of close vowels than of mid and open vowels. Big animals are associated with mid (81%) and open vowels (77%). As no effect of POS and VOI were found, H1 is only partially confirmed.

The bottom panel in Figure 3 reveals a significant relation between size and MOA ($\chi^2$ =28.433, p < .001). It shows that approximants (81%) and fricatives (82%) are most likely to be associated with big animals, whereas small animals are represented by clicks (100%), plosives (41%), trills (40%), and nasals (35%).

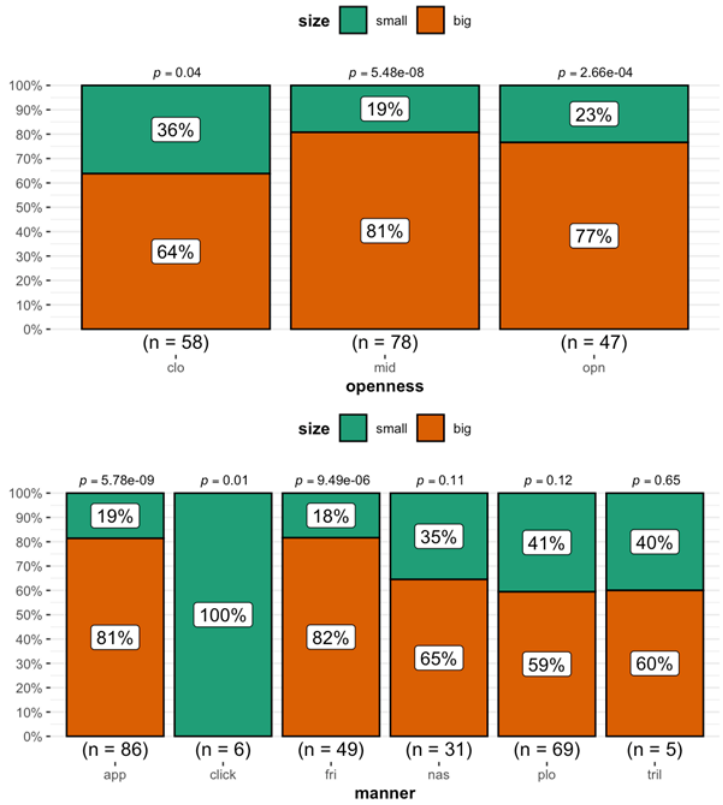



**Figure 3** – Effect of size on vowel openness and manner of articulation in consonants.

### 3.3 Animal Group

Our H3 is also confirmed as we found a significant relation between animal group and POA ($\chi^2$ = 35.26, p < .001) as well as MOA ($\chi^2$ = 33.358, p < .001). Figure 4 (top) shows that labial, alveolar, and back consonants are used in word forms for mammals (89%, 75%, 78% respectively). Alveolars are associated with reptiles as well (16%).

Figure 4 (bottom panel) shows that all MOAs are associated with mammals and birds but with varying frequencies. While the biggest proportion of all sounds is used when producing word forms for mammal (87% approximants, 79% fricatives, 94% nasals, 66% plosives), nasals and approximants are most strongly associated with mammals. Birds are seldom represented by approximants (8%), nasals (6%) and fricatives (4%). However, plosives are used to represent birds in two-thirds of the data (33%). Plosives (2%) and approximants (5%) are seldom associated with reptiles, whereas nasals are never used to represent them.

## 4 Discussion

Using the spontaneously produced data, we replicated the findings of perception studies that postulated a link between the size of a referent and its associated word form. Contrary to the findings of [8] and [9], we did not find significant effects of POS, POA and VOI on size. However, we did find MOA and OPN to be connected to size (H1). Furthermore, animacy and animal group play a role in the selection of consonant and vowel features (H2 & H3). In general, the vocalic features POS and ROU did not yield any significant effects in the iconic oral productions of word forms associated with animates. Contrary to the existing literature, VOI is also not as meaningful as expected, only being significantly related to animacy. POA and OPN each predict two of the three referent categories and therefore have to be considered as meaningful variables in iconic oral productions as in perception studies [1, 8, 9]. We found MOA to be a significant predictor for size, animacy, and animal group, making it the most

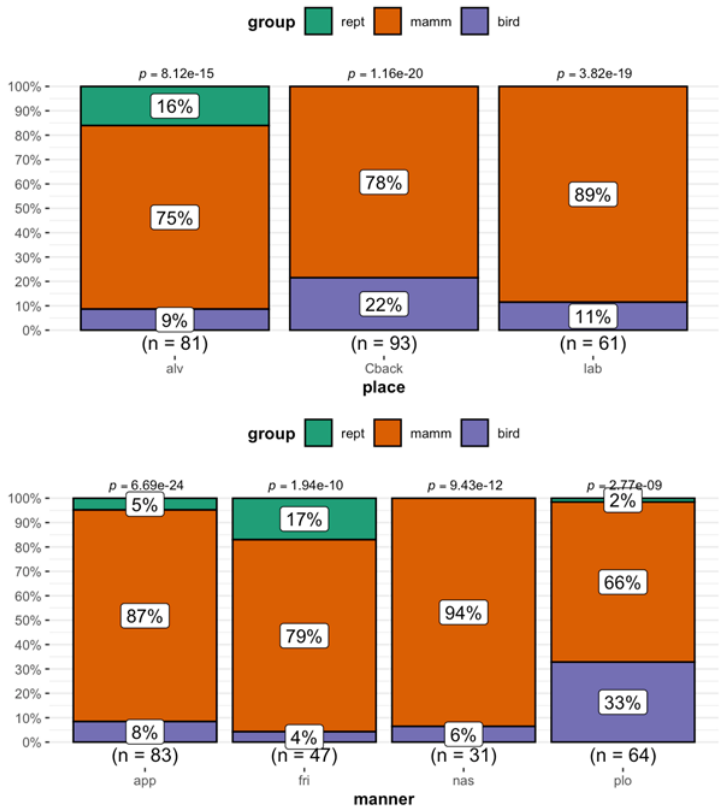



**Figure 4** – Role of animal group on place and manner of articulation.

meaningful predictor in our data set. These differences between our study and the previous findings indicate that sound symbolic relations in spontaneous productions differ from those in perception.

Our findings can be explained with reference to differences between items due to lexical interference [1]. Cats and dogs, for example, have lexicalized associated sounds (“meow”, “woof”) that are learned in childhood, while the word form associated with the noise of an emu does not have a well-known lexicalized form in English. The degree of lexicalization may possibly influence the choice of phonemes our participants employed to represent those animals.

## 5 Limitations and Future Work

Our analysis is delimited by the imbalanced data set, possibly inflating differences between the target variables. Due to the low quality of the recordings, our annotations are based on aural impressions. Thus subjectiveness in phoneme identification cannot be ruled out. In future, we aim to use the findings from spontaneous speech to conduct an experiment on laboratory speech to analyze the role of animacy, size, and animal group in a controlled experimental setup. This will allow us to collect a balanced data set and carry out a more objective acoustic analysis of word forms.

## 6 Conclusions

Despite its limitations, this production study on sound-symbolic relations between phoneme type and referents’ animacy, size, and animal group alludes to differences between iconicity in perception and production. This highlights the importance of using spontaneous speech to investigate this relation.

## References


[1] D'ONOFRIO, A.: *Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm*. *Language and speech*, 57(3), pp. 367–393, 2014.

[2] ĆWIEK, A., S. FUCHS, C. DRAXLER, E. L. ASU, D. DEDIU, K. HIOVAIN, S. KAWAHARA, S. KOUTALIDIS, M. KRIFKA, P. LIPPUS ET AL.: *The bouba/kiki effect is robust across cultures and writing systems*. *Philosophical Transactions of the Royal Society B*, 377(1841), p. 20200390, 2022.

[3] SPENCE, C. and C. V. PARISE: *The cognitive neuroscience of crossmodal correspondences*. *i-Perception*, 3(7), pp. 410–412, 2012.

[4] RAMACHANDRAN, V. S. and E. M. HUBBARD: *Synaesthesia–a window into perception, thought and language*. *Journal of consciousness studies*, 8(12), pp. 3–34, 2001.

[5] KILPATRICK, A., A. ĆWIEK, E. LEWIS, and S. KAWAHARA: *A cross-linguistic, sound symbolic relationship between labial consonants, voiced plosives, and pokémon friendship*. *Frontiers in Psychology*, 14, p. 1113143, 2023.

[6] CWIEK, A.: *Iconicity in language and speech*. Ph.D. thesis, Humboldt-Universität zu Berlin, 2022.

[7] OHALA, J. J., L. HINTON, and J. NICHOLS: *Sound symbolism*. In *Proc. 4th Seoul International Conference on Linguistics [SICOL]*, pp. 98–103. 1997.

[8] SCHMITZ, D.: *A 'size meets cuteness' relation in german vowels*. *Phonetik und Phonologie im deutschsprachigen Raum*, 2022.

[9] CHRISTOPH, N. and A. ĆWIEK: *Lautsymbolische größencodierung bei der benennung von hunden*. *Phonetik und Phonologie im deutschsprachigen Raum*, 2022.

[10] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer [computer program, [v. 6.3]*, 2022. Available at http://www.praat.org/ [retrieved 15.11.2022].

[11] ASSOCIATION, I. P.: *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[12] R CORE TEAM: *R: A language and environment for statistical computing[v. 4.2.0]*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

[13] VENABLES, W. N. and B. D. RIPLEY: *Modern Applied Statistics with S*. Springer, New York, fourth edn., 2002. URL `https://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

[14] ZEILEIS, A. and T. HOTHORN: *Diagnostic checking in regression relationships*. *R News*, 2(3), pp. 7–10, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

# PRODUKTION UND PERZEPTION DER VOKALQUANTITÄT IM ZÜRCHER DIALEKT

*Stephan Schmid[1], Camille Watter[1], Franka Zebe-Sheng[2]*

*[1]Phonetisches Laboratorium, Universität Zürich, [2]Freiburger Institut für Musikermedizin, Hochschule für Musik Freiburg und Medizinische Fakultät der Universität Freiburg*
*stephan.schmid@uzh.ch*

**Abstract:** Distinctive vowel quantity appears to be a pervasive feature of the Zurich German dialect. Based on the previous literature, we assume a vowel inventory of 20 phonemes, where the 10 stressed vowels [i e ɛ y ø œ æ ɒ o u] occur both short and long. For most vowel qualities, minimal pairs testify to the phonemicity of the quantity contrast. The current study reports the results of a production and a perception experiment conducted with 20 native speakers of the Zurich German dialect. 10 older participants (5 female) and 10 younger participants (5 female) first read 18 meaningful sentences with embedded target words in two speech rates (normal and fast). The same subjects also participated in a two-alternative forced-choice perception task, where they had to assign auditive stimuli to one of two words of a minimal pair; the vowel durations of the stimuli had been modified on a seven-scale continuum. The results of the production task show that the durations of short and long vowels are significantly different in both normal and fast speech rate; regarding timbre, short and long vowels of the same height display almost the same F1/F2 mean values. In the perception experiment, listeners reacted in an almost categorical manner to modifications of vowel durations; therefore, duration reveals to be a sufficient acoustic cue for the perception of vowel quantity.

## 1 Einleitung

Die distinktive Vokalquantität stellt ein typologisch markiertes Merkmal von Phoneminventaren dar, welches in etwa einem Fünftel der Sprachen der Welt vorkommt. So haben in der UPSID-Datenbank nur 62 von 317 Sprachen phonemische Daueunterschiede bei Vokalen; bei zunehmendem Umfang eines Vokalinventars steigt aber die Wahrscheinlichkeit, dass die Vokallänge distinktiv ist [1]. In gewissen Sprachen bestehen nicht für alle Vokale Quantitätsoppositionen, während in anderen Sprachen – z.B. im Standarddeutschen – Daueunterschiede mit unterschiedlichen Klangfarben einhergehen. Gerade in diesem Punkt unterscheidet sich der Zürcher Dialekt vom Standarddeutschen, da in betonter Silbe alle zehn Vokalqualitäten sowohl kurz als auch lang vorkommen [2]. Allerdings lag dazu bis jetzt nur beschränkte instrumentalphonetische Evidenz vor; immerhin hatte eine Pilotstudie mit drei Sprecher:innen ein durchschnittliches Dauerverhältnis (V/V:) von 0.56 ergeben [3].

Das Ziel dieses Beitrags liegt somit darin, die Validität der Ergebnisse der eben erwähnten Pilotstudie mit einer etwas größeren Datenbasis zu überprüfen. Deshalb wurden die in der Pilotstudie aufgenommenen Sätze zwei Jahrzehnte später mit 20 Sprecher:innen neu aufgenommen. Zudem wird in der hier vorliegenden Studie auch untersucht, welche Rolle Daueunterschiede bei der Wahrnehmung der Vokalquantität spielen. Zu diesem Zweck nahmen die gleichen 20 Gewährspersonen an einem Perzeptionsexperiment teil, bei welchem sie auditive Stimuli einem von zwei Wörtern eines Minimalpaars zuweisen mussten.

In den folgenden Abschnitten dieses Beitrags werden wir zunächst die Vokalinventare des Standarddeutschen und des Zürichdeutschen miteinander vergleichen (1.1-1.2), um die phonologischen und phonetischen Unterschiede zwischen den beiden Systemen herauszuarbeiten. Danach werden die Daten und Methoden des Produktions- und Perzeptionsexperiments beschrieben (2.1-2.5) und die Ergebnisse der darauf basierenden

Analysen vorgestellt (3.1-3.2), welche in der abschließenden Diskussion (4) kurz zusammengefasst werden.

## 1.1 Vokalqualitäten und Vokalquantität im Standarddeutschen

Die von Kohler [4] verfasste IPA-Skizze des Standarddeutschen (Abb. 1, links) geht von 16 Monophthongen aus, wobei die Vokaldauer bei den phonetischen Symbolen nur dann angegeben wird, wenn nicht gleichzeitig ein Unterschied in der Vokalqualität besteht (d.h. bei /ɛ/ ~ /ɛː/ sowie bei /a/ ~ /aː/).

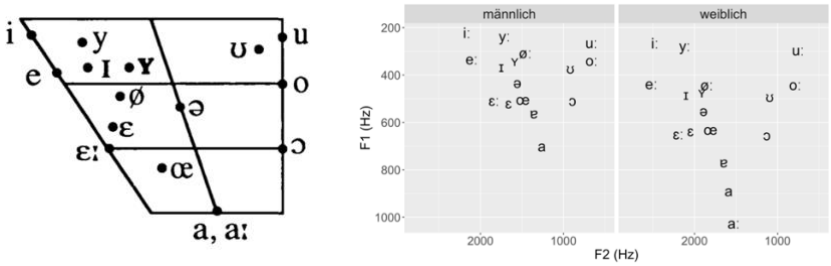



**Abbildung 1** - Symbolphonetische und akustische Darstellung der Vokale des Standarddeutschen

In den kürzlich von Kleber [5] vorgelegten Formantkarten (Abb. 1, rechts) werden hingegen die phonetischen Symbole aller Langvokale mit dem entsprechenden Längenzeichen versehen. Zudem ist ersichtlich, dass die halbhohen Kurzvokale [ɪ ʏ ʊ] sowohl bei männlichen Sprechern als auch bei weiblichen Sprecherinnen tiefer liegen als die mittelhohen Langvokale [eː øː oː].

Der Zusammenhang zwischen Vokalquantität und Vokalquantität stellt bekanntlich für die phonologische Analyse des deutschen Vokalsystems ein Problem dar, für welches unterschiedliche Lösungen vorgeschlagen wurden – etwa die Verwendung eines distinktiven Merkmals [±gespannt] [6, 7] oder die Annahme einer prosodischen Silbenschnittkorrelation [8, 9].

## 1.2 Vokalqualitäten und Vokalquantität im Zürcher Dialekt

Betrachtet man zum Vergleich die Vokalqualitäten des Zürichdeutschen, so zeigt sich sowohl symbolphonetisch (Abb. 2, links) als auch akustisch (Abb. 2, rechts) ein leicht anderes Bild.

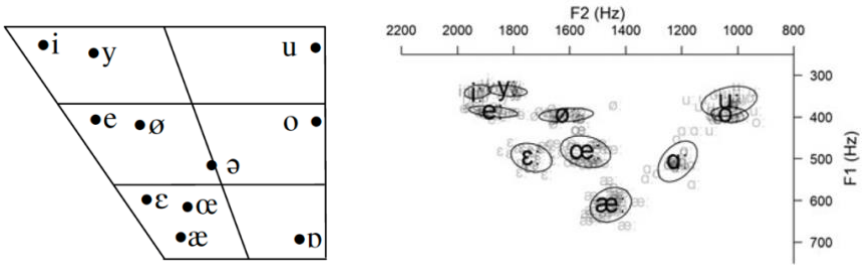



**Abbildung 2** - Symbolphonetische und akustische Darstellung der Vokale des Zürcher Dialekts

Das Vokaltrapez der IPA-Illustration [2] führt nur 11 Vokalqualitäten auf; so fehlen [ɪ ʏ ʊ] und auch die Tiefvokale haben andere Klangfarben als im Standarddeutschen. Die akustischen Messungen von Langvokalen von Zihlmann [10] ergeben ebenfalls eine leicht unterschiedliche Positionierung einzelner Vokalformanten: So fällt etwa die Nähe der mittelhohen zu den hohen Vokalen auf, zudem liegt der hintere Tiefvokal auf der Höhe der Mittelvokale. Aus phonologischer Sicht ist aber relevant, dass für alle betonten Vokale eine Quantitätsopposition besteht, die nicht mit einem Qualitätsunterschied verbunden ist. Für sieben von zehn Vokalen wird der Phonemstatus durch Minimalpaare belegt [3]; im Fall von [ø] und [u] liegen Semi-Minimalpaare vor, bei denen sich die Zielwörter zusätzlich in der Quantität des nachfolgenden

Konsonanten (*lenis* vs. *fortis*) unterscheiden: vgl. /ˈb̥øg̊ə/ ‘Bögen’ ~ /ˈb̥øːkə/ ‘Narren’, /b̥rux/ ‘Bruch’ ~ /b̥ruːɣ̊/ ‘Brauch’. Im Gegensatz zu [3] würden wir heute auch die Vokale /œ/ und /œː/ als unterschiedliche Phoneme betrachten, da – obwohl keine Minimalpaare vorliegen – die Länge des Vokals nicht aufgrund einer phonologischen Regel vorhersagbar ist (vgl. /b̥lœf/ ‘Bluff’ vs. /ʒœːf/ ‘Schaafe’).

# 2 Daten und Methoden des Produktions- und Perzeptionsexperiments

## 2.1 Teilnehmende

20 Muttersprachler:innen des Zürichdeutschen nahmen sowohl an einem Produktions- als auch an einem Perzeptionsexperiment teil. Sie waren in eine jüngere Altersgruppe zwischen 23 und 32 (Durchschnittsalter 28) und eine ältere Altersgruppe zwischen 56 und 75 (Durchschnittsalter 67) eingeteilt; beide Gruppen bestanden aus je 5 Frauen und 5 Männern.

## 2.2 Stimuli

Im Produktionsexperiment lasen die Versuchspersonen die gleichen 18 Sätze wie in der Originalstudie [3]. Die Sätze enthielten Zielwörter aus 7 Minimalpaaren und 2 Semi-Minimalpaaren (vgl. 1.2); für die beiden Vokale /œ/ und /œː/ wurden keine Daten erhoben.

Im Perzeptionsexperiment stammten die Stimuli von sieben Minimalpaaren, welche sich ausschließlich in der Vokalquantität unterscheiden. Die Stimuli waren in den Rahmensatz *Ich ha X uufgschribe* ‘Ich habe X aufgeschrieben’ eingebettet; die Hauptbetonung des Satzes lag somit immer auf der ersten Silbe des Zielwortes. Alle Sätze wurden von zwei Modellsprecher:innen (eine Frau und ein Mann) aufgenommen. Pro Minimalpaar wurde die Vokaldauer ausgehend vom Langvokal in sieben gleichmäßigen Stufen bis hin zur Dauer des Kurzvokals mithilfe eines Skripts in *Praat* [11] manipuliert. Dieser Schritt wurde für die beiden Modellsprecher:innen separat durchgeführt; dementsprechend waren die Vokaldauern der Modellsprecher: innen nicht identisch, wie aus Abb. 3 ersichtlich wird.

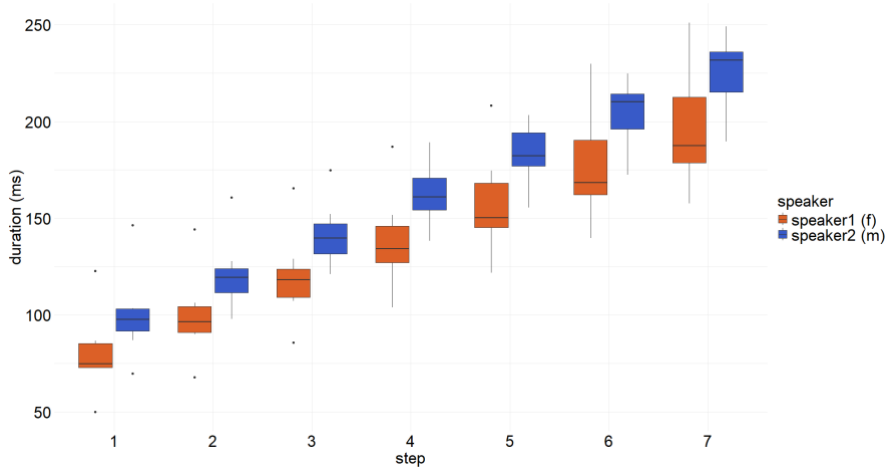



**Abbildung 3** - Vokaldauern in Millisekunden (y-Achse) der Perzeptionsstimuli für Stufen 1 (Kurzvokal) bis 7 (Langvokal) (x-Achse); Modellsprecherin 1 in orange, Modellsprecher 2 in blau

## 2.3 Ablauf

Im Produktionsexperiment lasen die Teilnehmenden die Sätze in zufälliger Reihenfolge, und zwar sowohl in normaler als auch in schneller Sprechgeschwindigkeit. Die Sätze wurden mithilfe der Software *SpeechRecorder* [12] in schweizerdeutscher Dialektschrift [13] dargestellt. Die Aufnahmen erfolgten in einer schallgedämpften Tonkabine.

Für das Perzeptionsexperiment wurde die Software *PsychoPy* [14] verwendet. Die Teilnehmenden hörten die Zielwörter im Rahmensatz in einer zufälligen Reihenfolge; allerdings folgten nie zwei Stimuli desselben Minimalpaares direkt aufeinander. Jede Stufe (1 bis 7) wurde pro Minimalpaar einmal abgespielt. Zur Ablenkung gab es zudem drei Minimalpaare, welche sich nur im wortmedialen Konsonanten (*lenis* vs. *fortis*) unterscheiden (ohne Abstufungen). Auf dem Bildschirm wurden zeitgleich beide Wörter des Minimalpaars in schweizerdeutscher Dialektschrift dargestellt. Die Teilnehmenden mussten sich für eines der beiden abgebildeten Wörter entscheiden, indem sie die linke oder die rechte Pfeiltaste auf der Tastatur drückten.

### 2.4 Datenaufbereitung

Die Audio-Dateien wurden mittels *Praat*-TextGrids manuell segmentiert und annotiert. Mithilfe eines *Praat*-Skripts wurden danach sowohl die Dauern der Vokale in Millisekunden als auch die Frequenzen für die Formanten F1 und F2 ermittelt und für alle Sprecher:innen sowie Stimuli in einer *.csv* Datei abgespeichert. Hervorgehend aus den Vokaldauern wurden zudem die Verhältnisse zwischen Kurz- und Langvokal (V/V:) berechnet.

Für die Perzeptionsergebnisse wurden pro Stimulus die von den Hörer:innen gegebenen Antworten binär kodiert (0 für ‘Kurzvokal’ und 1 für ‘Langvokal’). Auch hier wurden alle Ergebnisse in einer *.csv* Datei gespeichert.

### 2.5 Statistik

Die statistischen Analysen erfolgten in *R* [15]. Für die Produktionsdaten wurde ein lineares gemischtes Modell mithilfe der *R*-Pakete *lme4* [16] und *lmerTest* [17] gerechnet. Die Vokaldauer stellte dabei die abhängige Variable dar; die unabhängigen Variablen waren ‘Tempo’ (normal vs. schnell) und ‘Vokaltyp’ (kurz vs. lang), inklusive deren Interaktion. Zufällige Steigungen (*random slopes*) wurden für ‘Teilnehmende pro Vokaltyp’ und für ‘Wort pro Tempokondition’ hinzugefügt. Im Falle einer signifikanten Interaktion wurden paarweise Vergleiche (Tukey’s Test) mithilfe des R-Paketes *emmeans* [18] durchgeführt.

Für die Perzeptionsdaten wurde der Wert an der 50%-Grenze auf der Skala von 1 und 7 zwischen Kurz- und Langvokalen ermittelt. Dieser Wert wurde zunächst für alle Stimuli gemittelt und anschließend für die beiden Modellsprecher:innen separat berechnet. Die Perzeptionsergebnisse wurden hierbei deskriptiv ausgewertet.

## 3 Ergebnisse

### 3.1 Produktion

Die Formantkarten (F1/F2) aller Vokale für Frauen und Männer sind in Abb. 4 zu sehen. Hier wird deutlich, dass Kurz- und Langvokale jeweils sehr nah beieinanderliegen oder nahezu identische Werte vorweisen. Diese Messungen belegen somit die in 1.2 postulierte Unabhängigkeit von Vokalquantität und Vokalqualität.

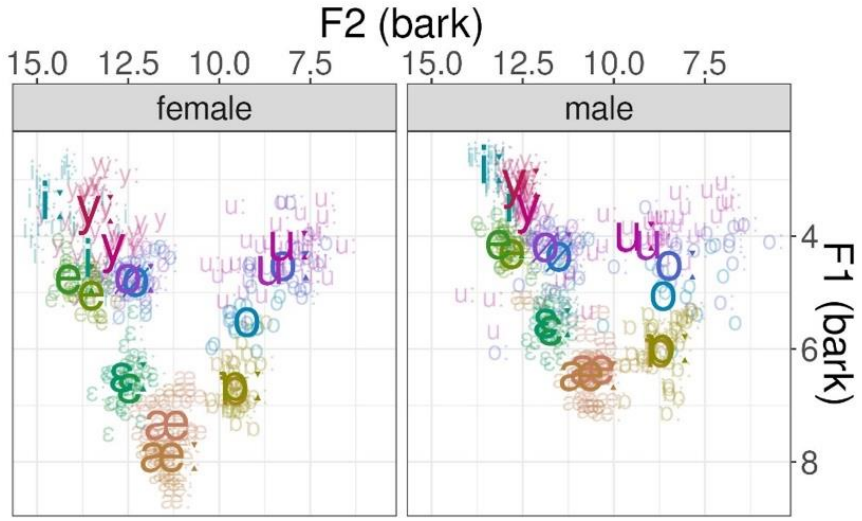



**Abbildung 4** - Vokalformanten F1 (y-Achse) und F2 (x-Achse) der weiblichen Sprecherinnen (links) sowie der männlichen Sprecher (rechts); erstellt mit *Visible Vowels* [19]

Das lineare gemischte Modell ergab, dass sich die Vokaldauern zwischen Kurz-und Langvokalen in der Produktion grundsätzlich signifikant voneinander unterscheiden. Zudem ist auch der Unterschied zwischen normalem und schnellem Sprechtempo signifikant. Darüber hinaus gibt es eine signifikante Interaktion zwischen Tempo und Vokallänge. Der Output des Modells ist in Tab. 1 abgebildet. Die durchschnittlichen Vokaldauern sind außerdem in Abb. 5 zu sehen.

**Tabelle 1** - Output des gemischten Modells für die Produktionsdaten

| | **Estimate** | **Std. Error** | **df** | **t value** | **Pr(>\|t\|)** |
|---|---|---|---|---|---|
| (Intercept) | 77.634 | 6.764 | 26.186 | 11.477 | 1.02e-11 *** |
| temponormal | 27.566 | 3.213 | 18.255 | 8.578 | 8.01e-08 *** |
| vowel_typelong | 34.305 | 9.062 | 22.325 | 3.785 | 0.000996 *** |
| temponormal:vowel_lengthlong | 38.894 | 4.531 | 18.061 | 8.583 | 8.64e-08 *** |

Aufgrund der signifikanten Interaktion zwischen Tempo und Vokallänge wurde ein paarweiser Vergleich zwischen Kurz-und Langvokal für beide Tempokonditionen einzeln berechnet. Dieser ergab einen signifikanten Unterschied zwischen Kurz- und Langvokal sowohl im normalen Sprechtempo ($z=-7.058$; $p<0.0001$) als auch im schnellen Sprechtempo ($z=-3.785$; $p=0.0002$). Hierbei erwies sich der Unterschied bei normaler Sprechgeschwindigkeit mit einem Dauerverhältnis von 0.59 als größer als bei schneller Sprechgeschwindigkeit, bei der das Dauerverhältnis 0.70 beträgt.

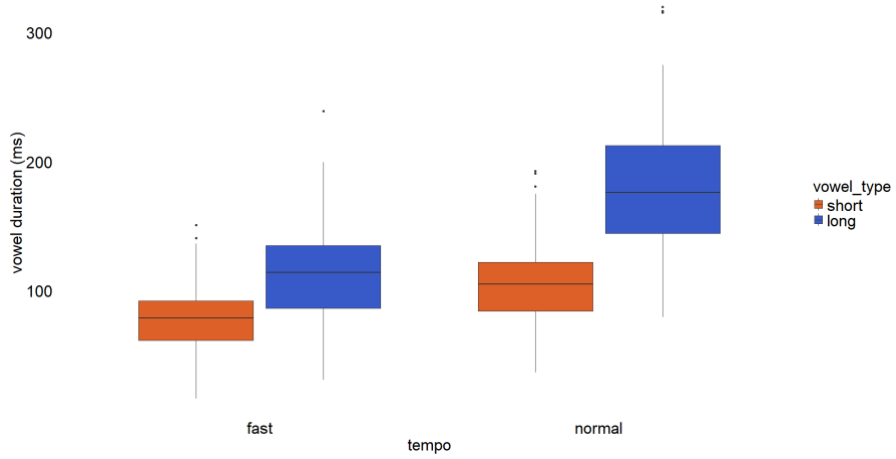



**Abbildung 5** - Durchschnittliche Vokaldauern in Millisekunden (y-Achse) für schnelles (links) sowie normales (rechts) Sprechtempo (x-Achse); orange für Kurzvokale und blau für Langvokale

## 3.2 Perzeption

In Abb. 6 wird der durchschnittliche Anteil an Antworten ‘Langvokal’ für die einzelnen Stufen dargestellt. Die 50%-Grenze zwischen Kurz- und Langvokal liegt bei 2.85. Dieser Wert ist bei 7 Stufen und damit 3.5 als Mittelpunkt relativ gering. Da die Modellsprecher:innen sich ja bezüglich ihrer Vokaldauern voneinander unterscheiden (vgl. Abb. 3), wurde zusätzlich die 50%-Grenze für beide Modellsprecher:innen separat ermittelt.

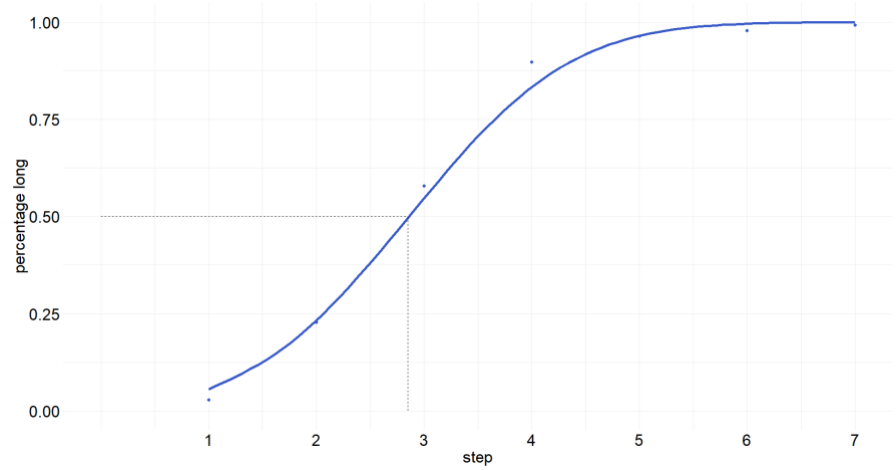



**Abbildung 6** - Prozentualer Anteil der Antworten ‘Langvokal’ für die Stufen 1 (Kurzvokal) bis 7 (Langvokal); gestrichelte Linie für die 50%-Grenze (alle Stimuli)

Die Ergebnisse für die Modellsprecher:innen separat sind in Abb. 7 ersichtlich. Daraus ist zu entnehmen, dass die Hörer:innen bei Modellsprecherin 1 deutlich später einen Langvokal wahrnahmen (bei einer 50%-Grenze von 3.17) als bei Modellsprecher 2 (bei einer 50%-Grenze von 2.51). Dieses Ergebnis ist in Übereinstimmung mit den von den Modellsprecher:innen unterschiedlich produzierten Vokaldauern, da Sprecherin 1 kürzere Dauern produziert als Sprecher 2 (vgl. Abb. 3).

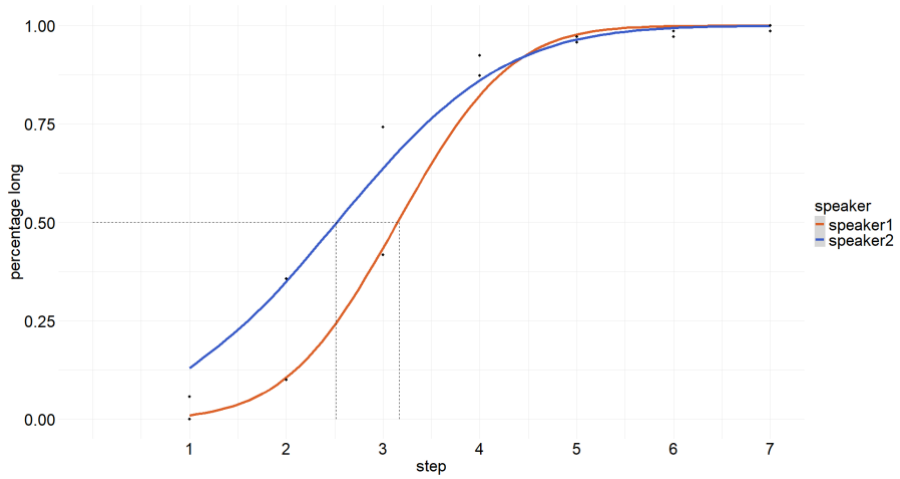



**Abbildung 7** - Prozentualer Anteil der Antworten ‘Langvokal’ für die Stufen 1 (Kurzvokal) bis 7 (Langvokal); gestrichelte Linie für die 50%-Grenze (alle Stimuli); Modellsprecherin 1 in orange, Modellsprecher 2 in blau

## 4 Diskussion und Fazit

Die Ergebnisse dieser Studie belegen die Vitalität und Stabilität der Vokalquantität im Zürcher Dialekt. Das Produktionsexperiment bestätigt insgesamt die Befunde der Pilotstudie vor Zwanzig Jahren [3], allerdings mit Daten von mehr Sprecher:innen. Kurz- und Langvokale weisen sowohl in normaler als auch in schneller Sprechgeschwindigkeit statistisch signifikante Daueunterschiede auf. Das proportionale Verhältnis der Dauern von Kurz- und Langvokalen liegt mit 0.59 in den Daten von 2024 (normale Sprechgeschwindigkeit) nur leicht über dem entsprechenden Verhältnis von 0.56 in den Daten von 2004. Die Daueunterschiede zwischen Kurz- und Langvokalen erweisen sich somit als ein synchron robustes und diachron stabiles phonetisches Korrelat der Vokalquantität im Zürcher Dialekt. Hingegen sind auch in den neuen Daten die spektralen Unterschiede zwischen Kurz- und Langvokalen für alle Vokalquantitäten entweder vernachlässigbar oder praktisch inexistent. Zum ersten Mal wurde in der vorliegenden Studie auch die Wahrnehmung der Vokalquantität im Zürcher Dialekt untersucht. Die Ergebnisse des Perzeptionsexperiment ergeben, dass die Wahrnehmung der Vokalquantität stabil ist und die Dauer als akustisches Korrelat für die Unterscheidung zwischen kurzen und langen Vokalen ausreicht. Somit stehen die Resultate des Perzeptionsexperiment im Einklang mit denjenigen des Produktionsexperiments. Aufgrund dieser phonetischen Evidenz bietet sich für die phonologische Beschreibung das Merkmal [±lang] an; der Zürcher Dialekt kann somit als ‘Quantitätssprache’ (im Gegensatz zu einer Silbenschnittsprache) bezeichnet werden.

## Dank

Diese Studie entstand im Rahmen des Forschungsprojekts *Vokal- und Konsonantenquantität in süddeutschen Varietäten*, welches vom Schweizerischen Nationalfonds finanziert wurde (Projekt-Nr. 190005).

## Literatur


[1] MADDIESON, I.: *Patterns of sounds.* Cambridge University Press, 1984.

[2] FLEISCHER, J. und S. SCHMID: *Illustrations of the IPA: Zurich German*. In *Journal of the International Phonetic Association* 36(2), S. 243-253, 2006. https://doi.org/10.1017/S0025100306002441

[3] SCHMID, S.: *Zur Vokalquantität in der Mundart der Stadt Zürich*. In *Linguistik online* 20, S. 93-116, 2004. https://doi.org/10.13092/lo.20.1065

[4] KOHLER, K.: Illustrations of the IPA: German. In *Journal of the International Phonetic Association* 20(1), S. 48–50, 1990. https://doi.org/10.1017/S0025100300004084

[5] KLEBER, F.: *Phonetik und Phonologie. Ein Lehr- und Arbeitsbuch*. Narr, Tübingen, 2023.

[6] RAMERS, K.: *Vokalquantität und –qualität im Deutschen.* Niemeyer, Tübingen, 1988.

[7] BECKER, T.: *Das Vokalsystem der deutschen Standardsprache.* Peter Lang, Frankfurt, 1998.

[8] AUER, P., P. GILLES, und H. SPIEKERMANN: *Silbenschnitt und Tonakzente.* Niemeyer, Tübingen, 2002.

[9] RESTLE, D.: *Silbenschnitt – Quantität – Kopplung.* Fink, München, 2003.

[10] ZIHLMANN, U.: *Vowel quality in four Alemannic dialects and its influence on the respective varieties of Swiss Standard German*. In *Journal of the International Phonetic Association* 53(1), S. 1-28, 2023. https://doi.org/10.1017/S0025100320000377

[11] BOERSMA, P. und D. WEENINK: *Praat* (Version 6.2.10) [Software], 2023.

[12] DRAXLER, C. und K. JÄNSCH: *SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software*. In *Proc. of the IV International Conference on Language Resources and Evaluation*, S. 559–562, 2004.

[13] DIETH, E.: *Schwyzertütschi Dialäktschrift* (Ed 2.), Verlag Sauerländer, Aarau, 1986.

[14] PEIRCE, J. W., J. R. GRAY, S. SIMPSON, M. R. MACASKILL, R. HÖCHENBERGER, H. SOGO, E. KASTMAN, und J. LINDELØV.: *PsychoPy2: experiments in behavior made easy*. In *Behavior Research Methods*, 2019.

[15] RSTUDIO TEAM: *RStudio: Integrated Development for R. RStudio* [Computer software], 2020.

[16] BATES, D., M. MÄCHLER, B. BOLKER, und S. WALKER: *Fitting Linear Mixed-Effects Models Using lme4*. In *Journal of Statistical Software 67*(1), S. 1-48, 2015. https://doi.org/10.18637/jss.v067.i01

[17] KUZNETSOVA, A., P.B. BROCKHOFF, und R.H.B. CHRISTENSEN*: lmerTest Package: Tests in Linear Mixed Effects Models.* In: *Journal of Statistical Software*, *82*(13), S. 1–26, 2017. https://doi.org/10.18637/jss.v082.i13

[18] LENTH, R. V.: *emmeans: Estimated Marginal Means, aka Least-Squares Means* [Computer software], 2021. https://CRAN.R-project.org/package=emmeans

[19] HEERINGA, W. und H. VAN DE VELDE: *Visible Vowels: A tool for the visualization of vowel variation*. In *Proceedings CLARIN annual conference*, S. 8-10, 2018.

# HOMOPHONOUS SEMANTIC MINIMAL PAIRS DIFFER IN THEIR SUBPHONEMIC ACOUSTIC DURATIONS: THE CASE OF GENERIC AND SPECIFIC MASCULINES IN GERMAN

*Dominic Schmitz*
*Heinrich-Heine-Universität Düsseldorf*
*dominic.schmitz@uni-duesseldorf.de*

**Abstract:** Previous research showed that subphonemic durational variation is modulated by lexical and morphological differences. The present study takes research on subphonemic differences one step further and asks: are there subphonemic durational differences in phonologically, morphologically, and lexically identical members of semantic minimal pairs? To answer this question, a sentence reading task on German was conducted. Target words were generic masculines and specific masculines ending in *-er*. Items were preceded by a context and embedded in a sentence, with similar contexts and sentences for the forms of the same item. 40 participants produced 30 targets each. The duration of the *-er* suffix, i.e. /ɐ/, was analysed in linear mixed-effects regression models. The results showed that generic masculines come with longer /ɐ/ durations than specific masculines. These findings add to the body on subphonemic durational differences unaccounted for by established theories of speech production.

## 1 Subphonemic differences and the generic masculine in German

Previous research identified subphonemic durational differences conditioned by either morphological or lexical differences. Most prominently, word-final /s/ in English shows different durations depending on its morphological nature. Non-morphemic /s/ appears to be longest, followed by suffix /s/, which in turn is followed by auxiliary clitic /s/ [1]. Further durational differences have been found, for example, in prefixes [2], in homophonous free and bound (pseudo-)stems [3], and in homophonous words [4].

Such subphonemic durational differences are unexpected from a traditional theoretical view, as morphological and lexical information is assumed to be unavailable to speech production [5]. Psycholinguistic models of speech production similarly assume that morphological information is no longer available once the phonological stage is reached [6]. Exemplar-based models may account for subphonemic durational differences, suggesting that the experienced durational differences are stored and with that also produced [7], but lack an explanation as to how exactly these durational differences come into existence in the first place. Thus far, it is the Discriminative Lexicon Model and its computational implementations [8] that are able to offer explanations for these durational differences based on meaning- and form-related features of the elements under investigation [9, 10].

As evidence on subphonemic durational differences caused by morphological and lexical differences amasses, the present study takes research on such fine-phonetic differences one step further and asks whether are there subphonemic durational differences in phonologically, morphologically, and lexically identical members of semantic minimal pairs.

With specific and generic masculines, German–among other languages–offers a good testcase to gain first insights into whether fine-semantic differences also cause fine-phonetic differences. In German, grammatically masculine role nouns with feminine counterparts may not

only be used to refer to male individuals but also to individuals of any other gender [11]. If such a masculine is used to refer to a male individual, e.g. *Max ist Lehrer* 'Max is a teacher', this usage is called *specific*. If, however, the same role noun is used to refer to an individual of another gender, e.g. *Tina ist Lehrer* 'Tina is a teacher', or to a referent of undeclared gender, e.g. *mein Kind ist Lehrer* 'my child is a teacher', this usage is called *generic*. Henceforth, I will call the generic usage referring to gender-specified referents *specified generic* and the generic usage referring to referents with undeclared gender *unspecified generic*.

*Lehrer*, just as many other role nouns in German, ends in the suffix *-er*, produced as /ɐ/, as this suffix is attached to the stem of a verb to derive a role noun.[1] All role nouns derived this way are grammatically masculine and come with a feminine counterpart.[2]

Making use of such specific masculines, specified generic masculines, and unspecified generic masculines, the present study aims to find answers to the following two research questions:

**RQ1** Does the fine-semantic difference between specific masculines and generic masculines lead to fine-phonetic durational differences in word-final *-er*?

**RQ2** Does the difference in specification between specified generic masculines and unspecified generic masculines lead to fine-phonetic durational differences in word-final *-er*?

## 2 Methodology

### 2.1 Materials

For the present study, a sentence reading task was conducted. Items were embedded in simple target sentences, all of which were preceded by a context phrase or sentence setting the scene. Both phrases/sentences in combination made clear whether a masculine role noun was used specifically or generically, and context sentences did not differ between different role noun forms of the same target. That is, except for the person who was referred to, as this was required to obtain the different types of role noun readings. An illustration is given in Example 1, with changing referents and related pronouns in italics and the target in bold.[3]

Ex.1 *Matteos Vater/Mein Kind/Marlenes Mutter* kann richtig gut nähen. *Er/Es/Sie ist* **Schneider** von Beruf.
*Matteo's dad/My child/Marlene's mum* is really good at sewing. *He/It/She* is a **tailor** by profession.

Items were 20 role nouns ending in *-er*, 10 of which were stereotypically female, while the other 10 were stereotypically male. All were taken from the list provided by [12], which includes information on stereotypicality. Table 1 lists all items.

### 2.2 Procedure

During the reading task, participants were presented with one set of context and target phrase/ sentence at a time. They were instructed to first read both quietly, before then producing them aloud. The experiment progressed in a self-paced fashion, and participants were told that progressing quickly was not the main objective of their task. Each participant produced 30 sentences with target items, i.e. 10 sentences per type of masculine, and additionally 10 sentences in which the role noun was grammatically feminine. The latter functioned as filler items.

[1] Henceforth I will use *-er* and /ɐ/ interchangeably.

[2] The feminine counterpart is formed by attaching the feminine suffix *-in*, e.g. *Lehrerin* 'female teacher'.

[3] Find a full list at https://osf.io/dvrq4/.

**Table 1** – Overview of all target words, including English translations.

| | | | | |
|---|---|---|---|---|
| **stereotypically female** | | | | |
| *Balletttänzer* | *Eiskunstläufer* | *Flugbegleiter* | *Geburtshelfer* | *Haushälter* |
| *(ballet dancer)* | *(ice skater)* | *(flight attendant)* | *(obstetrician)* | *(housekeeper)* |
| *Hellseher* | *Kosmetiker* | *Pfleger* | *Schneider* | *Verkäufer* |
| *(clairvoyant)* | *(beautician)* | *(caregiver)* | *(tailor)* | *(vendor)* |
| **stereotypically male** | | | | |
| *Bauarbeiter* | *Elektriker* | *Fußballspieler* | *Kranführer* | *Maurer* |
| *(construction worker)* | *(electrician)* | *(footballer)* | *(crane operator)* | *(mason)* |
| *Programmierer* | *Rennfahrer* | *Reporter* | *Schreiner* | *Wahrsager* |
| *(programmer)* | *(race driver)* | *(reporter)* | *(carpenter)* | *(fortune-teller)* |

After the reading task, participants were asked to provide information on their gender, age, and attitude towards the use of generic masculines, pair forms, and gender star forms. Pair forms are phrases consisting of both the masculine and feminine form, e.g. *Lehrerin und Lehrer*; gender star forms constitute a novel role noun form making use of a new allegedly gender-neutral suffix, e.g. *Lehrer*in*. Attitudes towards the three forms were given using a 5-point Likert scale.

### 2.3 Participants

40 participants with German as L1 took part in the experiment. The mean age of the participants was 29.1 years, with an age-range from 20 to 64. Half of the participants were recruited at the author's institution, whereas the other half were recruited online.

### 2.4 Acoustic and statistical analyses

The acoustic analyses of all recordings took place in Praat [13]. For all targets, the duration of *-er*, i.e. /ɐ/, was determined using both the spectrogram and the waveform, following segmentation criteria based on the phonetic literature [14]. All boundaries were placed at the nearest zero crossing. Recordings with production errors, stutter, and laughter were excluded. Overall, 1113 of 1200 recordings were retained for further analysis. Based on the annotations in Praat, durations were extracted using the rPraat package [15] in R [16].

The duration of word-final *-er* was statistically analysed following the analysis of word-final *-s* in English as conducted by Schmitz et al. [17].[4] That is, the duration of *-er* was predicted in linear mixed-effects regression models using the lme4 package [18] in R. As fixed effects, the following variables were included:

TYPEOFER. This is the variable of interest. It encodes whether a word-final *-er* is that of an unspecified generic masculine (`GMu`), a specified generic masculine (`GMs`), or a specific masculine (`SM`).

DURBASE. Indicating a local speaking rate [e.g. 1], base duration was measured. Base duration here is equal to the summed duration of all word-internal segments preceding the *-er*.

PRETYPE, FOLTYPE. Different types of preceding and following segments may result in different co-articulatory effects. Hence, the type of the preceding and following speech sound was included in the analysis. These variables take the values `fricative`, `liquid`, `nasal`, `plosive`, and `vowel`.

SYLNUMBER. The number of syllables was included to account for effects of syllable shortening [19]. The number was assessed by counting the prototypical number of syllables of

[4] Find the relevant data and R script at https://osf.io/dvrq4/.

the target words. As participants were instructed to read clearly and in an average pace, this prototypical number of syllables is also the number that was produced.

NUMBER. As it is unclear whether the number of the target word makes a difference regarding the duration of word-final *-er*, this information was included in the analysis as a binary variable with the values `singular` and `plural`.

STEREOTYPICALITY. Stereotypicality was included in the analysis, as it might make a difference whether a given masculine noun is stereotypically female or male. Stereotypicality data are adopted from [12].

SPEECHRATE. As speech rate typically shows a direct influence on the duration of segments, where a higher speech rate comes with generally shorter segments, this variable was incorporated into the analysis. Speech rate was measured using a Praat script by [20].

TRIALNUMBER. The number of the trial was included to account for potential effects of priming, training, and fatigue. The order of trials was fully randomised.

GENDER. As the usage of generic masculines is directly connected to the notion of gender, speaker gender was part of the model formula. This variable takes the values `female`, `male`, and `nonbinary`.

AGE. Speaker age was included for two reasons. First, it may show an effect on phonetic realisation. Second, age may act as a proxy for the attitude towards more progressive topics, such as gender and language.

ATTGM, ATTPF, ATTST. To get a more direct idea of a speaker's attitude towards gender and language and its potential influence on phonetic realisation, speakers' ratings on the usage of generic masculines, pair forms, and gender star forms were included. Each of the three variables takes integer values of the range $[1,5]$, where 1 represents the opinion that the usage of a given form is *very bad* and 5 represents the opinion that the usage of a given form is *very good*.

Further, the following variables were introduced as random intercepts:

SPEAKER. To account for inter-speaker differences not accounted for by the already introduced speaker-dependent variables, the ID of each speaker was specified as a random intercept term.

WORD. Similarly, to account for inter-word differences not reflected in the already introduced variables, the word was specified as a random intercept term.

To avoid issues of collinearity that might lead to unreliable model estimates [9], all predictor variables were checked for problematic levels of correlation (i.e., $|\rho| \geq 0.5$). Strong correlations were found for DURBASE and SYLNUMBER ($\rho = 0.75$), and ATTST and ATTGM ($\rho = -0.67$). To proceed, SYLNUMBER and ATTGM were disregarded.

Then, continuous variables were checked for their distributions. It was found that neither `durEr` nor `durBase` nor `speechRate` followed a normal distribution. To achieve a more normal distribution and, with that, a better model fit, `durEr` was log-transformed. The other two variables were still not normally distributed after log-transformations, and thus remained in their original forms.

With the retained and transformed variables, two initial linear mixed-effects regression models were fitted to log-transformed dependent variables. One was fitted to the absolute duration of word-final *-er*, `durEr`, and one was fitted to the relative duration of word-final *-er*, `durErRel`. The latter model was included following previous studies on the duration of different types of word-final /s/ [1, 17, 21] and is motivated by the finding that differences in duration may play out differently in absolute and relative segment duration [22]. For both dependent variables, the initial model was specified with the aforementioned fixed effects and random

intercepts.

The initial model was then reduced stepwise using the `step` function provided by the lmerTest package [23]. This stepwise reduction process fits models repeatedly, performing a combination of forward selection and backward elimination, adding and removing predictors iteratively. The fitted models are compared via their AIC values. The result of this process is the model formula resulting in the lowest AIC, i.e. best model fit.

Finally, the reduced model was trimmed based on its residuals to obtain an even better model fit [24]. Data points with residuals further than 2.5 standard deviations from the mean were removed. The result of this last step were the final models.

## 3 Results

The final model fitted to the absolute duration of *-er* contained only the predictor of interest, TYPEOFER, and the two random intercepts, SPEAKER and WORD. The model fitted to the relative duration of *-er* additionally contained DURBASE as fixed effect. During the residual trimming process of both models, $n = 19$ data points (1.7%) were removed. Table 2 provides summaries for both models.

**Table 2** – Summary of the final model fitted to the absolute *-er* durations.

| **Absolute duration model** | **Estimate** | **Std. Error** | **df** | ***t*-value** | ***p*-value** |
|---|---|---|---|---|---|
| Intercept | -2.483 | 0.038 | 46.52 | -65.580 | <0.001 |
| TYPEOFERGMu | <0.001 | 0.018 | 1036.00 | -0.052 | 0.959 |
| TYPEOFERSM | -0.250 | 0.018 | 1036.00 | -13.998 | <0.001 |
| **Relative duration model** | **Estimate** | **Std. Error** | **df** | ***t*-value** | ***p*-value** |
| Intercept | -2.494 | 0.055 | 106.0 | -45.658 | <0.001 |
| TYPEOFERGMu | 0.001 | 0.018 | 1035.0 | 0.056 | 0.955 |
| TYPEOFERSM | -0.249 | 0.018 | 1037.0 | -13.924 | <0.001 |
| DURBASELOG | -1.015 | 0.052 | 319.8 | -19.336 | <2e-16 |

Bonferroni-corrected pairwise comparisons for the levels of TYPEOFER in both models showed that the durational differences between the specific masculine and either of the generic masculines was highly significant ($p < 0.001$), while the difference between the specified and unspecified generic masculine was not ($p = 1$). The effect size of TYPEOFER is large with $\eta^2 = 0.20$ with 95% $CI = [0.17, 1.00]$ for the absolute duration and with $\eta^2 = 0.20$ with 95% $CI = [0.48, 1.00]$ for the relative duration of *-er*.

The effects of TYPEOFER and DURBASE are illustrated in Figure 1. In both models, specific masculines come with significantly shorter *-er* durations than both types of generic masculines. For DURBASE it is found that shorter base durations come with shorter *-er* durations.

## 4 Discussion

The present study set out to find a first answer to whether fine-semantic differences in words with identical lexical and morphological makeup lead to fine-phonetic differences. Analysing the productions of specific masculines, specified generic masculines, and unspecified generic masculines embedded within disambiguating sentences, evidence was found in favour of such fine-phonetic differences caused by fine-semantic differences.

**RQ1**, which asked whether there are durational differences between specific masculines and generic masculines, may be answered positively. In both absolute and relative durations,

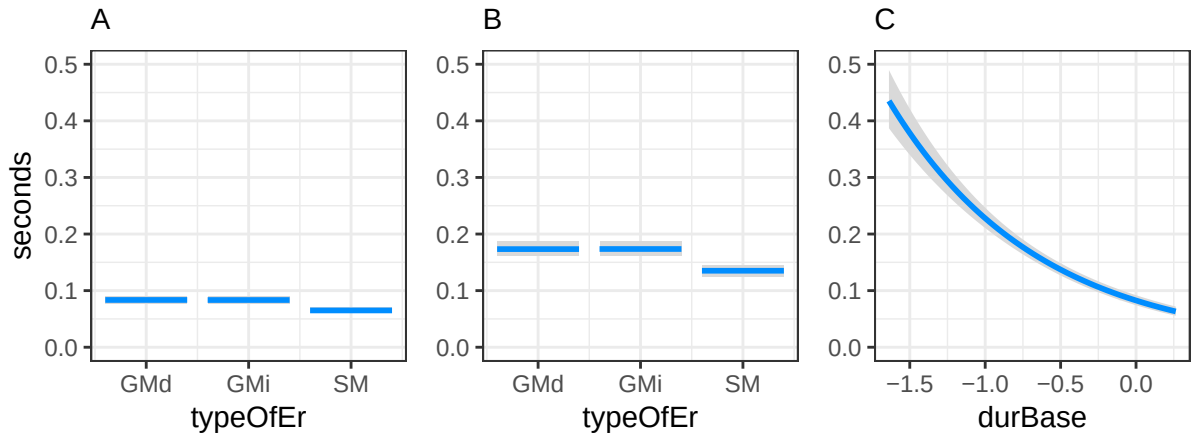



**Figure 1** – Partial effects of TYPEOFER as predicted by the model fitted to absolute *-er* duration (Panel A) and fitted to relative *-er* duration (Panel B), as well as the partial effect of DURBASE in the model fitted to relative *-er* duration (Panel C).

*-er* is produced significantly longer when it is part of a generic masculine than when it is part of a specific masculine.

Regarding **RQ2**, which asked whether durational differences are also to be found between specified and unspecified generic masculines, no such difference was found. Both specified and unspecified generic masculines are produced with highly similar durations.

Taking the results together, it appears that the fine-semantic difference between specific and generic usage leads to a durational difference in a role noun's word-final suffix, i.e. in *-er*. Whether the generic usage is referring to a gender-specified individual or to an individual whose gender is not specified seems to not make a difference.

These novel findings raise the question of why these differences exist. One potential explanation may lie in the cognitive load required to parse specific and generic masculines. While specific masculines are relatively easily parsed–a masculine form is used to refer to a male individual–generic masculines are not. Here, a grammatically masculine form that may be used to refer to male individuals is used to refer to non-male individuals, even though a feminine counterpart for referring to female individuals exists. Comprehension efforts are assumably higher ['cognitive markedness', cf. 25], which in turn may lead to longer durations in production. One potential framework to test this idea in is that of the Discriminative Lexicon [8].

Finally, the present findings add to the results of previous research on subphonemic durational differences caused by morphological or lexical differences. That is, not only lexical or morphological differences in phonologically identical words or segments lead to fine-phonetic durational differences, but so do fine-semantic differences in lexically, morphologically, and phonologically identical words. This, then, adds to the evidence calling into question assumptions made by traditional feed-forwards models and psycholinguistic models of speech production [5, 6], as these do not account for such differences.

## 5 Conclusion

This study was the first to investigate whether fine-semantic differences cause fine-phonetic differences in lexically, morphologically, and phonologically identical words. Making use of specific and generic masculines in German in a reading task, it was found that such fine-semantic differences come with subphonemic durational differences. These findings add to the body of evidence that calls into question established models of speech production. The results call not only for a revision of established models of speech production, but also for further research in this area, exploring why these differences exist.

## References


[1] PLAG, I., J. HOMANN, and G. KUNTER: *Homophony and morphology: The acoustics of word-final S in English*. *Journal of Linguistics*, 53(1), pp. 181–216, 2017.

[2] HEDIA, S. B. and I. PLAG: *Gemination and degemination in english prefixation: Phonetic evidence for morphological organization*. *Journal of Phonetics*, 62, pp. 34–49, 2017.

[3] BLAZEJ, L. J. and A. M. COHEN-GOLDBERG: *Can we hear morphological complexity before words are complex? Journal of Experimental Psychology: Human Perception and Performance*, 41(1), p. 50, 2015.

[4] LOHMANN, A.: *Time and thyme are not homophones: A closer look at gahl's work on the lemma-frequency effect, including a reanalysis*. *Language*, 94, pp. e180–e190, 2018.

[5] KIPARSKY, P.: *Lexical morphology and phonology*. In I.-S. YANG (ed.), *Linguistics in the morning calm: Selected papers from SICOL1*, pp. 3–91. 1982.

[6] LEVELT, W. J., A. ROELOFS, and A. S. MEYER: *A theory of lexical access in speech production*. *Behavioral and brain sciences*, 22(1), pp. 1–38, 1999.

[7] PIERREHUMBERT, J. B.: *Exemplar dynamics: Word frequency, lenition and contrast*. *Typological studies in language*, 45, pp. 137–158, 2001.

[8] CHUANG, Y.-Y. and R. H. BAAYEN: *Discriminative learning and the lexicon: NDL and LDL*. *Oxford Research Encyclopedia of Linguistics*, 2021.

[9] TOMASCHEK, F., P. HENDRIX, and R. H. BAAYEN: *Strategies for addressing collinearity in multivariate linguistic data*. *Journal of Phonetics*, 71, pp. 249–267, 2018.

[10] SCHMITZ, D., I. PLAG, D. BAER-HENNEY, and S. D. STEIN: *Durational differences of word-final /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning*. *Frontiers in Psychology*, 12, 2021.

[11] DIEWALD, G.: *Zur Diskussion: Geschlechtergerechte Sprache als Thema der germanistischen Linguistik – exemplarisch exerziert am Streit um das sogenannte generische Maskulinum*. *Zeitschrift für germanistische Linguistik*, 46, pp. 283–299, 2018.

[12] MISERSKY, J., P. M. GYGAX, P. CANAL, U. GABRIEL, A. GARNHAM, F. BRAUN, T. CHIARINI, K. ENGLUND, A. HANULIKOVA, A. ÖTTL, J. VALDROVA, L. V. STOCKHAUSEN, and S. SCZESNY: *Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak*. *Behavior Research Methods*, 46, pp. 841–871, 2014.

[13] BOERSMA, P. and D. WEENINK: *Praat: Doing phonetics by computer*. 2019. URL `http://www.praat.org/`.

[14] LADEFOGED, P.: *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Wiley, 2003.

[15] BORIL, T. and R. SKARNITZL: *Tools rPraat and mPraat*. In P. SOJKA, A. HORÁK, I. KOPECEK, and K. PALA (eds.), *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*, pp. 367–374. Springer International Publishing, Cham, 2016.

[16] TEAM, R. C.: *R: A language and environment for statistical computing*. 2021. URL https://www.r-project.org/.

[17] SCHMITZ, D., D. BAER-HENNEY, and I. PLAG: *The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords*. *Phonetica*, 78(5-6), pp. 571–616, 2021.

[18] BATES, D., M. MÄCHLER, B. M. BOLKER, and S. C. WALKER: *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software*, 67, 2015.

[19] SCHMITZ, D., H.-E. CHO, and H. NIEMANN: *Vowel shortening in German as a function of syllable structure*. In *Proceedings 13. Phonetik und Phonologie Tagung (P&P13)*, pp. 181–184. 2018.

[20] DE JONG, N. and T. WEMPE: *Praat script syllable nuclei [Praat script]*. 2008. URL https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2.

[21] SCHMITZ, D. and D. BAER-HENNEY: *Morphology renders homophonous segments phonetically different: Word-final /s/ in German*. In *Proceedings of Speech Prosody 2024 (SP2024)*. Universiteit Leiden, Netherlands, 2024.

[22] BAAYEN, R. H., Y.-Y. CHUANG, E. SHAFAEI-BAJESTAN, and J. P. BLEVINS: *The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning*. *Complexity*, 2019, p. 4895891, 2019.

[23] KUZNETSOVA, A., P. B. BROCKHOFF, and R. H. B. CHRISTENSEN: *lmertest package: Tests in linear mixed effects models*. *Journal of Statistical Software*, 82, 2017.

[24] BAAYEN, R. H. and P. MILIN: *Analyzing reaction times*. *International Journal of Psychological Research*, 3, pp. 12–28, 2010.

[25] HASPELMATH, M.: *Against markedness (and what to replace it with)*. *Journal of Linguistics*, 42(1), p. 25–70, 2006. doi:10.1017/S0022226705003683.

# SPRACHGEOGRAPHISCHE VERTEILUNG DES *SEGMENTAL ANCHORING* IN STANDARDINTENDIERTER LESEAUSSPRACHE DES DEUTSCHEN – EINE PILOTSTUDIE

*Beat Siebenhaar*

*Universität Leipzig, Institut für Germanistik*
*siebenhaar@uni-leipzig.de*

**Kurzfassung:** Der Beitrag stellt eine Pilotstudie zur Sprachgeographie der Intonation in der Leseaussprache im gesamten zusammenhängenden deutschen Sprachgebiet vor. Für die sprachgeographische Verteilung kann auf das Deutsch-heute-Korpus zurückgegriffen werden. Aus dem da aufgenommen *Nordwind-und-Sonne*-Lesetext wird hier eine einzelne Stelle phonetisch analysiert, die von fast allen Sprechern mit einer steigenden Nukleussilbe (je nach Ansatz und Position als L+H* oder L*+H interpretiert) und dann einer Weiterweisung mit einer gleichbleibend hohen bzw. steigenden Kontur realisiert wurde. Dabei zeigt sich für die Dauer der Silbe eine längere Dauer im Süden. Der Anfang der Akzentuierung, bestimmt als F0-Minimum, wird auf einer geographischen Nordwest-Südost-Achse nach hinten verschoben. Die Position des F0-Maximums erweist sich als problematischer, da dessen Funktion weniger einheitlich ist. Die Darstellung dieser Ergebnisse dient daher eher der Problematisierung als einer resultierenden Festlegung. Die Daten zur Silbenlänge bestätigen frühere Ergebnisse. Die Ergebnisse zum F0-Minium sind dagegen neu und konsistent. Damit zeigt das Pilotprojekt, dass die Methode in dieser Beziehung weitergeführt werden kann.

## 1 Forschungseinblick und Forschungsfragen

Forschung zur arealspezifischen Prosodie des Deutschen ist noch immer eher spärlich zu finden. Nach den großräumig angelegten Arbeiten zur substandardsprachlichen Intonationskontur von Gilles [3], Peters [14] und Pistor [15] sowie Kügler [10] zum Sächsischen und Schwäbischen sind einige Aufsätze erschienen, die meist zwei oder drei Regionen vergleichen: Atterer/Ladd [1] und Kleber/Rathcke [8] zum segmental anchoring im regionalen Standard, Gilles [4] zum Vergleich der Intonationkonturen im luxemburgisch-moselfränkischen Grenzgebiet oder Leemann [11] mit Fujisaki-Modellierung der Intonation von vier Schweizer Dialekten. Ähnlich dürftig sieht die Situation in Bezug auf die temporale Verteilung aus. Eine feingliedrige sprachgeographische Darstellung der temporalen Organisation der Schweizer Dialekte auf der Basis von crowdsourced Aufnahmen bietet Leeman [12]. Die zeitliche Strukturierung des gesamten deutschsprachigen Raums hat Hahn [5] in der Monographie zur Sprachgeographie von Sprechgeschwindigkeit und Reduktion im Blick. Sprachgeographisch orientierte Untersuchungen zur Intonation bzw. zur F0-Kontur, die den gesamten deutschen Sprachraum berücksichtigen, sind bislang nicht publiziert worden.

Mit dem vorliegenden Pilotprojekt wird überprüft, ob die bestehenden Aufnahmen des Deutsch-heute-Projekts [9], welche für das SpuRD-Projekt [6] segmentiert worden sind, für eine automatisierte sprachgeographische Analyse, die sich an [1, 8] orientiert, genutzt werden können. Dazu werden erste Resultate präsentiert. Folgende Fragen stehen im Zentrum:

1. Gibt es eine sprachgeographische Strukturierung der Silbenlänge?
2. Gibt es eine sprachgeographische Verteilung der zeitlichen Binnengliederung der Silbe?
3. Gibt es eine sprachgeographische Strukturierung der segmentalen Verankerung des F0-Tiefpunktes der Silbe und des F0-Hochpunktes der Silbe?
4. Ist die angesetzte Methode brauchbar, um Intonationsmuster im Raum zu erkennen?

## 2 Datenbasis

Datengrundlage für die Untersuchung bietet der Lesetext *Der Nordwind und die Sonne*, welcher für das Deutsch-heute-Projekt [9] zwischen 2006 und 2009 aufgenommen worden ist und für das SpuRD-Projekt [6] aufgearbeitet wurde. Aus dem gesamten Korpus werden nur die Aufnahmen in normalem Lesetempo der jungen männlichen Probanden analysiert. Das sind 327 Aufnahmen von Abiturienten aus 165 gleichmäßig über das deutsche Sprachgebiet verteilten Orten. Üblicherweise wurden je zwei Sprecher pro Ort aufgenommen.

Für die Pilotstudie, die (quick and dirty) einen quantitativen und möglichst automatisierten Ansatz verfolgt, wurde eine Textstelle ausgewählt, die intonatorisch normalerweise als Weiterweisung mit dem Akzentton (L+H* oder L*+H) auf der Nukleussilbe und einem hohen Grenzton realisiert worden ist. Es ist dies die Stelle *als ein Wanderer*, die somit im gesamten Korpus vergleichbar ist. Zentral für die Untersuchung ist die Nukleussilbe [ˈvan].

## 3 Methode

### 3.1 Datenerhebung

Die Datenerhebung ist in [9] beschrieben. Die Abiturienten wurden in einem ruhigen Raum der Schule mit Headsetmikrophonen aufgenommen. Zur Datenaufbereitung wurde jede Aufnahme über das Online-Tool WebMAUS [7] auf der Lautebene automatisch vorsegmentiert und anschließend manuell ergänzt und korrigiert (zum genauen Verfahren [5]).

### 3.2 Anteile der Laute an der Silbe, Normalisierung auf 100% (siehe Abb. 1)

Für die vorliegende Untersuchung wird die Dauer der Nukleussilbe [ˈvan] berechnet sowie die Anteile von Silbenonset, -nucleus und -coda. Die Dauer der Segmente kann aus der manuellen Segmentierung übernommen werden. Das Silbenende wird regulär vor dem Burst von [d] gesetzt. So wird die Schwierigkeit, dass die meist stimmhafte Okklusionsphase von [d] und [n] oft nicht zu trennen sind, systematisch gelöst. Wenn [d] ganz elidiert wird, wird das Ende von [n] als Silbengrenze gesetzt. In Einzelfällen sind auch [n] und [v] am Silbenanfang kaum unterscheidbar oder vollständig zu [n] oder [m] assimiliert. In diesen Fällen wird die Silbengrenze beim Vokalbeginn gesetzt. Um die Intonationsparameter bei unterschiedlicher Sprechgeschwindigkeit [5] vergleichbar zu machen, wird die Silbendauer, die zwischen Messwerten von 140 und 340 ms schwankt, auf 100% normalisiert. Das Verfahren orientiert sich an [1] und [8].

### 3.3 Berechnung des Tiefs (L) und des Hochs (H) je in % der Silbendauer (siehe Abb. 1)

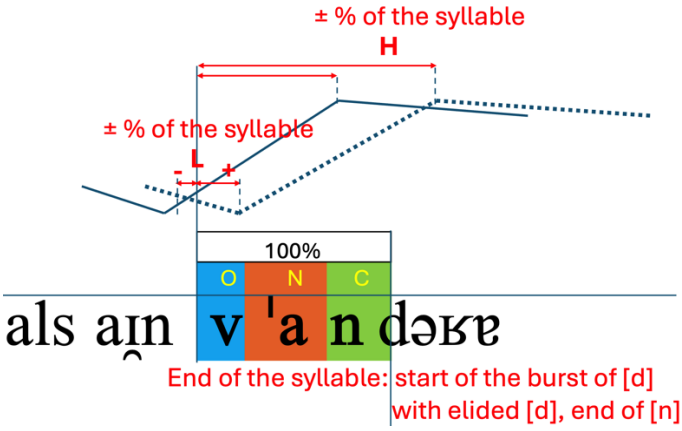



**Abbildung 1** – Messpunkte für die Studie

Die Markierungen von L und H werden auf der Basis der F0-Extraktion von Praat [2] gesetzt. Der Tiefton (L) wird auf das lokale F0-Minimum der Kontur um den Beginn der akzentuierten Silbe gelegt. Der Hochton (H) wird da markiert, wo in 10 ms kein weiterer Anstieg von mehr als 1 Hz erfolgt. In Einzelfällen wird in der Silbe und auch in der Folgesilbe kein Maximum erreicht. Die Weiterweisung wird also mit einem kontinuierlichen Anstieg bis zur Phrasengrenze realisiert, wie sie Gilles [3] mit

unterschiedlicher räumlicher Geltung für das gesamte Deutsche beschrieben hat. Wegen ihrer abweichenden Funktion werden diese Belege für die Analyse von H nicht berücksichtigt. Selten ist in der Phrase auch auch kein L sichtbar, weshalb auch diese Belege ausgeschlossen werden. Abb. 1 zeigt die Markierungsmethode. Um die Daten zu ver-gleichen, werden die beiden Messpunkte auf den Silbenbeginn referenziert und der Abstand zum Silbenbeginn als Prozentwert der Gesamtsilbendauer berechnet.

### 3.4 Berechnung der Mittelwerte pro Ort und deren Abbildung im Raum.

Zuerst wird gezeigt, dass bei einem Half-Split-Test, für den die jeweils höheren Werte eines Ortes in eine Gruppe und die tieferen in die andere Gruppe zusammengefasst werden, für beide Gruppen räumlich ähnliche Strukturen erscheinen. Die areale Verteilung ist damit bei örtlich individuellen Unterschieden im Wesentlichen also gegeben, so dass die Daten pro Ort gemittelt werden können und konsistentere Werte erreicht werden. Dafür werden die Silbendauer, sowie Positionswerte für L und H übernommen und jeweils der Mittelwert für jeden dieser drei Werte berechnet. Die Normalisierung der Dauer erfolgt auf der Basis dieser Mittelwerte pro Ort.

Um die Daten im im geographischen Raum zu visualisieren, werden für jede Analyse fünf gleich große Klassen gebildet. Die Ortspunkte werden nach diesen Klassen farblich von blau über grün nach rot dargestellt. Weil die einzelnen Punkte durch individuelle Variation geprägt sein können und pro Ort meist nur zwei, manchmal auch nur eine, Aufnahme vorliegen, wird zusätzlich eine ausgleichende Flächendarstellung mittels Kontourdiagrammen gewählt, welche die Raumstruktur besser sichtbar macht. Die Karten sollen entsprechend nicht auf die einzelnen Punkte hin gelesen werden, sondern auf die übergreifenden Flächenstrukturen.

## 4 Resultate

Im Folgenden werden die Daten für die Silbendauer sowie für Position von L und H dargestellt. Nach einer ersten statistischen Analyse in Bezug auf die traditionelle Dialekteinteilung des Deutschen, werden die Daten im Raum visualisiert. Vor der sprachgeographischen Interpretation werden die Daten auf ihre Konsistenz hin überprüft.

### 4.1 Silbendauer in traditioneller Dialekteinteilung

Abbildung 2 zeigt, dass die Silbendauern nach einer klassischen Aufteilung in Dialektregionen mit einer ANOVA signifikante Unterschiede aufweisen, $F(5, 322) = 4{,}83$, $p < 0{,}001$. Ein Post-hoc Students *t*-Test präzisiert, dass sich die beiden oberdeutschen Gebiete von den westnieder- und westmitteldeutschen hochsignifikant ($p < ,01$) unterscheiden, während die ostnieder- und ostmitteldeutschen Gebiete keine signifikanten Unterschiede zu den anderen Gebieten aufweisen. Hier unterscheiden sich die vorliegenden Resultate von [8], wo für das Ostmitteldeutsche längere Silbendauern als für norddeutsche und bairische Aufnahmen belegt sind.

Aus Abbildung 3 wird deutlich, dass eine Zunahme der Silbendauer von [vˈan] im Süden, wie sie in Abbildung 2 gegeben ist, v. a. auf einen längeren Nucleus zurückzuführen ist. ANOVAs, die die drei Silbenteile in Bezug zur Dialektregion setzen, weisen für den Onset und die Coda bei Weitem nicht signifikante Unterschiede nach, während für den Nucleus ein höchstsignifikanter Unterschied vorliegt, $F(5, 322) = 25{,}59$, $p < 0{,}001$. Die längeren Silbendauern im Süden sind also im Wesentlichen auf längere Vokale zurückzuführen.

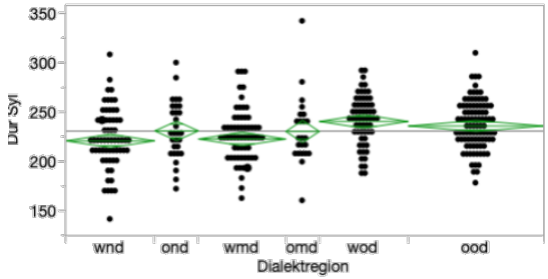



**Abbildung 2** – Silbendauer in ms (Mittelwert und Konfidenzintervall in der Diamantdarstellung) nach Dialektregion (w = West-, o=Ost-, n =nieder-, m=mittel-, o=ober-, d=-deutsch)

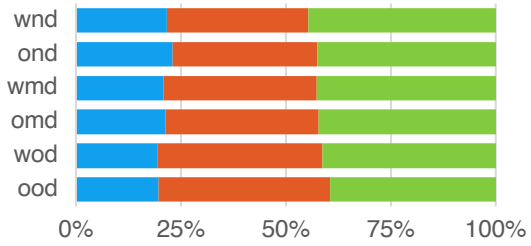
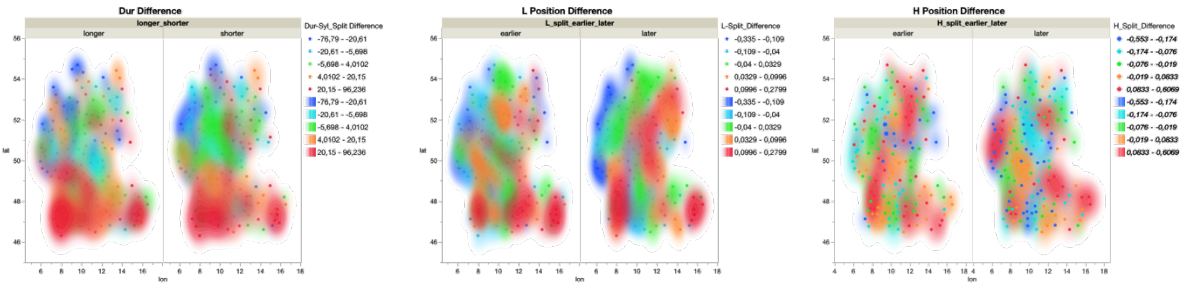



**Abbildung 3 –** Anteil von Onset (blau), Nucleus (orange) und Coda (grün) an der Silbe nach Dialektregion (w= West-, o=Ost-, n=nieder-, m=mittel-, o=ober-, d=-deutsch)

Hahn [5] hat nachgewiesen, dass die klassischen Dialekteinteilungen, die hauptsächlich auf laut-segmentalen, teilw. auf morphologischen Variablen beruhen, für prosodische Aspekte keine relevante Bezugsgröße sind, sondern dass prosodische Variablen – insb. temporale Aspekte, Reduktionen und Elisionen – eigenständige Raumstrukturen bilden. Deshalb wird folgend, auch wenn sich wie oben Unterschiede feststellen lassen, auf den Bezug zur traditionellen Dialektklassifizierung verzichtet. Stattdessen werden die Daten direkt im Raum dargestellt.

## 4.2 Datenkonsistenz

Da insgesamt, wie oben dargestellt, pro Ort meist nur zwei Aufnahmen vorliegen (manchmal auch nur eine) und die Prosodie einer großen individuellen Variation unterliegt, soll zuerst die Vermutung zurückgewiesen werden, dass die Werte zufällig im Raum verteilt sind. Wenn das gelingt, kann davon ausgegangen werden, dass die Daten sprachgeographische Muster abbilden. Dazu werden die Werte aller Orte, für die zwei oder mehr Datensätze vorliegen, im Sinne einer Split-half-Methode in zwei gleich große Gruppen aufgeteilt: Eine Gruppe enthält für die untersuchten Variable die höheren Werte, die andere Gruppe die niederen Werte. Orte mit nur einem Beleg werden aus dieser Analyse ausgeschlossen, ebenso mittlere Belege von Orten mit mehr als zwei Aufnahmen. Für jede dieser beiden Gruppen wird der jeweilige Mittelwert errechnet und die individuellen Abweichungen von diesem Mittelwert. Als Erstes wurde die Korrelation zwischen den Werten der beiden Gruppen überprüft. Für alle drei Variablen finden sich gute Korrelationswerte (Silbendauer, r = 0,59, p < 0,001; Position von L, r = 0,5, p < .0,001, Position von H, r = 0,51, p < 0,001), so dass der Zusammenhang gegeben ist. Anschließend werden diese Abweichungen vom jeweiligen Mittelwert in den Raum projiziert.




**Karten 1–3 –** Half-split Abweichungen vom jeweiligen Mittelwert für Silbendauer (links), Position von L (Mitte) und Position von H (rechts)

Karten 1 bis 3 zeigen jeweils auf der linken Seite die höheren Werte für die Silbendauer bzw. für die früheren L-oder H-Werte und auf der rechten die jeweils tieferen. Für die Interpretation sind nicht lokale Übereinstimmungen oder Differenzen wichtig, sondern der Gesamteindruck,

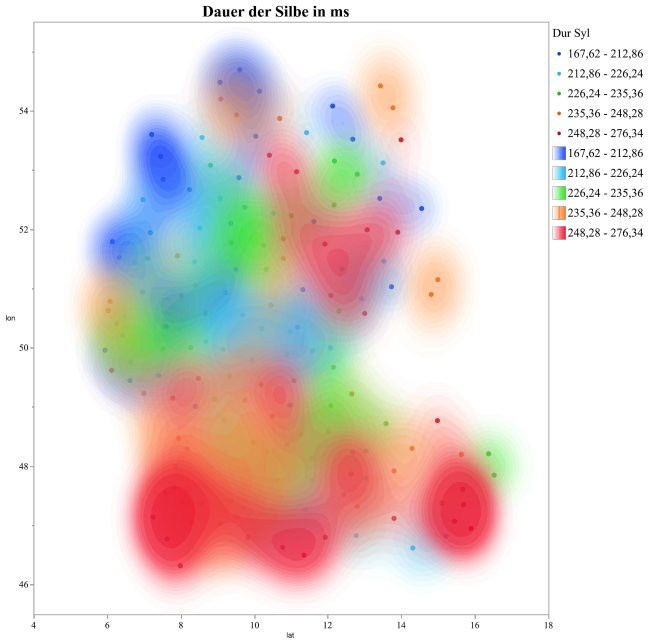
weshalb die kleine Abbildung ausreichend ist. Besonders für die Silbendauern (Karte 1) ergibt sich ein grundsätzlich konstantes Bild, Unterschiede finden sich im zentralen Bereich, wo für die Gruppe der Sprecher mit den kürzeren Dauern eine relativ stabile mittlere Dauer zu sehen ist, die in der anderen Gruppe variabler erscheint. In Bezug auf die Position von L (Karte 2) ist der Westen und Norden konstant mit Werten, die ein frühes L belegen. Im Süden fallen drei Regionen mit späteren Werten auf, die durch Regionen mit weniger späten unterbrochen sind. Hier liegt also keine Einheitlichkeit vor, aber eine etwas kleinräumigere Konstanz. Die Gebiete im Nordosten und im mittleren Osten sind im Vergleich der beiden Gruppen nicht ganz so stabil. Die Verteilung für die Position von H (Karte 3) ist weniger einheitlich. Insbesondere im Westen und Norden sind die Unterschiede deutlich, im Süden sind sie eher gradueller Natur.

Auch wenn die beiden Gruppen nicht identische sprachgeographische Verbreitungsmuster aufweisen, deuten die in Grundzügen ähnlichen Karten doch darauf, dass für die Silbendauer und die Position von L von einer sprachgeographischen Struktur ausgegangen werden kann. Auf die Besonderheiten für H wird weiter unten (4.5) eingegangen.

Für die folgenden Karten werden die Daten aller Aufnahmen pro Ort gemittelt, was nach den Einschätzungen der Half-split Analysen von Karte 1 und Karte 2 gerechtfertigt erscheint. Durch die Zusammenlegung der Werte pro Ort werden zudem individuelle Züge ausgeglättet. Ebenfalls können so Orte, für die nur ein Beleg vorliegen, mitberücksichtigt werden.

## 4.3 Silbendauer im Raum

Karte 4 zeigt die gemittelten Silbendauern pro Ort im Raum. Sie verdeutlicht die schon in Karte 1 erkennbaren Raumstrukturen und deren Stütze durch die Abbildung 2. Im Süden finden wir die längeren Silbendauern, wobei v. a. im Osten auch Gebiete mit mittleren Werten erscheinen. Der westmittel- und der westniederdeutsche Raum zeigen kurze Silbendauern, Abweichungen mit etwas längeren Silben finden sich im moselfränkischen Raum sowie im Ostfälischen mit dem Übergang zum Brandenburgischen. Im ostmittel- und ostniederdeutschen Raum führt die Zusammenfassung der Daten pro Ort nur beschränkt zu einem konsistenteren Bild. Im Osten und Westen dieses Gebiets mit einer Brücke durch Sachsen finden sich eher längere Silben, dazwischen liegen Gebiete mit kürzeren Silben. Die Daten aus dieser Region deuten auf eine relativ starke Variation hin, wobei trotzdem jeder Ortspunkt Nachbarn mit ähnlichen Dauerwerten hat, die Verteilung also auch da nicht ganz zufällig ist.

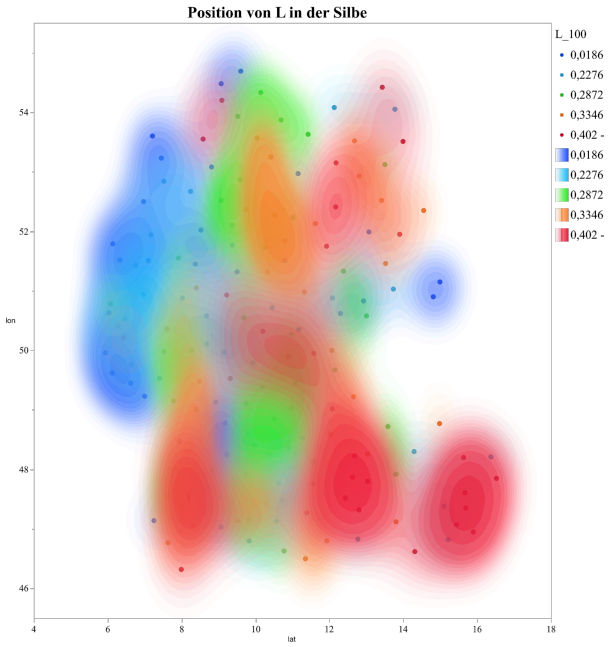



**Karte 4** – Silbendauer in ms (dunkelblau = kurz, hellblau, grün, orange, rot = lang) im Raum

Mit Fokus auf das Ostmitteldeutsche widerspricht auch diese Karte den Daten von [8], die für das Ostmitteldeutsche längere Silben als im Süden feststellen. Jedoch deckt sich das Bild mit der Verteilung in Hahn [5]. Das ist nicht unerwartet, da die Daten derselben Erhebung entstammen. Allerdings berücksichtigt [5] die gesamte Aufnahmedauer, während hier nur eine einzelne Silbe gemessen wird.

## 4.4 Position von L im Raum

Neu ist die Untersuchung intonatorischer Aspekte. Karte 5 zeigt die Position von L als Abstand in ms vom Silbenbeginn, wobei die frühesten Werte kurz nach Silbenbeginn liegen und die spätesten nach der Mitte der Silbe. Insgesamt bestätigt und klärt die Zusammenfassung der Werte pro Ort das Raumbild der Half-split-Analyse. Die Zusammenfassung individueller Werte pro Ort ergibt ein deutliches Raumbild eines kontinuierlichen West-Ost-Übergangs mit frühen Werten im Westen und späten im Osten. Durchbrochen wird dieses Muster im Südwesten durch späte Werte im Alemannischen und Rheinfränkischen. Das alemannische und bairische Gebiet mit den späten Werten ist nicht als durchgehendes südliches Gebiet zu interpretieren, da dazwischen im Schwäbischen sowie an der südlichsten bairisch-alemannischen Grenze mittlere Werte vorkommen. Im Osten finden sich als Abweichung frühe Werte im obersächsischen Raum.



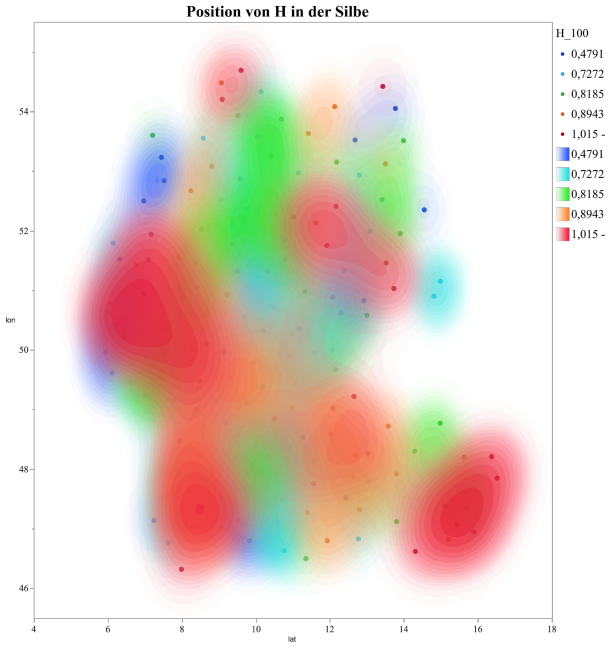

**Karte 5 –** Position von L in der Silbe als Abstand in ms vom Silbenbeginn (dunkelblau = früh, hellblau, grün, orange, rot = spät) im Raum

Diese räumliche Verteilung bestätigt die Daten in [1], wo für nordwestdeutsche Sprecher frühe und für bairische Sprecher späte Werte belegt sind. Für das Ostmitteldeutsche ist die Interpretation in Bezug auf [8] etwas komplexer. Die gegenüber dem Bairischen deutlich früheren Werte des Ostmitteldeutschen können bestätigt werden. Doch dass die Werte im Ostmitteldeutschen auch im Vergleich zum Niederdeutschen frühere Position kennzeichnen, wird eher nicht gestützt: Je nach regionaler Herkunft der Sprecher aus dem westniederdeutschen Raum, der blaue und grüne Flächen und auch wenige rote und orange Ortspunkte umfasst, könnten die ostmitteldeutschen blauen Punkte, welche im Wesentlichen den obersächsischen Raum umfassen, etwas frühere L markieren. Wenn man allerdings der Mehrzahl der Belege dieser Karte berücksichtigt, so ist die sehr frühe Position der ostmitteldeutschen Sprecher von [8] eher zu hinterfragen. Generell wird deutlich, dass die grobmaschige Einteilung nach Dialektgebieten von [1] und [8] der prosodischen Variation im Raum nicht gerecht werden kann.

## 4.5 Position von H im Raum

Während die Daten für die Silbendauer und für die Position von L auf der Basis der Half-split-Analysen (Karten 1 und 2) gut pro Ort zusammenzufassen sind und so aussagekräftigere Strukturen ergeben, ist die Analyse von H mit großer Vorsicht zu bewerten. Einerseits ist die Feststellung des Messpunktes durch den sich oft kontinuierlich abschwächenden Anstieg relativ willkürlich gesetzt und andererseits durch mögliche Messfehler sowie den konsonantischen Einfluss auf F0 beeinflusst. Zudem mussten mehrere Datensätze ausgeschlossen werden, weil nach dem L kein deutliches H als Teil eines Akzents realisiert wird, sondern nur ein Grenzton, welcher meist auf der letzten Silbe liegt, aber manchmal auch früher sein F0-Maximum erreicht.

Die Unterscheidung des hohen Akzenttons und des Grenztons ist auch für die menschliche Kodierung eine Herausforderung; eine automatische Unterscheidung, wie sie hier mit der Bedingung, dass H nicht später als das Ende der nächsten Silbe /ər/ sein kann, vorgenommen wurde, ist notwendigerweise fehlerbehaftet (aber wenigstens systematisch). Der Blick auf die Karte 3 hat die unterschiedlichen Raumstrukturen für die beiden Gruppen deutlich gemacht, was sich nun auch im Vergleich mit Karte 6 wieder bemerkbar macht. Die Bildung der Mittelwerte pro Ort glättet die individuellen Unterschiede aus. Dadurch dass Kartenbilder in 3 sich unterscheiden erscheint hier in Karte 6 ein weiteres, das durch die Einzelbelege nicht gestützt ist. Auf eine Interpretation dieser Daten wird deshalb verzichtet. Im Rahmen des Pilotprojekts wird aber deutlich, dass die verwendeten automatisierten Verfahren an ihre Grenzen stoßen und weitere Zugänge, evtl. doch auch wieder eine manuelle Annotation, nötig werden.




**Karte 6 –** Position von H in der Silbe als Abstand in ms vom Silbenbeginn (dunkelblau = früh, hellblau, grün, orange, rot = spät) im Raum

## 5 Diskussion und Conclusio

Die erste Forschungsfrage konnte deutlich bejaht werden, die Raumstruktur der Silbendauer deckt sich mit globaleren Maßen aus [5]. Auch die zweite Frage nach der zeitlichen Binnengliederung der Silbe kann in der Kürze bejaht werden. Die räumliche Strukturierung ist hier nur im Bezug auf die Dialekteinteilung gegeben, die sich für prosodische Untersuchungen als zu grob erweist. Für detailliertere Ergebnisse sind weitere Arbeiten notwendig. Die dritte Frage nach der Sprachgeographie der segmentalen Verankerung von L und H, die im Zentrum des Interesses steht, kann für die hier vorgenommene Untersuchung teilweise positiv beantwortet werden. Für L ergeben sich relativ stabile sprachgeographische Muster, die sich nur beschränkt an der traditionellen Dialekteinteilung orientieren. Sie bestätigen damit die grundsätzliche Einschätzung zur Sprachgeographie temporaler Strukturen [5]. Die großräumigere Abdeckung stellt andererseits die punktuelle Analyse von [8] in Bezug auf das Ostmitteldeutsche in Frage. Weitere Analysen aus dem bestehenden Korpus können hierzu Klärung bringen. Die Ergebnisse für die Verankerung von H sind dagegen mit den vorliegenden Analysen nicht befriedigend, was auf methodische Schwierigkeiten zurückzuführen ist. Damit ist auch schon Forschungsfrage vier angesprochen. Ein weitgehend automatisiertes Verfahren für die Intonation auf der Basis einer bestehenden Segmentierung auf lautlicher Ebene hat sich für die Analyse von L als plausibel erwiesen. Basis dafür ist eine intonationsphonologisch einheitliche Realisierung, was für L mit der hier untersuchten Intonationsphrase gegeben ist. Für die Analyse von H haben sich dagegen mehrere technische Probleme ergeben: teilw. nur schwache Ausprägung der Position von H, womit kleine Abweichungen bedeutsam werden und auch kleine Messfehler die Resultate beeinflussen können. Zudem realisieren die Sprecher die funktionale Weiterweisung intonationsphonologisch unterschiedlich, womit der Analyse verschiedene phonologische Muster zugrunde liegen, die vor der Analyse auseinandergehalten werden müssen. Hier müssen weitere Verfahren wie z. B. in [13, 15] entwickelt, beigezogen werden. Grundsätzlich

zeigt die Pilotstudie das Potenzial des Ansatzes. Dieser kann weiterverfolgt und verfeinert werden sowie auf weitere Akzentmuster, wie z. B. [3] auf fallende Nukleussilben, ausgedehnt werden. Die Daten bieten eine adäquate Grundlage dafür.

## Dank

Wesentliche Vorarbeiten entstammen dem Projekt „Sprechtempo und Reduktion im Deutschen“ (SpuRD), das von der Deutsche Forschungsgemeinschaft unter der Projektnummer 383554615 gefördert wurde.

Ein besonderer Dank geht auch an die Kolleg:innen an der P&P, insb. an Stefanie Jannedy und an Simon Oppermann, die mir mit kritischen Nachbohren geholfen haben, meinen Beitrag zu verbessern.

## Literatur

[1] ATTERER, M. und D. R. LADD. 2004. On the phonetics and phonology of "segmental anchoring" of F0: evidence from German. *Journal of Phonetics* 32. 177–197.
[2] BOERSMAA, P. & D. WEENINK. 2024. *Praat: doing phonetics by computer* [Computer program]. Version 6.4.10. online. (http://www.praat.org (29.10.2024))
[3] GILLES, P. 2005. *Regionale Prosodie im Deutschen. Variabilität der Intonation von Abschluss und Weiterweisung*. Berlin, New York: de Gruyter.
[4] GILLES, P. 2015. Variation der Intonation im luxemburgisch-moselfränkischen Grenzgebiet. In M. ELMENTALER, M. HUNDT, und J. E. SCHMIDT (Hg.), *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder*, 135–150. Stuttgart: Steiner.
[5] HAHN, M. 2022. *Sprechgeschwindigkeit und Reduktion im deutschen Sprachraum*. Hildesheim, Zürich, New York: Georg Olms.
[6] HAHN, M. und B. SIEBENHAAR. 2019. Spatial Variation of Articulation Rate and Phonetic Reduction in Standard-Intended German. In S. CALHOUN, P. ESCUDERO, M. TABAIN, und P.WARREN (Hg.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 2695–2699. Melbourne.
[7] KISLER, T., U. D. REICHEL, und F. SCHIEL. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347.
[8] KLEBER, F. und T. RATHCKE. 2008. More on the “segmental anchoring” of prenuclear rises: Evidence from East Middle German. In *Speech Prosody 2008*, 123–126. Aix en Provence. (http://aune.lpl.univ-aix.fr/~sprosig/sp2008/papers/id123.pdf)
[9] KLEINER, S. 2015. „Deutsch heute“ und der Atlas zur Aussprache des deutschen Gebrauchsstandards. In R. KEHREIN, A. LAMELI und S. RABANUS (Hg.), *Regionale Variation des Deutschen – Projekte und Perspektiven*, 489–518. Berlin/Boston: De Gruyter.
[10] KÜGLER, F. 2007. The intonational phonology of Swabian and Upper Saxon. Tübingen: Max Niemeyer.
[11] LEEMANN, A. 2012. *Swiss German Intonation Patterns*. Amsterdam: Benjamins.
[12] LEEMANN, A. 2017. Analyzing geospatial variation in articulation rate using crowdsourced speech data. *Journal of Linguistic Geography* 4(2). 76–96. (doi:10.1017/jlg.2016.11)
[13] MERTENS, P. 2022. The Prosogram Model for Pitch Stylization and Its Applications in Intonation Transcription. In J. BARNES und S. SHATTUCK-HUFNAGEL (Hg.), *Prosodic Theory and Practice*, 259–286. Cambridge (MA): MIT Press. (https://doi.org/10.7551/mitpress/10413.003.0010)
[14] PETERS, J. 2006. *Intonation deutscher Regionalsprachen*. Berlin, New York: De Gruyter. (https://doi.org/10.1515/9783110201871)
[15] Pistor, T. 2022. *Universelle Intonationsmuster. Ein empirischer Nachweis konstanter prosodischer Strukturen in Regionalsprachen des Deutschen und darüber hinaus*. Stuttgart: Steiner.

# TOR GLEICH TOR? UNTERSUCHUNG ZUR WAHRNEHMUNG DER EXPRESSIVITÄT VON FUßBALL-LIVEKOMMENTAREN

*Philipp Simon, Ines Bose, Sven Grawunder*

*Martin-Luther-Universität, Halle/Saale*
*philipp.simon@student.uni-halle.de, {bose, grawunder}@sprechwiss.uni-halle.de*

**Kurzfassung:** Der vorliegende Beitrag untersucht, ob die häufig in den Social-Media geteilte Ansicht - der Frauenfußball wäre unattraktiver - sich auch in der Sprechweise der Live-Kommentatoren und -Kommentatorinnen der ARD/ZDF wiederfindet. Dazu wurden zunächst 30 Audiodateien von Torkommentierungen von der Männer-Fußball-WM 22 und der Frauen-Fußball-WM 23 ausgewählt und von Probanden hinsichtlich der Expressivität der Kommentatoren/innen bewertet. Die Ergebnisse zeigen, dass die Männer-WM expressiver wahrgenommen wird. Im Anschluss wurden diese 30 Audios phonetisch analysiert und die Ergebnisse mit dem Resultat der Probandenbefragung korreliert, um die Muster hinter der Entscheidung der Probanden zu finden. Die Ergebnisse zeigen, dass die Männer-WM begründet als expressiver wahrgenommen wird. Dies zeigt sich in höheren Werten der Männer-WM bei der Differenz zwischen mittlerer $F_0$ in konsekutiven Spannungsphasen (SP3 vs. SP5)[5], beim Stimmumfang in der gesamten Audiodatei und beim $F_0$-Mittelwert der SP5. Dies alles sind Kovariaten, die in der Literatur mit expressiver Sprechweise [6] verknüpft sind.

## 1 Hintergrund

In vielen Sportforen oder auf Social-Media gibt es die Auffassung, dass der Frauenfußball langsamer und weniger attraktiv ist als der Männerfußball. Der Vergleich der Zuschauerzahlen der letzten beiden Weltmeisterschaften zeigt ein deutlich höheres Interesses der Fans an der Männer-WM (ca. 5 Mrd.) als an der Frauen-WM (ca. 1,1 Mrd.) [1, 2]. Neben sportlichen Aspekten, die mitunter auch diskutiert werden, lassen zudem auch auditive Höreindrücke vermuten, dass die Livekommentierung im Fernsehen beim Männerfußball eine andere ist als beim Frauenfußball.

Obwohl sich schon viele linguistische Untersuchungen im Kontext der Performance [14] mit dem Thema Fußball und Fußballübertragungen beschäftigten, sodass es sich eine regelrechte Fußballlinguistik [13] herausgebildet hat und die Zahl der Corpora mit dem Fokus Sportkommentierungen insbes. Fußball ansteigt [15], gibt es nur wenige Arbeiten zu den sprecherischen Realisierungen solcher Formate. Kern [3] untersucht beispielsweise die von den Kommentatoren eingesetzten Mittel zur Steigerung der Spannung und beim Torerfolg. Trouvain [4] analysiert die Unterschiede bei stimmlichen Parametern zwischen Radio- und Fernsehübertragung. Eib/Bose [5] analysierten die unterschiedlich eingesetzten sprecherischen Mittel von Radioreportern in verschiedenen Radioübertragungsformaten eines einzelnen Fußballspiels miteinander.

Der vorliegende Beitrag ist im Kontext einer Masterabschlussarbeit (des Erstautors) entstanden, die sich zum Ziel gestellt hatte, einen direkten Vergleich zwischen der sprecherischen Gestaltung der Kommentatoren/innen im Männer- und Frauenfußball anzustellen. Hierbei wurde die Frage gestellt, ob „Tor gleich Tor“ gilt. D.h.: Wird in der ARD/ZDF-Fernsehübertragung die Torerzielung bei der Männer-WM expressiver kommentiert als bei der Frauen-WM? Zur Beantwortung werden dafür Merkmale und Parameter der Produktion und Perzeption bzw. des Eindrucks herangezogen.

## 2 Expressivität

Expressivität wird hier aufgefasst als das Ausmaß der Emotionalität einer Person, das sich im Einsatz von sprecherischen Mitteln und Sprache manifestiert [6]. Diese Mittel dienen dazu, eine bestimmte Wirkung zu erreichen, z. B.:
die eigene Emotionalität authentisch zu manifestieren und zu zeigen,
vom Publikum als emotional wahrgenommen zu werden (Selbstkundgabe). D.h.: als der leidenschaftliche Kommentator wahrgenommen zu werden.

- ein Tor als besonders spektakulär zu beschreiben und dadurch Emotionen beim Publikum zu wecken. D.h.: die Fans mit sprecherischen Mitteln an der Atmosphäre im Stadion teilhaben zu lassen.

Die Übergänge zwischen den drei Aspekten der Expressivität sind dabei fließend und hängen wiederum stark mit der Persönlichkeit, dem Selbstverständnis und der Rolle des Kommentators /der Kommentatorin zusammen [7]. Die gewählten sprecherischen Mittel zum Ausdruck der Expressivität bei der Livefußballkommentierung von Toren sind primär die Stimmhöhe [8], aber auch Sprechgeschwindigkeit, Endmelodieverlauf, Lautstärke und Intensität [3,4,5] werden in der wissenschaftlichen Literatur diskutiert.

## 3 Übersicht Spannungsphasen

Eib/Bose [5] beschreiben die verschiedenen Spielphasen und den dazugehörigen Sprechausdruck. Dabei gibt das Fußballspiel i. d. R. vor, wie die Stimme eingesetzt wird. Das Geschehen oder das von dem/der Kommentator/in antizipierte Geschehen wird der Situation entsprechend mit einer charakteristischen Sprechweise begleitet. Die Sprechphasen sind dabei nicht immer deckungsgleich mit der Spannungsphase. Obwohl die Untersuchung im Zusammenhang von Radioreportern erstellt wurde, lässt sie sich auch auf die Fernsehkommentator/innen übertragen. Die hierzu von Eib/Bose [5] entwickelten Spannungsphasen (SP) dienen als systematische und möglichst eindeutige Einteilung der verschiedenen Phasen im Spiel. Insgesamt gibt es bei Eib/Bose [5] fünf verschiedene Spannungsphasen, die in der Regel aufeinander aufbauen. Die für die vorliegende Untersuchung bedeutsamen Spannungsphasen sind SP3 bis SP5, also die entscheidenden Phasen für die Entwicklung eines Tores.

**Tabelle 1 – 5 Spannungsphasen eines Fußballspiels aus der Perspektive eines Livereporters [5].**

| Spannungsphase | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Kurzbeschreibung | Ball aus dem Spiel | Ruhiges Spiel (im Mittelfeld) | Ballbewegung Richtung Tor | Torchance/ Torschuss | Torerzielung |
| Ball im Spiel | nein | ja | ja | ja | ja |
| Ball im Angriffsdrittel | nein | nein | ja | ja | ja |
| Torchance zu erwarten | nein | nein | nein | ja | ja |
| Torerzielung | nein | nein | nein | nein | ja |

## 4 Korpus

Das Untersuchungskorpus wurde aus den in der ARD/ZDF-Mediathek verfügbaren Spielen der Männer-WM 2022 in Katar und der Frauen-WM 2023 in Australien und Neuseeland zusammengestellt. Dabei wurde direkt nach den Turnieren sowohl die Video- als auch die Audiodatei gesichert. Insgesamt umfasst das Korpus 194 der 336 erzielten Tore beider Turniere, davon 71 Tore der Männer-WM und 123 Tore der Frauen-WM. Da nicht alle Tore untersucht werden konnten, wurden Auswahlkriterien für die Tore aufgestellt. Zum einen wurden nur Tore von Kommentatoren und Kommentatorinnen verwendet, die sowohl bei der Männer-WM als auch bei der Frauen-WM im Einsatz waren. Dies waren zwei Kommentatorinnen: Christina Graf (CG) und Claudia Neumann (CN) und zwei Kommentatoren: Martin Schneider (MS) und Oliver Schmidt (OS). Außerdem wurden nur Tore aus dem Spiel heraus untersucht (Spannungs-

phasen SP3 bis SP5). D.h. keine Standardsituationen (Freistoß, Eckball, Elfmeter), kein mögliches Abseits, keine Verletzung, kein Eigentor und keine deutsche Beteiligung. Übrig blieben 63 Tore. Um möglichst vergleichbare Tore pro Kommentator/in zu verwenden, wurden außerdem nur Tore der Gruppenphase verwendet und solche mit einem gleichen oder ähnlichen Ergebnis (1:0, 1:1 etc.). Am Ende des Auswahlprozesses standen insgesamt 30 Tore. 15 pro Turnier und zwei bis fünf Audios pro Kommentator/in. Die Tonspur der Audiodatei startet ca. 10 Sekunden vor dem Tor und endet ca. 15 Sekunden nach dem Tor. Der Start war dabei entweder eine Sprechpause oder, falls der Start in eine Kommentierungsphase fiel, wurde er so nach vorne verlegt, dass der angefangene Satz zumindest als Teiläußerung mit aufgezeichnet wurde und somit verständlich wird. Nach dem Tor wurde so lange aufgezeichnet, bis die Expressivität in der Stimme der Kommentator/innen abgeklungen war. Dies markiert das Ende der SP5.

## 5 Hörerbefragung

Um den subjektiven Eindruck von Social-Media – die Männer-WM wäre attraktiver – zu überprüfen, wurde eine Onlinebefragung mittels Limesurvey durchgeführt, in der die Probanden die 30 Audiobeispiele hinsichtlich der Expressivität des/der Kommentators/in bewerten sollten. Die Probanden mussten nach dem Hören der Audiodatei die Kommentierung auf einer Skala von 1 nicht expressiv bis 10 sehr expressiv bewerten. An der Umfrage nahmen 69 Personen teil. Neben dem Alter und Geschlecht wurde außerdem noch der Fußballkonsum abgefragt (18-65 Jahre; weiblich=23, männlich=43, divers=1, k. A.=2; Fußballkonsum mehrmals wöchentlich n=21, wöchentlich n=11, gelegentlich n=13, selten=14, gar nicht =10).

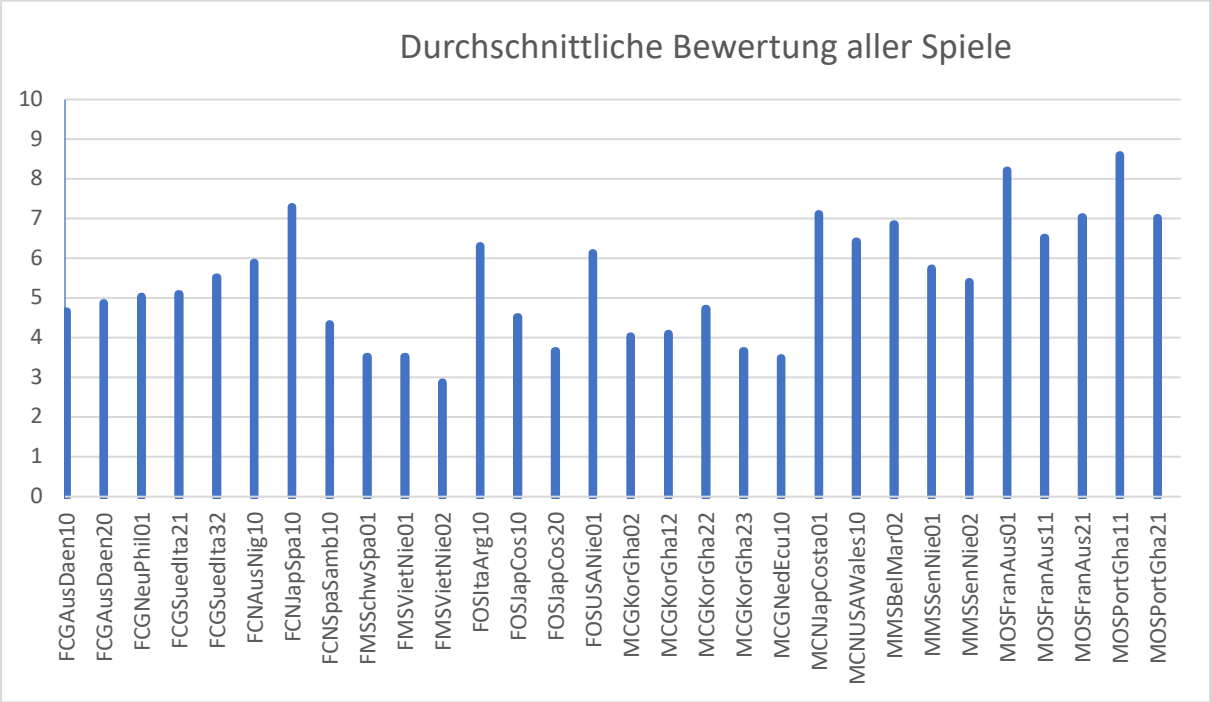



**Abbildung 1 – Durchschnittliche Bewertung aller Spiele. Die Bezeichnung der Spiele setzt sich zusammen aus: F/ M steht für Frauen- bzw. Männer-WM, gefolgt von dem/der Kommenator/in, dem Spiel und dem jeweiligen Spielstand.**

Die Abbildung 2 zeigt die durchschnittliche Expressivitätsbewertung aller 30 Spiele von allen 69 Teilnehmer/innen. Die Kommentierungen mit der höchsten Expressivitätswahrnehmung sind die Spiele MOSPortGha11 (8,60), MOSFranAus01 (8,21) und FCNJapSpa10 (7,28). Die Spiele, die am wenigsten expressiv wahrgenommen werden, sind FMSVietNie02 (2,88), MCGNedEcu10 (3,49), FMSVietNie01 (3,52) und FMSSchwSpa01 (3,52).

**Tabelle 2 – Übersicht über die Ergebnisse der Hörerbefragung**

| | Frauen-WM | | | Männer-WM | | |
|---|---|---|---|---|---|---|
| Hörer-gruppen | Gesamt | w / m | Selten, kei-nen /regel-mäßig | Gesamt | w / m | Selten, keinen /regelmäßig |
| Gesamt | 4,89 | 5,16 / 4,68 | 5,49 / 4,57 | 5,93 | 6,04 / 5,82 | 6,37 / 5,70 |
| CG (weiblich) | 5,04 | 5,39 / 4,80 | 5,60 / 4,71 | 4,01 | 4,17 / 3,87 | 4,61 / 3,70 |
| CN (weiblich) | 5,84 | 5,99 / 5,69 | 6,43 / 5,46 | 6,77 | 7,09 / 6,48 | 7,25 / 6,44 |
| MS (männlich) | 3,31 | 3,54 / 3,10 | 3,92 / 3,05 | 6,00 | 6,09 / 5,97 | 6,41 / 5,85 |
| OS (männlich) | 5,16 | 5,46 / 4,96 | 5,84 / 4,85 | 7,48 | 7,45 / 7,41 | 7,76 / 7,31 |
| | | | | | | |
| Geringste Expressivität | | | | | | Höchste Expressivität |

Die Ergebnisse zeigen, dass es eine geringe Differenz gibt im Vergleich der Geschlechter. D.h. weibliche Konsumenten nehmen die Kommentierungen ca. 0,2 bis 0,4 Punkte expressiver wahr als die männlichen Konsumenten. Außerdem nehmen Konsumenten, die selten bis keinen Fußball schauen, die Kommentierung 0,8 bis 0,9 Punkte expressiver wahr als solche die regelmäßig einschalten.
Des Weiteren ist interessant, dass das Alter bei der Bewertung keine Rolle spielt. Die Expressivitätswahrnehmung ist über alle Altersgruppen hinweg sehr stabil und entspricht der Feld-Gesamtheit.

## 6 Ergebnisse der phonetischen Analyse

Die 30 Audiobeispiele wurden mittels Audacity [11] und BAS Web Audio Enhancement Services [9] bearbeitetet und optimiert, um die Stadion- und Hintergrundgeräusche zu minimieren. Anschließend wurden die Dateien in Praat [12] transferiert, die Silbengrenzen gesetzt, der $F_0$-Wert pro Silbe eingetragen und die Spannungsphasen festgelegt. Pro Spannungsphase wurden folgende Parameter ermittelt und mit den Ergebnissen der Befragung korreliert: Sprechgeschwindigkeit (Silben/Sekunde), Stimmhöhe $F_0$-Wert (Mittelwert), Stimmumfang und -anstieg, außerdem die Differenz der $F_0$-Mittelwerte zwischen SP3 und SP5. Die stärksten Zusammenhänge finden sich in der Differenz von mittlerer $F_0$ in SP3 und SP5 (0,65), im Tonhöhenumfang der Kommentator/innen in der gesamten Audiodatei (0,65) und in der mittleren $F_0$ beim Torjubel (SP5) (0,58).
Die Ergebnisse (s. Tabelle 3) zeigen, dass die Männer-WM begründet als expressiver kommentiertwahrgenommen wird. Dies zeigt sich in höheren Werten der Männer-WM bei der Differenz zwischen mittlerer $F_0$ in SP3 vs. SP5, beim Stimmumfang in der gesamten Audiodatei und beim $F_0$-Mittelwert der SP5. Ebenfalls zeigt sich hier der Studioeffekt, da MS und OS insbesondere bei der Differenz $F_0$-Mittelwerte SP3 vs. SP5 und der $F_0$-Mittelwerte SP5 (Hz) deutlich höhere Werte bei der Männer-WM aufweisen. Ob dies allein die Differenz zwischen den beiden Turnieren erklären kann, lässt sich aus den vorliegenden Daten nicht abschließend bestimmen. Vermutlich spielen persönliche Vorlieben und die Spielauswahl ebenfalls eine Rolle.

**Tabelle 3 – Übersicht über die Parameter mit der höchsten Korrelation zu der Hörerbefragung.* Diese Parameter beruht nur auf 22 der 30 Audios da nicht bei allen Audios eine SP3 vorhanden war. **Halbtöne mit Referenzton 100Hz**

| | Differenz $F_0$-Mittelwerte SP3 vs. SP5 (Hz/ HT)* | | Tonhöhenumfang gesamte Audiodatei (Hz /HT) | | $F_0$-Mittelwerte SP5 (Hz / HT)** | |
|---|---|---|---|---|---|---|
| | Frauen-WM | Männer-WM | Frauen-WM | Männer-WM | Frauen-WM | Männer-WM |
| Gesamt | 115 / 9 | 173 / 13 | 269/19 | 288/20 | 311 / 19,6 | 342 / 21,2 |
| CG | 170 /11 | 116 / 8 | 292/16 | 240/15 | 362 / 22,2 | 314 / 19,8 |
| CN | 46 / 3 | 237 / 13 | 307/18 | 380/26 | 322 / 20,2 | 410 / 24,4 |
| MS | 107 / 11 | 157 / 12 | 226/21 | 235/17 | 245 / 15,5 | 324 / 20,3 |
| OS | 109 / 9 | 224 / 18 | 257/21 | 330/25 | 289 /18,3 | 352 / 21,7 |

Da u.a. die Stimmhöhe stark abhängig vom Geschlecht ist, wurde die Ergebnisse nochmal aufgeteilt in Kommentatoren (blau) und Kommentatorinnen (orange).

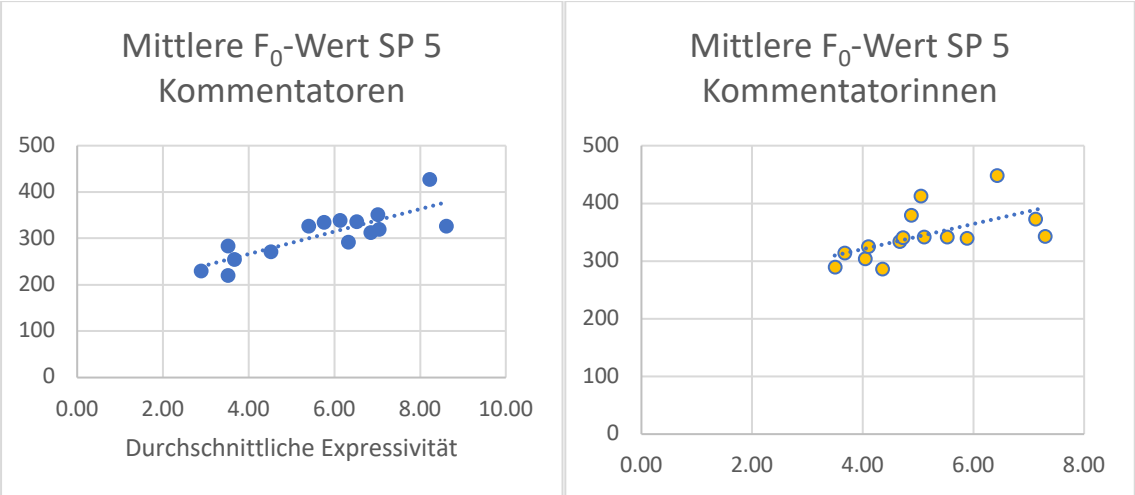



**Abbildung 2 – Mittlerer $F_0$-Wert während der Torkommentierung SP 5**

Die geschlechterspezifische Analyse zeigt, dass es bei den Kommentatoren (0,82) eine noch stärkere Korrelation zwischen der Stimmhöhe während der Torkommentierung und den Ergebnissen der Befragung gibt als bei den Kommentatorinnen (0,58). Außerdem fällt auf, dass die Kommentatoren beim Torjubel teilweise ähnliche mittlere $F_0$-Wert erreichen wie die Kommentatorinnen. Hier liegt der Verdacht nahe, dass die Kommentatorinnen die Stimme beim Torjubel kontrollieren. Auch Aussagen von Claudia Neumann deuten in die Richtung [10]. Das Ausmaß und die Auswirkungen davon wäre eine weitere spannende Forschung im Bereich der Livekommentierung.

## 7 Fazit und Diskussion

Das Hauptergebnis aus der Befragung der Teilnehmer bezüglich der Wahrnehmung der Expressivität der WM-Torkommentierung ist, dass ein Unterschied zwischen der Torkommentierung der Männer- und Frauen-WM wahrgenommen wird. Die Ergebnisse der Befragung deuten stark darauf hin, dass die Torkommentierung der Männer-WM expressiver ist als die der Frauen-WM. Insbesondere der individuelle Unterschied zwischen den Kommentatoren und beim

gleichen Kommentator zwischen den Turnieren ist teilweise sehr hoch. Während CG bei der Frauen-WM-Kommentierung expressiver (ca. 1 Punkt) wahrgenommen wird, empfinden die Umfrageteilnehmer/innen bei CN, MS und OS die Kommentierung der Männer-WM als expressiver (ca. 1 - 2,7 Punkte). Dabei lässt sich diese Expressivität auch durch die Verläufe der $F_0$-Werte in den Spielphasen belegen.

Unklar ist jedoch, welchen Anteil die Studiokommentierung an der Differenz der Wahrnehmung zwischen Männer- und Frauen-WM von MS und OS hat (2,3 und 2,7 Punkte weniger als bei der Männer-WM), da beide die Frauen-WM nicht live aus dem Stadion, sondern aus dem Studio kommentiert haben. CG und CN dagegen haben beide Turniere live aus dem Stadion kommentiert. Ohne entsprechende Vergleichswerte von CG und CN aus dem Studio oder MS und OS live aus einem Stadion der Frauen-WM lässt sich die Differenz nicht bestimmen.

Da für die Zuschauer/innen nicht ersichtlich ist, ob der/die Kommentator/in im Stadion oder in der Studiokabine sitzt (ausgenommen von einer Erwähnung durch die Kommentatoren selbst), aber der gleiche auditive Anspruch an eine Kommentierung gestellt wird, wäre es, für das gleichwertige Erlebnis der Expressivität bei beiden Turnieren, erforderlich, die Kommentator/inn/en die Spiele vor Ort kommentieren zu lassen.

Durch die Anzahl von 15 Vergleichsspielen pro Turnier und zwei bis fünf Audiobeispielen pro Kommentator sind die Ergebnisse nur ein erster Vergleich und müssten durch weitere Untersuchungen bestätigt werden. Ein nächster Schritt wäre, alle archivierten Spiele der Männer-WM 22 und der Frauen-WM 23 zu analysieren und diese mit den Bewertungen aus dieser Arbeit zu vergleichen.

## 8 Literatur


[1] https://www.euromonitor.com/press/press-releases/july-20232/womens-world-cup-2023-viewership-to-cross-2-billion-double-from-2019-euromonitor-international (17.10.2024)

[2] https://inside.fifa.com/fifa-world-cup-qatar-2022-commercial (17.10.2024)

[3] KERN, F.: Speaking dramatically - The prosody of live radio commentary of football matches. In: Barth-Weingarten, Dagmar (Hrsg.): Prosody in interaction. John Benjamins Publishing Amsterdam, 217–237, 2010.

[4] TROUVAIN, J.: Between excitement and triumph – Live Football commentaries in Radio vs. TV. In: Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII) Hong Kong, 2022-2025, 2011.

[5] EIB, F., BOSE, I.: „dass der Hörer das Kopfkino einschalten kann“ – Sprachliche und sprecherische Gestaltung von Spannungsphasen und Formatbezug in Fußballreportagen. In: Bose, Ines / Finke, Clara Luise / Schwenke, Anna (Hrsg.): Medien – Sprechen – Klang: Empirische Forschungen zum medienvermittelten Sprechen. Frank&Timme: Berlin, 242-279. (SSP 27), 2021.

[6] PUSTKA, E.: Was ist Expressivität? In: Pustka, Elissa / Goldschmitt, Stefanie (Hrsg.): Emotionen, Expressivität, Emphase. Erich Schmidt Verlag Berlin, 11-39, 2014.

[7] KUNERT, J., KUNI, P.: Tension between Journalistic and Entertainment Values in Live Soccer TV Commentary: The Commentator’s Perspective. *Journalism and Media* 4: 631–647, 2023.

[8] SAMLOWSKI, B., KERN, F., TROUVAIN, J.: Perception of Suspense in Live Football Commentaries from German and British Perspectives In: 9th International Conference on Speech Prosody 2018, Poznań, 40-44, 2018.

[9] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/AudioEnhance (17.10.2024)

[10] https://www.blick.ch/sport/fussball/fussball-kommentatorin-claudia-neumann-ich-will-einfach-meinen-job-machen-id15817757.html (17.10.2024)

[11] Audacity Team: Audacity (Version 3.6.0) [Software], 2024. https://www.audacityteam.org/

[12] Boersma, P. und Weenink, D. Praat: doing phonetics by computer [Computer-Programm Version 6.4.20]. 2024. Verfügbar unter: http://www.praat.org/

[13] Meier-Vieracker, S., (Hrsg.): Reingegrätscht: eine kleine Linguistik des Fußballs. Tübingen: Narr Francke Attempto; 2024.

[14] Lavric, E., Pisek, G., Skinner A., Stadler W., (Hrsg.): The linguistics of football. Tübingen: Gunter Narr Verlag; 2008.

[15] Callies, M., Levin, M., (Hrsg.): Corpus approaches to the language of sports: texts, media, modalities. London, New York etc.: Bloomsbury Academic; 2021.

[16] Boczek, K., Dogruel, L., Schallhorn, C.: Gender byline bias in sports reporting: Examining the visibility and audience perception of male and female journalists in sports coverage. Journalism, 24(7), 1462–1481, 2023.

[17] Christopherson, N., Janning, M., McConnell, E.: Two kicks forward, one kick back: A content analysis of media discourses on the 1999 Women's World Cup Soccer Championship. Sociology of Sport Journal, 19(2), 170–188, 2002.

# GENDER ROLES SHAPING PHONETIC PATTERNS IN GEORGIAN AND GERMAN

*Nato Sulaberidze, Adrian P. Simpson, Melanie Weirich*

*Friedrich-Schiller-University, Jena*
*nato.sulaberidze@uni-jena.de*

**Abstract:** This analysis explores differences in stop production and fundamental frequency in relation to a continuous spectrum of self-reported gender role identity (masculinity / femininity) in German and Georgian speakers. The potential relationships between gender role identity and phonetic measures, such as mean fundamental frequency (f0) and voice onset time (VOT), are examined in the speech of male speakers of these languages. In Georgian, VOT is measured for ejectives and their voiceless aspirated congeners. In German, fortis stops are analysed, including contexts where these stops exhibit ejective-like behaviour due to coarticulation with glottalised vowels. Gender role identity is assessed using the Personal Attributes Questionnaire (M+, F+) and Traditional Masculinity Femininity (TMF) scales, which measure femininity / masculinity based on subjects' self-reported characteristics. Analysis reveals an unexpected, significant negative correlation between mean f0 and femininity on the TMF scale across both language groups. A significant positive correlation between VOT and TMF scores for Georgian stops is also observed, although this pattern is not consistent across all positions studied.

## 1 Introduction

An association between higher mean fundamental frequency and female or feminine speech has been observed cross-linguistically [1][2]. In addition, other phonetic features, such as vowel space, consonantal contrasts, segment durations, and speech rate, also display gender-specific patterns [3]. Variation in these speech patterns has been explained by both physiological differences [4] and social factors [5][6]. Socio-phonetic arguments commonly cited in the literature suggest that gender-specific phonetic variation differs across languages and social strata [7][8], assuming no major anatomical differences across these different language and social groups. Moreover, gender identity as a social construct has been shown to influence speech characteristics, regardless of the sex assigned at birth. Of particular relevance to this paper are the recent socio-phonetic studies on gender-specific correlates in speech in German, which showed a relationship between higher fundamental frequency (f0) and higher femininity (e.g. in heterosexual male speakers [2] and female lesbian speakers [9]). In terms of stop production, a positive correlation between feminine identity and the voice onset time (VOT) of /t/, as well as the aspiration of /k/ in gay male speakers of German [10] was observed.

As part of a broader study investigating the production of phonological (Georgian), epiphenomenal (German), and socio-phonetic (English) ejectives, this work focuses on the relationship between VOT of stops, f0, and gender-role identity in German and Georgian speakers. Georgian features ejective stops in its phonological inventory, which contrast with voiceless aspirated and voiced stops at bilabial, dental, and velar places of articulation. German, on the other hand, lacks ejectives phonologically, but these sounds may occur phonetically in specific contexts: final fortis stops are often produced as ejectives before glottalized onset open vowels [11].

The present analysis explores the relationship between gender role identity and phonetic parameters, including mean f0 and VOT, in German and Georgian male speakers. We analyse socio-phonetic patterns in these languages to explore how acoustic parameters correlate with gender-specific characteristics in these two linguistically and culturally distinct groups. To our knowledge, Georgian has not yet been subjected to a socio-phonetic study in terms of gender

identity. In addition, this analysis represents the first attempt to study the relationship between gender role identity and ejective production.

# 2 Method

## 2.1 Speakers

The speech and gender identities of nine Georgian (age: M = 31.33, SD = 11.4) and twelve German (age: M = 27.25, SD = 6.06) male participants, all assigned male at birth and self-identifying as male, were analysed.

Most of the Georgian participants (eight out of nine) were students or had already completed a bachelor's degree, and all were residing in Jena, Germany, at the time of recording. Eight of the nine reported speaking German as a second language at a minimum B1+ level. Of the Georgian group, two were born in Tbilisi (the capital, located in Eastern Georgia), six in various regions of Western Georgia, and one in Switzerland. However, five cited Tbilisi as their longest place of residence, and two reported spending half of their lives there.

Among the German participants, eleven held the general higher education entrance qualification, and nine were students in Jena. Nine of the twelve were born in East Germany, with eight reporting it as their longest place of residence.

## 2.2 Speech material and recordings

Voice onset time is measured for ejectives and their voiceless aspirated congeners in Georgian. In German, final fortis stops are examined. In the context of a following glottalised onset vowel such plosives are often produced as ejectives due to coarticulation. In the analysed reading material, stops were embedded in distinct sentence contexts due to the primary purposes of the study.

In Georgian voiceless aspirated and ejective stops in labial, dental and velar places of articulation / p t k p' t' k' / were embedded in sentences in vocalic contexts of nouns. Close and open vowels preceded and followed the plosive in both symmetric and asymmetric contexts: /aCa, iCi, aCi, iCa/. Sentence-initial stops had only followed vowels /i/ or /a/. Stops appeared sentence-initially (1), word-initially (2), and word-medially (3), e.g. in:

(1) პალატა ტიტების სურნელმა აავსო. / **p'a**lat'a t'it'ebis surnelma aavso / (The ward was filled with the scent of tulips)
(2) სუფთა პანელი დავაყენე. / supt**a p'a**neli davaq'ene / (I installed a clean panel)
(3) წვერის საპარსი ბასრია. / t͡s'veris s**ap'a**rsi basria / (The beard razor is sharp)

The original reading material of Georgian also contained target voiced plosives, which were not included in the VOT analysis to ensure comparability with the German data and to minimize additional complexity. There was a total of 324 target items in the original Georgian sentence material (3 places of articulation x 3 ways of articulation x 3 conditions x 4 nouns x 3 repetitions). The number of read sentences totalled 243 (81 sentences contained two target words).

In the German sentence material, fortis stops /p t k/ were embedded in verbs in three positions: before an open onset vowel (1), before a schwa (2), and before a schwa in nasal environment (3), e.g., in:

(1) Der Junge h**at a**uch gelacht. [deːɐ̯ jʊŋə h**at ʔa**ʊ̯x gəlaχt] (The boy also laughed)
(2) Das Mädchen h**atte** ein Buch dabei. [das mɛːtçən h**atə** ʔaɪ̯n buːx dabaɪ̯] (The girl had a book with her)
(3) Die Kinder h**atten** alle Spaß. [diː kɪndɐ h**atən** ʔalə ʃpaːs] (The children all had fun)

According to the primary study design, read material included a fourth condition in which the word-final fortis plosive, followed by a nasal (e.g., hat nie /hat niː/ 'had never') was produced. Since this condition is similar to the condition /haten/, which tends to be produced as a syllabic

nasal [hatn̪], it is not included in the analysis. Each sentence was elicited three times in randomised order, however speakers read the material in the consistent sequence. As a result, the total number of 216 sentences was read (3 stops x 4 conditions x 6 verbs x 3 repetitions).

The data was collected in the phonetic laboratory at the University of Jena by the same female experimenter (first author) for all subjects. For the original purpose of the study subjects were fitted with the electrodes of the electroglottograph as well as with bruxing plates with a pressure transducer attached [11].

### 2.3 Gender Identity measurement

Gender role identity is assessed in the SoSci Survey [12] using the following scales: TMF (The Traditional Masculinity-Femininity scale) [13] and German extended version of the PAQ (Personal Attributes Questionnaire) [14], GEPAQ (German Extended Personal Attributes Questionnaire) [15]. To assess gender role identity in Georgian men, we produced Georgian translations of TMF, GEPAQ-M+ and GEPAQ-F+ questionnaires.

The TMF questionnaire, consisting of six items, evaluates individuals' self-concept in relation to their gender roles (femininity / masculinity) on the levels of their own interests, behaviour and appearance. On a seven-point scale, respondents rated how they are traditionally perceived, how they perceive themselves and how they ideally wish to be. GEPAQ records the self-assessed femininity (expressivity), or masculinity (instrumentality) based on eight positive or negative characteristics that are stereotypically attributed to women or men. In our experiment we used questionnaires that included male and female positive traits (GEPAQ-M+ and GEPAQ-F+ respectively). Male positive traits consisted of independence, activity, superiority, competitiveness, decisiveness, persistence, self-confidence, and resilience, while female traits were emotionality, devotion, gentleness, kindness, empathy, understanding, and warmth. Ratings were made on a seven-point scale [2].

### 2.4 Preprocessing and analysis

For the acoustic analysis, audio files were transcribed using WebMAUS [16] and manually corrected and annotated using Praat [17]. VOT was measured between the points of stop oral release to the first pulse of glottal activity, associated with the following sound. In the German data, VOT of epiphenomenal ejectives was defined from the release of the oral closure to the clearly visible glottal pulse, associated with the glottal stop or the onset of the creaky voice of the following vowel. In Georgian ejectives VOT was measured from the oral release to the glottal activity associated with the following vowel. Mean VOT was calculated for each speaker, separated by stop and sentence context. Mean f0 was calculated for each speaker as well.

Statistical analysis was performed using R [18] with the visualizations of ggplot2 package [19]. Cronbach's Alpha was calculated to assess the internal consistency of three scales. For the GEPAQ-F+ scale, $\alpha = 0.75$, indicating acceptable internal consistency. The GEPAQ-M+ scale showed $\alpha = 0.79$, indicating good reliability, and the TMF scale yielded $\alpha = 0.89$, reflecting high internal consistency. Data normality was evaluated using the Shapiro-Wilk test, confirming that the variables were normally distributed ($p > 0.05$). As normality assumptions were met, Pearson correlation coefficients were computed to examine the relationships between VOT or f0 and the scales TMF, GEPAQ-F+, and GEPAQ-M+. To assess differences in mean GEPAQ-M, GEPAQ-F, and TMF scores between the Georgian and German groups, independent samples t-tests (Welch's variant) were conducted. Statistical significance was evaluated at an alpha level of .05.

# 3 Results

## 3.1 Gender identity

The analysis of gender identity using the TMF, GEPAQ-M+, and GEPAQ-F+ scales provide insights into how Georgian and German male speakers self-assessed their gender roles (masculinity and femininity) and how these assessments relate across different scale levels. Figure 2 illustrates the relationships among the scales, where higher mean TMF and GEPAQ-F+ scores reflect higher femininity, and higher mean GEPAQ-M+ scores reflect higher masculinity. Correlation tests did not reveal significant relationships between the scores, however certain noteworthy patterns emerged. Specifically, there appears to be a trend toward a positive relationship between the scales TMF and GEPAQ-F+, which both measure self-ascribed feminine gender role. In both Georgian and German speakers, higher TMF and GEPAQ-F+ scores (indicative of femininity) tend to be negatively correlated with traditional masculine traits measured by GEPAQ-M+ scores. In our sample, Georgian participants displayed a wider range of femininity scores on the TMF scale compared to Germans, with two Georgian subjects showing relatively high femininity scores. Conversely, some German speakers had high scores on the M+ scale, lying at the upper end of the range. Figure 2 shows differences between the scale scores of German and Georgian men. Independent samples t-tests revealed no statistically significant differences in mean GEPAQ-M+ scores, $t(17.78) = -0.67$, $p = .509$, mean GEPAQ-F+ scores, $t(16.61) = -0.36$, $p = .724$, or mean TMF scores, $t(11.80) = 0.71$, $p = .494$, between the Georgian and German language groups.

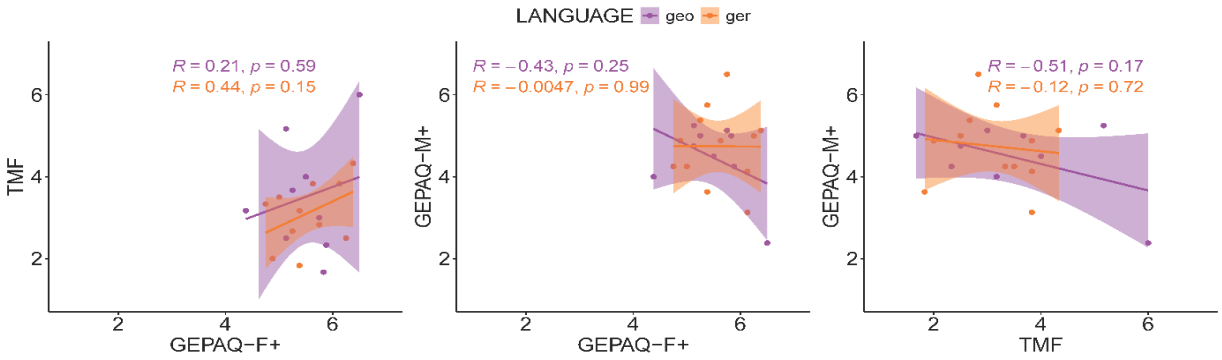



**Figure 1 -** Self-rated gender identity in German (orange) and Georgian (violet) speakers. Relationship between mean scores of GEPAQ-F+ and TMF (left), GEPAQ-F+ and GEPAQ-M+ (centre), and TMF and GEAPQ-M+. Higher TMF and GEPAQ-F+ scores indicate higher femininity, and higher GEPAQ-M+ scores indicate higher masculinity.

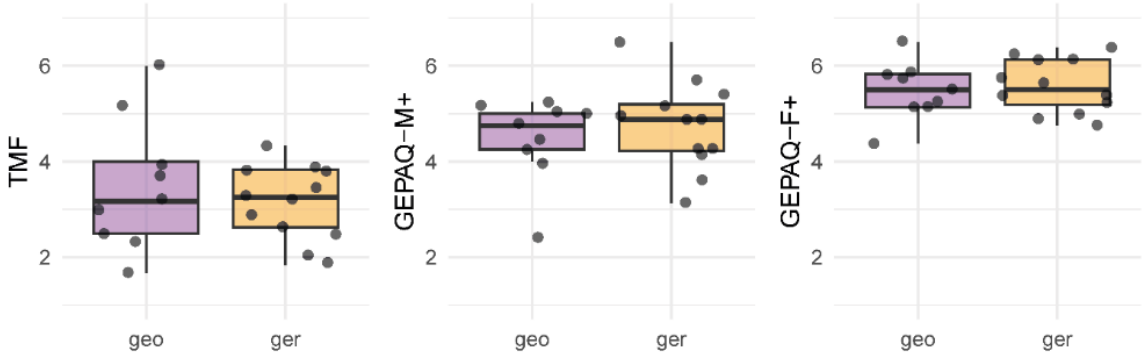



**Figure 2 -** Self-rated gender identity in German (yellow) and Georgian (violet) speakers. Mean scores of TMF (left), GEPAQ-M+ (centre) and GEPAQ-F+ (right), and TMF and GEAPQ-M+. Higher TMF and GEPAQ-F+ scores indicate higher femininity, and higher GEPAQ-M+ scores indicate higher masculinity.

### 3.2 F0

Figure 2 illustrates relationships between self-assigned gender identity (as measured by TMF and GEPAQ-F+ and GEPAQ-M+) and f0 in both Georgian and German speakers. In both Georgian and German, a significant negative correlation was found between mean f0 and TMF scores, indicating that speakers who rated themselves higher in femininity (based on the TMF scale) tended to have lower mean fundamental frequency.

No significant correlations were observed within the other scales. While the correlations between masculinity (GEPAQ-M+) and mean f0 were not statistically significant, a positive trend was observed in both groups. The GEPAQ-F+ scale showed different patterns in these two language groups: a positive trend between f0 and self-ascribed feminine traits was found in Georgian speakers, while a negative trend was noted in German speakers.

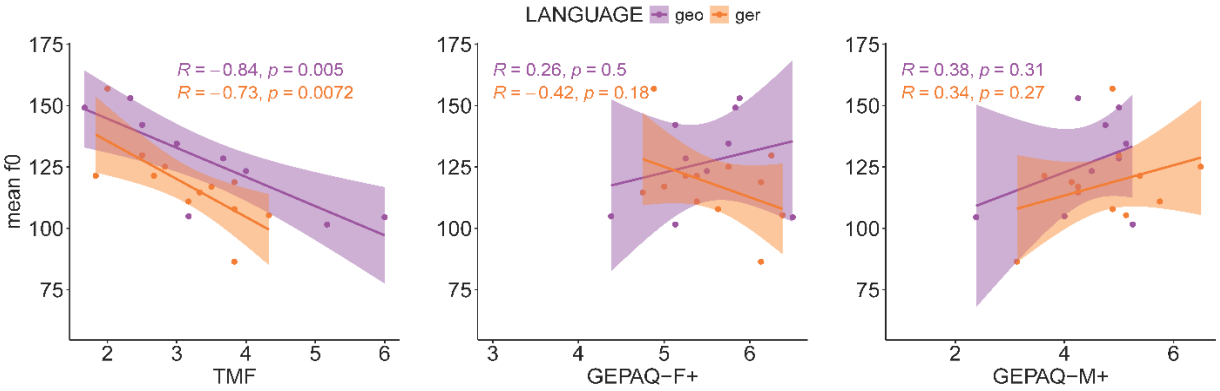



**Figure 3 -** Correlations between mean fundamental frequency and mean score of self-ascribed gender identity given in TMF (right), GEPAQ-F+ (centre) and GEPAQ-M+ (right) in Georgian (purple) and German (orange) subjects. Higher TMF and GEPAQ-F+ scores indicate higher femininity, and higher GEPAQ-M+ scores indicate higher masculinity.

### 3.3 VOT

The analysis of voice onset time (VOT) in relation to gender role identity reveals significant patterns across different stop types and positions in Georgian. Both ejectives and their voiceless aspirated counterparts show a significant relationship with feminine traits as measured by TMF-scale in Georgian speakers, although significance is not consistent across all observed sentence contexts. Specifically, we observe significant positive correlations between TMF scores and VOT for initial (sentence-initial and word-initial) velar ejectives, sentence-internal (word-medial and word-initial) voiceless aspirated velar stops, and for sentence-initial and word-initial /p/ as well as word-medial /t/. Figure 4 illustrates the significant relationships between TMF scores and mean VOT values across different Georgian stops in various sentence contexts. In contrast, German stops in our sample do not display significant correlations in any sentence context. Table 1 presents the results of correlation tests for all stops and contexts across the three gender identity scales in Georgian. Note that the reported significant correlations between VOT and TMF are unadjusted and should be interpreted as exploratory, as multiple testing on the same dataset increases the likelihood of false positives.

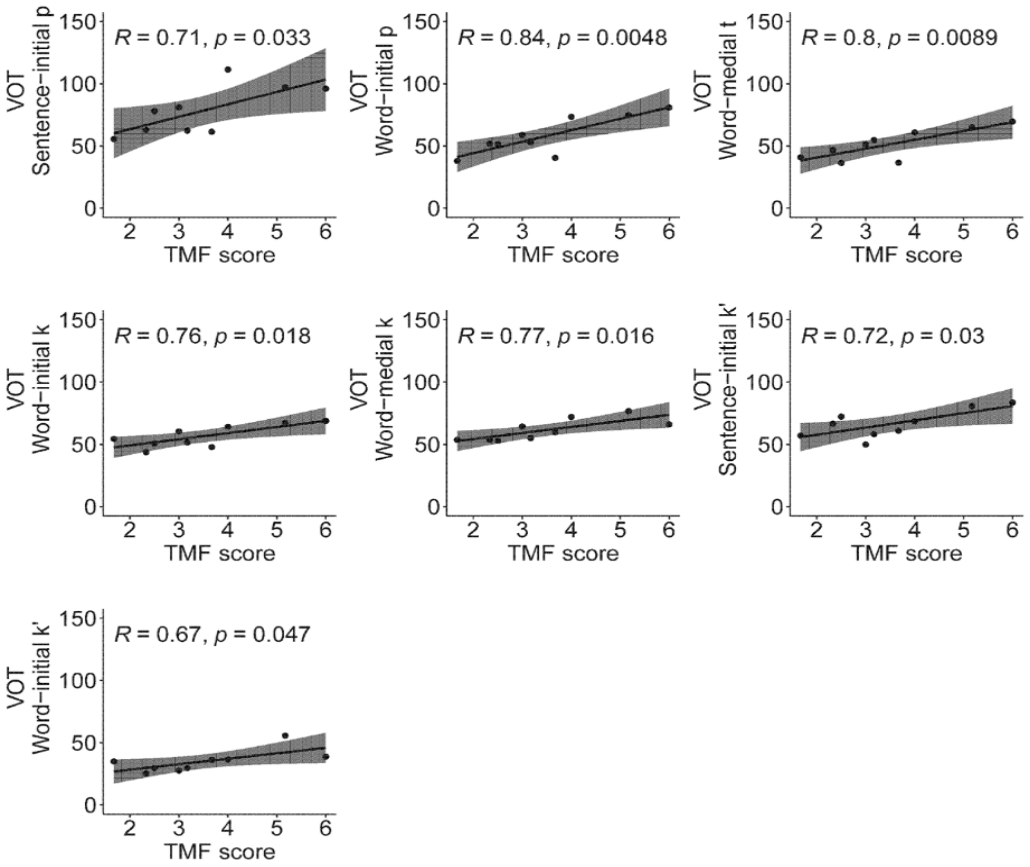



**Figure 4 -** Significant positive correlations between VOT and TMF-scores (higher score = higher femininity) in Georgian voiceless aspirated and ejective stops at labial, alveolar and velar place of articulations in different sentence contexts.

**Table 1** - Correlations of mean voice onset time and mean scores of TMF, GEPAQ-F+ and GEPAQ-M+ in Georgian speakers in different sounds and contexts. Bold p-values indicate significance <0.05 (unadjusted).

| | **TMF** | | **GEPAQ-F+** | | **GEPAQ-M+** | |
|---|---|---|---|---|---|---|
| **Sound** | **r** | **p-value** | **r** | **p-value** | **r** | **p-value** |
| sentence-initial p | 0.71 | **.03** | 0.22 | .58 | -0.21 | .58 |
| word-initial p | 0.84 | **.00** | 0.28 | .47 | -0.48 | .19 |
| word-medial p | 0.15 | .71 | 0.15 | .71 | 0.02 | .95 |
| sentence-initial t | 0.59 | .09 | 0.02 | .95 | -0.17 | .66 |
| word-initial t | 0.37 | .33 | 0.26 | .49 | 0.01 | .98 |
| word-medial t | 0.80 | **.01** | 0.23 | .56 | -0.52 | .15 |
| sentence-initial k | 0.62 | .08 | 0.04 | .92 | -0.23 | .56 |
| word-initial k | 0.76 | **.02** | 0.32 | .41 | -0.26 | .50 |
| word-medial k | 0.77 | **.02** | 0.12 | .76 | 0.03 | .94 |
| sentence-initial p' | 0.46 | .22 | 0.40 | .28 | -0.23 | .55 |
| word-initial p' | -0.01 | .98 | 0.25 | .51 | 0.17 | .67 |
| word-medial p' | 0.37 | .33 | -0.47 | .20 | -0.05 | .91 |
| sentence-initial t' | 0.45 | .22 | 0.19 | .62 | 0.13 | .74 |
| word-initial t' | 0.16 | .68 | 0.05 | .90 | 0.15 | .70 |
| word-medial t' | 0.57 | .11 | -0.28 | .46 | 0.02 | .97 |
| sentence-initial k' | 0.72 | **.03** | 0.24 | .53 | -0.48 | .19 |
| word-initial k' | 0.67 | **.05** | -0.04 | .92 | 0.11 | .78 |
| word-medial k' | -0.10 | .79 | 0.27 | .49 | 0.50 | .17 |

## 4 Discussion

The present analysis provides insights into the relationship between gender identity and phonetic features like mean f0 and VOT in Georgian and German speakers, revealing interesting patterns across the two languages. Contrary to initial expectations, a significant negative correlation was observed between f0 and TMF scores in both Georgian and German speakers. Although this pattern was surprising given the findings of [2] for German males, similar observations have been reported, for example, in a study on male Japanese-English bilingual speakers in their English [20]. Various factors could account for these contradictory effects, including cultural or language-specific differences, participant characteristics such as age, and study design elements like the choice of speech material or even the experimenter's gender, which can influence participants' speech behaviour [21].

In terms of VOT, significant positive correlations were found between TMF scores and VOT in specific phonetic contexts, including sentence- and word-initial velar ejectives (/k'/) and labial voiceless aspirated stops (/p/), as well as word-initial and word-medial voiceless aspirated velar (/k/) and word-medial labial (/k/). These findings suggest that self-reported femininity is associated with longer VOT durations, aligning with the findings of [10].

The limitations of this analysis, including a small sample size, the exclusive focus on male speakers, and the use of read speech material originally designed for investigating ejective production, should be acknowledged. Moreover, the significant correlations between VOT and TMF in the Georgian data warrant cautious interpretation due to the risk of Type I errors associated with multiple testing, however for the exploratory reasons p-values were left unadjusted in this analysis.

Despite these limitations, the observations presented underscore the need for further investigation with respect to the phonetic correlates of gender. In the future research we will incorporate a more diverse sample, including female speakers and spontaneous speech data, to explore whether similar patterns emerge.

## Acknowledgements

The Project is funded by the German Research Foundation (DFG, funding code: SI743/8). We are grateful to the German and Georgian participants for their participation in the extensive recordings and to our student assistant Lilia Kurnosova for her help with the labelling of audio files.

## References

[1] TRAUNMÜLLER, H. & ERIKSSON, A.*: The frequency range of the voice fundamental in the speech of male and female adults*. Stockholm, Department of Linguistics, University of Stockholm, 1995.

[2] WEIRICH, M., SIMPSON, A. P.: *Gender identity is indexed and perceived in speech.* PLOS ONE, 13. e0209226. 10.1371/journal.pone.0209226, 2018.

[3] SIMPSON, A. P.: *Phonetic differences between male and female speech.* Language and Linguistics Compass. 3, p. 621–640, https://doi.org/10.1111/j.1749-818X.2009.00125.x, 2009

[4] TITZE, I. R.: *Physiologic and acoustic differences between male and female voices*. The Journal of the Acoustical Society of America, 85(4), p. 1699-1707, https://doi.org/10.1121/1.397959, 1989.

[5] HENTON, C.: *Cross-language variation in the vowels of female and male speakers*. In *Proc. The 13th International Congress of Phonetic Sciences (ICPhS)*, p. 420–423, Stockholm, 1995.

[6] WEIRICH, M., SIMPSON, A. P.*: Changes in parents' phonetic parameters during the first year of a child's life. Comparative study of German and Swedish.* In *Proc. The 20th*

*International Congress of Phonetic Sciences (ICPhS),* 2023.
[7] PÉPIOT, E.: *Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers*. In: *Speech Prosody 2014*. ISCA; p. 305–309, 2014.
[8] STUART-SMITH, J.: *Changing perspectives on /s/ and gender over time in Glasgow*. Linguistics Vanguard, 6(s1):20180064, https://doi.org/10.1515/lingvan-2018-0064, 2020.
[9] KACHEL, S., SIMPSON, A. P., STEFFENS, M. C.: *Acoustic correlates of sexual orientation and gender-role self-concept in women's speech*. The Journal of the Acoustical Society of America, 141(6): p. 4793–4809, 2017.
[10] KACHEL, S., A. P. SIMPSON, und M. C. STEFFENS: *"Do I sound straight?": Acoustic correlates of actual and perceived sexual orientation and masculinity/femininity in men's speech.* Journal of Speech, Language, and Hearing Research, 61, p. 1-19, 2018.
[11] BRANDT, E., SIMPSON, A. P.: *The production of ejectives in German and Georgian*. Journal of Phonetics, 89, https://doi.org/10.1016/j.wocn.2021.101111, 2021.
[12] LEINER, D. J.: *SoSci Survey*. *[Computer Software].* URL: https://www.soscisurvey.de, 2019.
[13] KACHEL S., STEFFENS, M.C., NIEDLICH, C.: *Traditional Masculinity and Femininity: Validation of a New Scale Assessing Gender Roles*. Frontiers in Psychology, 7, DOI=10.3389/fpsyg.2016.00956, 2016.
[14] SPENCE, J. T., HELMREICH, R. L.: *Masculinity and Femininity: Their Psychological Dimensions, Correlates, and Antecedents*. University of Texas Press, 1978.
[15] RUNGE T. E., FREY, D., GOLLWITZER, P. M., HELMREICH, R. L., SPENCE, J.T.: *Masculine (Instrumental) and Feminine (Expressive) Traits: A Comparison between Students in the United States and West Germany.* Journal of Cross-Cultural Psychology, 12(2), p. 142–162, 1981.
[16] KISLER, T., REICHEL, U., SCHIEL, F.: *Multilingual processing of speech via web services.* Computer Speech & Language. 1(45), p. 326–347, 2017.
[17] BOERSMA, P., WEENINK, D.: *Praat*, *Doing phonetics by computer. [Computer Software].* URL: http://www.praat.org, Version 6.0.40, 2022.
[18] RCORE: *R: A language and environment for statistical computing. [Computer Software].* URL: https://www.R-project.org, 2022.
[19] WICKHAM, H.: *ggplot2*: *elegant graphics for data analysis*, URL: https://ggplot2.tidyverse.org, 2016.
[20] PASSONI, E., DE LEEUW, E., LEVON, E.: *Bilinguals Produce Pitch Range Differently in Their Two Languages to Convey Social Meaning.* Lang Speech, 65(4), p. 1071–1095, 2022
[21] KACHEL, S., SIMPSON, A. P., STEFFENS, M. C.: *Speakers' vocal expression of sexual orientation depends on experimenter gender.* Speech Communication, 156:103023, 2024.

# LANGUAGE REDUNDANCY EFFECTS ON PROSODIC PROMINENCE AND F0 PATTERNS IN STANDARD GERMAN

*Tianyi Zhao*

*University of Konstanz, Konstanz*
*tianyi.zhao@uni-konstanz.de*

**Abstract:** Previous research suggests that words with high language redundancy (i.e., predictability based on lexical, syntactic, semantic or pragmatic factors) tend to be produced with lower acoustic salience. The Smooth Signal Redundancy Hypothesis proposed that this inverse relationship between language redundancy and acoustic salience is controlled via prosodic prominence and boundary structure, and that f0, as one of the acoustic parameters, should also correlate with language redundancy. Following a previous study on durational measures, the current investigation focuses on the relationship between lexical frequency, repetition and prosodic prominence (including f0 contours and pitch accent types). Results showed that f0 can be affected by language redundancy in Standard German, but also indicated that effects of different language redundancy factors may correspond to different f0 patterns compared to duration.

## 1 Introduction

To ensure efficient and robust communication, speakers are assumed to manipulate signal redundancy by spreading it evenly throughout an utterance to maximise recognition likelihood of different speech sections [1]. According to the Smooth Signal Redundancy Hypothesis ([2, 3], henceforth SSRH), signal redundancy of linguistic items is determined via the inverse relationship between two levels: a) language redundancy, i.e., the predictability based on, e.g., lexical, syntactic, semantic factors and b) acoustic redundancy, i.e., acoustic salience based on acoustic properties such as duration, f0 and acoustic correlates that indicate phonemic distinctions like formant structure and spectral properties. These two levels are predicted to be inversely correlated: Words with low language redundancy, e.g., low lexical frequency, as they are less predictable than frequent words, would be produced more saliently (e.g., with longer durations) and vice versa. Furthermore, such correlation is proposed to be mediated via prosodic prominence (lexical and phrasal prominence) and boundary structure [3].

Previous research provided several lines of evidence for language redundancy effects on the durations of words, morphemes or syllables [2, 4, 5, 6]. Recent studies offered additional evidence for localized acoustic redundancy in the form of durational measures at the prosodic boundaries [7, 8]. While the existing experimental results revealed a consistent picture of durational measures, much less is known about how f0 patterns are related to language redundancy. Research has so far focused exclusively on English. With recordings of game moves in a conversational setting, [9] found that more predictable game moves can be reflected by durational reduction, smaller f0 excursions on words as well as fewer intonational phrase boundaries in utterances. The experiment distinguished between predictability and importance (in terms of how relevant and necessary the move is for winning the game) and results showed that importance was preferably signalled by intensity than by other acoustic cues.

Such potential distinction between different kinds of predictability was also discussed in [10], where it was discovered that predictability like semantic focus was associated with f0 in addition to word and vowel duration, whereas discourse mention mainly resulted in durational changes. A recent study [11] provided some support for the relationship between lexical frequency and f0 movement on different parts of the target words. Specifically, mean f0 values at the prosodic phrase boundary differed significantly across frequency condition, pointing to a larger span of boundary tones for infrequent targets. Despite a fine-grained analysis down to individual sonorant intervals of each target word, the observed f0 movement did not always behave in the same way as duration did in a previous study using the same dataset [12]. In general, durational effects were much more regular and consistent, while f0 seemed to show a less clear picture across studies.

The present study used the same data from one a previous experiment [8], where we confirmed the complementary relationship between lexical frequency, repetition and prosodic boundary strength, measured by duration and pause. As duration and f0 are known to be interactive [13, 14, 15], these data were used as a starting point to explore how language redundancy, in the form of lexical frequency and repetition, affects prosodic prominence and f0 patterns in Standard German. The use of optional intermediate phrasal breaks after the target words were also taken into account to view the prosodic constituent structure.

## 2 Methods

The current study focused on the prosodic prominence by means of accent types and f0 contours on the target words, using data collected in [8] which were used to analyse how language redundancy affects duration at prosodic word boundaries.

### 2.1 Participants

Thirty German native speakers (male = 6, female = 24; mean age = 24.2, *SD* = 4.3) were recruited and participated in the experiment at the University of Konstanz. None of them had any reported hearing or speaking impairments.

### 2.2 Materials

The target sentences were the same as in [8]. There were ten target pairs that were repeatedly produced three times in the experiment. Each pair consisted of a real target word - representing the frequent target, and a non-word as the infrequent target in that pair (see Table 1).

**Table 1** – An example of the target pair *Berlinerinnen* ('Berliners')-*Berbinerinnen* ('Berbiners') in the structurally similar carrier sentences.

| | | | |
|---|---|---|---|
| **Frequent (real word)** | | | |
| Drei | Berlinerinnen | fahren | zusammen |
| three | Berliner.FEM-PL | drive-PRS-PL | together |
| "Three Berliners drive together." | | | |
| **Infrequent (non-word)** | | | |
| Drei | Berbinerinnen | kochen | zusammen |
| three | *Berbiner.FEM-PL | cook-PRS-PL | together |
| "Three Berbiners cook together." | | | |

The target words were minimal pairs: The real words were derived from ten geographic

attributive words[1] and the non-words were created based on the real words by changing only the onset of the stressed syllable. The idea of the design was to assume an evident distinction in lexical frequency by the manipulation of real and non-words, as participants should have fundamentally more exposure to real words as compared to non-words. All non-words were phonotactically permissible in German. Target words were checked in a pilot experiment to ensure that the real words were familiar to the participants and that all target words could be pronounced properly.

In the initial experimental design [8], the target words were designed to be long (five syllables with the second syllable stressed) to disentangle the lengthening effects on lexically prominent syllables and prosodic boundaries, and to have an overview of the durational changes throughout the target words and their corresponding boundaries. This design benefits the current investigation as well allows for a detailed view on the f0 contour change over time. Each target sentence had four prosodic words and maximally two phonological phrases.

### 2.3 Experimental procedure

The experiment included a background questionnaire (including basic information and language background of participants) and a reading task. To aid the pronunciation of the long target words, participants were instructed to listen to the list of target words (produced by a female German native speaker) before the recording session. The word list was presented only once in audio form.

During the recording, each target sentence was displayed individually on the screen in a soundproof recording booth, while the participants were instructed to read them out loud. To minimise the possible contrastive focus on the target word, each sentence was followed by two filler sentences that were neither lexically nor syntactically related to the targets. Targets and fillers were then randomized into one block and each block was repeated three times by all participants as a within-subjects design. Productions were recorded with a condenser microphone in the Phonlab at the University of Konstanz (with a sampling rate of 48 kHz, 16-Bit, mono).

### 2.4 Data annotation and analysis

As described above, the ten target sentence pairs were produced three times by each speaker. For a clearer comparison across repetition, the duration as well as the f0 analysis only used the data of the first and third repetition. Target sentences with wrong stress patterns or hesitations within individual words were excluded, which left a total of 976 data points for the statistical analysis. Following [8], all productions were automatically segmented by MAUS [18, 19] and then manually corrected with Praat [20] based on standard segmentation criteria [21]. The target words were annotated syllable-wise, with the exception of the first syllables due to segmental reliability. The closure durations of the target onsets were merged with the boundary-related intervals that were part of the preceding word, e.g., *rei* ([ʁeɪ]) of *drei* ('three'). Rhymes of the first target syllables were segmented separately.

To analyse the f0 variation and its relation with the prosodic prominence on target words, we extracted 50 measurements per interval with ProsodyPro [22], i.e., 300 measurements per word (pre-target boundary inclusive). The retrieved time-normalized f0 contours were compared across lexical frequency and repetition. Given the non-linear and continuous nature of f0 trajectories, we used general additive mixed models (GAMMs, [23, 24]) for f0 analysis. The R

[1]The real words were selected based on frequency measures retrieved from the online corpus *Referenz und Zeitungskorpora* (since 04.03.2024 integrated into *Gegenwartskorpora*) in DWDS (*Digitales Wörterbuch der deutschen Sprache*, [16, 17]).

package *mgcv* was used for model fitting [25, 23] and the package *itsadug* was used to plot the model results [26]. There were two models fitted for lexical frequency and repetition condition respectively, both models were corrected for auto-correlation using a parameter determined by the function *ac_resid()* in the package *itsadug*. A third model that additionally included the smooth term to capture the interaction of frequency and repetition was dispreferred based on the output of *compareML()*.

Accent types of the target words and the optional phrasal breaks (break indices corresponding to 3 or 4 according to GToBI [27]) following the target words were manually annotated. To check the reliability of the perceptual annotations, a second annotator annotated approximately 20% (N = 200 sentences, 100 frequent and infrequent targets) of the full dataset. Interrater reliability was assessed by Cohen's Kappa [28] (using the *irr* package in R [29]). The interrater agreement for both phrasal break and accent type was "substantial" [30, p. 165] ($\kappa = 0.9$): The two annotators agreed in 96% of the cases for phrasal break and 97% for accent type, indicating reliable annotations. The annotations of the first annotator were used for further statistical analysis. The number of phrasal breaks and accent types was compared by $\chi^2$-tests.

# 3 Results

## 3.1 Phrasing and pitch accent

There were significantly fewer phrasal breaks following the frequent target words (frequency: $\chi^2 = 7.27$, df = 1, $p < 0.01$), especially upon the third repetition ($\chi^2 = 39.16$, df = 1, $p < 0.0001$). The accent status (prenuclear or nuclear) of the target words was also significantly associated with both lexical frequency and repetition (frequency: $\chi^2 = 6.79$, df = 1, $p < 0.01$; repetition: $\chi^2 = 28.97$, df = 1, $p < 0.0001$). There were very few cases, mostly in the third repetition, where the target word seemed to be deaccented (ca 1% of all data).

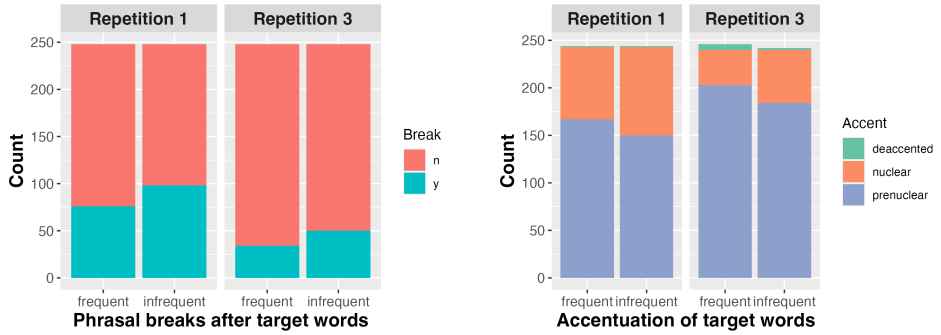



**Figure 1** – Descriptive statistics for phrasal break (left) and pitch accent type (right) of target words.

## 3.2 F0 contours

Overall, the average f0 contours showed higher distinctiveness between repetitions (Figure 2) than between frequent and infrequent targets (which is omitted here due to limited space). The illustrated peak on the unstressed syllable immediately after the stressed syllable could suggest the tendency for peak delay in the first compared to the third repetition, with a larger f0 range there for the first mention. In both panels, greater divergence of the f0 trajectories can be observed towards the end of the target words, which pointed to differences in the boundary tones as suggested also by the use of phrasal breaks. While the contours of frequent targets seemed to be divergent for almost the whole target word, the infrequent targets tend to be

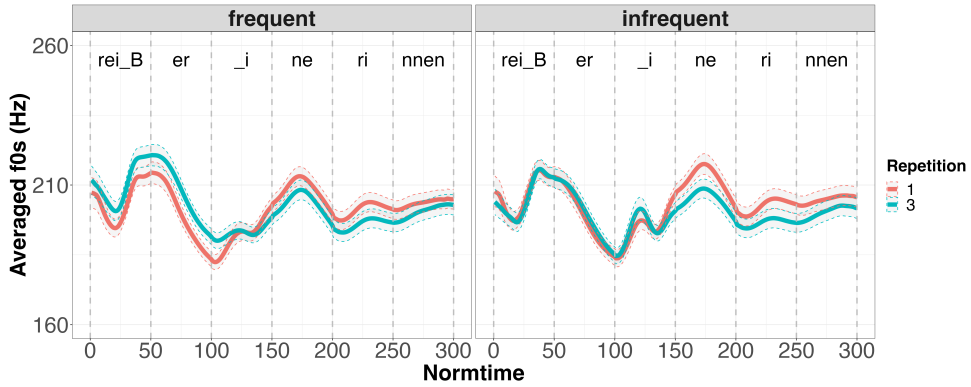



**Figure 2** – Time normalized average f0 contours by repetition on frequent and infrequent targets. Annotated intervals of the example pair *Berlinerinnen-Berbinerinnen* are provided for illustration.

distinguished more after the stressed syllables (given the possible peak delay) and in the latter half of the words.

The factor smooth for interaction between lexical frequency and repetition was not necessary based on ML score and AIC by model comparison (*compareML()*). In order to have a separate view of lexical frequency and repetition, we included them respectively as fixed effects in two different models, both with a factor smooth for the interaction of frequency/repetition condition over (normalized) time (*s(Normtime, by = frequency/repetition)*) and smooths for subjects and items (random slopes). The final models accounted for 74% of the deviance for lexical frequency and 74.8% for repetition.

As shown by Figure 3 (right panel), the estimated f0 differences between repetitions were significant throughout most intervals of the target word except for part of the boundary interval and the stressed syllable. However, no such statistical difference was found for the entire target word depending on frequency condition. A closer inspection within the repetition subset did not reveal different findings.

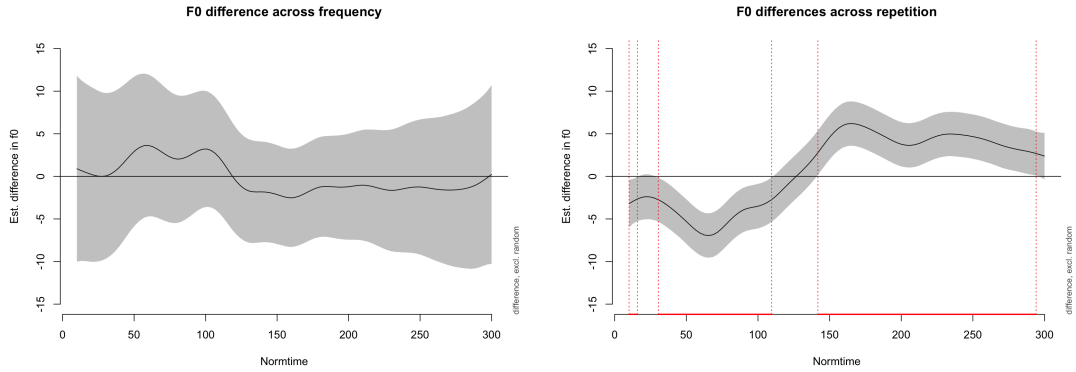



**Figure 3** – F0 difference curves of the target words based on GAMMs output, including the boundary interval at the left word edge. The frequency and repetition conditions are presented in left and right panels separately.

## 4 Discussion and conclusion

The results confirmed the robust effects of prosodic phrasing (e.g., in the form of phrasal breaks) but were not able to reveal a clear picture of f0 movement. Firstly, together with the previous findings on duration [8], prosodic constituent structure seems to be more sensitive to lexical frequency and repetition, the two language redundancy factors in this experiment, which provides

further support for the claim that language redundancy affects duration at prosodic boundaries [3].

Second, similar to the existing research on English varieties [9, 10, 11], lexical frequency and repetition yielded different f0 patterns. On the one hand, perceptual analysis showed that words with lower language redundancy were more likely to receive nuclear pitch accents, suggesting a connection between language redundancy and accent status of the linguistic item. On the other hand, the current f0 analysis showed that prosodic prominence can be reduced by repetition, but not necessarily by lexical frequency. This is in fact in line with the lack of significant duration differences on prominent syllables in [8], yet contradictory to the proposal of SSRH on the modification of prosodic prominence [2, 31]. However, it remains unclear whether this unexpected finding is due to experimental design, potential interference of peak delays, or different strategies of implementing prosodic prominence and constituency to accommodate to the reading task. As the target sentences were rather simple, speakers did not necessarily need to employ all resources available. Since the experiment was primarily constructed to test prosodic boundary strength, it is restricted to the real-non-word design with simplified sentence structure. The real and non-words are acknowledged to differ in their representations in the mental lexicon as non-words cannot be associated with specific meanings, which may lead to a lack of motivation to achieve a smooth signal in the utterance. For future research, longer utterances with varying syntactic structure or in a more conversational setting should be considered as they allow for more natural and intuitive f0 productions.

More importantly, the discrepancies in both language and acoustic redundancy measures further demonstrate that different types of redundancy factors need to be controlled and inspected separately during experimental manipulation. They may correspond to different f0 patterns or motivate acoustic cues in varying ways.

In summary, this study provides some useful inspirations for investigating the correlation between language redundancy measures and f0 in German data, which has so far received less attention in empirical research. Following [10, 11], the current findings confirm that language redundancy can affect f0 in German in some ways, but also raise further questions for f0 manipulations regarding other language redundancy measures and the intrinsic differences between language redundancy factors.

## Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with project BO 3388/4-1 and by the Arts and Humanities Research Council (AHRC) with project AH/W010801/1. We would like to thank the PhonLab at the University of Konstanz for the use of their recording facilities and the participants at P&P 2024 in Halle for the inspiring discussion.

## References

[1] LINDBLOM, B.: *Explaining Phonetic Variation: A Sketch of the H&H Theory*. In W. J. HARDCASTLE and A. MARCHAL (eds.), *Speech Production and Speech Modelling*, vol. 55, pp. 403–439. Kluwer Academic Publishers, Dordrecht, 1990. doi:https://link.springer.com/chapter/10.1007/978-94-009-2037-8_16.

[2] AYLETT, M. and A. TURK: *The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and*

*Duration in Spontaneous Speech*. *Language and Speech*, 47(1), pp. 31–56, 2004. doi:10.1177/00238309040470010201.

[3] TURK, A.: *Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis*. *Laboratory Phonology*, pp. 227–262, 2010.

[4] PLUYMAEKERS, M., M. ERNESTUS, and R. H. BAAYEN: *Articulatory planning is continuous and sensitive to informational redundancy*. *Phonetica*, 62, pp. 146–159, 2005.

[5] BELL, A., J. M. BRENIER, M. GREGORY, C. GIRAND, and D. JURAFSKY: *Predictability effects on durations of content and function words in conversational English*. *Journal of Memory and Language*, 60(1), pp. 92–111, 2009.

[6] LAM, T. Q. and D. G. WATSON: *Repetition is easy: Why repeated referents have reduced prominence*. *Memory and Cognition*, 38(8), pp. 1137–1146, 2010.

[7] ANDREEVA, B., B. MÖBIUS, and J. WHANG: *Effects of surprisal and boundary strength on phrase-final lengthening*. In *Proceedings of the 10th International Conference on Speech Prosody 2020*, pp. 146–150. 2020.

[8] ZHAO, T., T. BÖGEL, A. TURK, and R. N. DE SOUZA: *Language redundancy effects on the prosodic word boundary strength in Standard German*. In *Proccedings of Speech Prosody 2024*, pp. 1085–1089. Leiden, the Netherlands, 2024. doi:10.21437/SpeechProsody.2024-219.

[9] WATSON, D. G., J. E. ARNOLD, and M. K. TANENHAUS: *Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production*. *Cognition*, 106(3), pp. 1548–1557, 2008. doi:10.1016/j.cognition.2007.06.009.

[10] TURNBULL, R.: *The role of predictability in intonational variability*. *Language and Speech*, 60(1), pp. 123–153, 2017.

[11] ZHANG, C., C. LAI, R. NAPOLEÃO DE SOUZA, A. TURK, and T. BÖGEL: *Language Redundancy Effects on F0: A Preliminary Controlled Study*. In *Proceedings of the 20th International Congress of Phonetic Sciences ICPhS 2023*. Guarant International, Prague, Czech Republic, 2023. URL `https://guarant.cz/icphs2023/877.pdf`.

[12] BÖGEL, T. and A. TURK: *Frequency effects and prosodic boundary strength*. In S. CALHOUN, P. ESCUDERO, M. TABAIN, and P. WARREN (eds.), *Proceedings of the International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia*, pp. 1014–1018. 2019.

[13] VAN DOMMELEN, W. A.: *Does dynamic F0 increase perceived duration? New light on an old issue*. *Journal of Phonetics*, 21(4), pp. 367–386, 1993.

[14] ALAN, C.: *Tonal effects on perceived vowel duration*. *Laboratory Phonology*, 10(4), pp. 151–168, 2010.

[15] NIEBUHR, O. and J. WINKLER: *The Relative Cueing Power of F0 and Duration in German Prominence Perception*. In *Proceedings of Interspeech 2017*, pp. 611–615. 2017.

[16] BERLIN-BRANDENBURGISCHEN AKADEMIE DER WISSENSCHAFTEN: *Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*. `https://www.dwds.de/`, 2023. Last accessed: 12.04.2023.

[17] KLEIN, W. and A. GEYKEN: *Das Digitale Wörterbuch der Deutschen Sprache (DWDS)*. *Lexicographica*, 26, pp. 79–96, 2010. doi:10.1515/9783110223231.1.79.

[18] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In *Proc. of the ICPhS*, pp. 607–610. San Francisco, 1999.

[19] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 45, pp. 326–347, 2017. doi:10.1016/j.csl.2017.01.005.

[20] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer [computer program, Version 6.4.06]*, 2024. Available at http://www.praat.org/.

[21] TURK, A., S. NAKAI, and M. SUGAHARA: *Acoustic segment durations in prosodic research: a practical guide*. In S. SUDHOFF, D. LENERTOVÁ, R. MEYER, S. PAPPERT, P. AUGURZKY, I. MLEINEK, N. RICHTER, and J. SCHLIESSER (eds.), *Methods in Empirical Prosody Research*, pp. 1–28. De Gruyter, Berlin, New York, 2006.

[22] XU, Y.: *ProsodyPro-A tool for large-scale systematic prosody analysis*. In *Proceedings of TRASP 2013*. Laboratoire Parole et Langage, France, 2013. URL `https://core.ac.uk/download/pdf/16692792.pdf`.

[23] WOOD, S. N.: *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017. doi:https://doi.org/10.1201/9781315370279.

[24] WIELING, M.: *Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English*. *Journal of Phonetics*, 70, pp. 86–116, 2018. doi:https://doi.org/10.1016/j.wocn.2018.03.002.

[25] WOOD, S. N.: *Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models*. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1), pp. 3–36, 2011.

[26] VAN RIJ, J., M. WIELING, R. H. BAAYEN, and D. VAN RIJN: *itsadug: Interpreting time series and autocorrelated data using GAMMs*. `https://cran.r-project.org/web/packages/itsadug/itsadug.pdf`, 2022. R package.

[27] GRICE, M. and S. BAUMANN: *Deutsche intonation und GToBl*. *Linguistische Berichte*, pp. 267–298, 2002.

[28] COHEN, J.: *A coefficient of agreement for nominal scales*. *Educational and psychological measurement*, 20(1), pp. 37–46, 1960.

[29] M. GAMER AND J. LEMON AND P. SINGH: *irr: Various coefficients of interrater reliability and agreement*. `https://CRAN.R-project.org/package=irr`, 2012.

[30] LANDIS, J. R. and G. G. KOCH: *The Measurement of Observer Agreement for Categorical Data*. *Biometrics*, 33(1), pp. 159–174, 1977. doi:http://www.jstor.org/stable/2529310.

[31] AYLETT, M. and A. TURK: *Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei*. *Journal of the Acoustical Society of America*, 119(5), pp. 3048–3058, 2006. doi:https://doi.org/10.1121/1.2188331.

# Index