Mel-cepstral distortion of German vowels in different information density contexts

Erika Brandt, Frank Zimmerer, Bistra Andreeva, Bernd Möbius

Universität des Saarlandes, Saarbrücken, Germany

ebrandt@coli.uni-saarland.de

Abstract

This study investigated whether German vowels differ significantly from each other in mel-cepstral distortion (MCD) when they stand in different information density (ID) contexts. We hypothesized that vowels in the same ID contexts are more similar to each other than vowels that stand in different ID conditions. Read speech material from PhonDat2 of 16 German natives (m = 10, f = 6) was analyzed. Bi-phone and word language models were calculated based on DeWaC. To account for additional variability in the data, prosodic factors, as well as corpus-specific frequency values were also entered into the statistical models. Results showed that vowels in different ID conditions were significantly different in their MCD values. Unigram word probability and corpus-specific word frequency showed the expected effect on vowel similarity with a hierarchy between non-contrasting and contrasting conditions. However, these did not form a homogeneous group since there were group-internal significant differences. The largest distance can be found between vowels produced at fast speech rate, and between unstressed vowels.

Index Terms: information density, phonetic encoding, mel-cepstral distortion

1. Introduction

Linguistics has borrowed concepts from information theory [1] to explain differences in linguistic encoding. These concepts quantify the amount of information passed through the channel, also called information density (ID), as well as the channel capacity in communication processes. In recent years, numerous studies have shown that ID has an impact on linguistic structures on different linguistic levels, see [2] for extensive review.

Phonetic structures in high ID contexts are increased, while reduced phonetic structures are usually found in low ID contexts. Easily predictable vowels are less dispersed than vowels that are less predictable [3, 4]. Vowel dispersion is defined as the Euclidean distance between the average centre of the vowel space and formant values for each vowel [5]. This widely used measure includes formant measurements and extensive data cleaning: Formant tracking is prone to errors, especially for vowels which have close first and second formants, or close second and third formant. Manual verification of formant data is prohibitive for the analysis of large amounts of speech material.

As an alternative to vowel dispersion, the distance metric mel-cepstral distortion (MCD) can be used. MCD is the Euclidean distance between two vectors that describe the global spectral characteristics. Here, we used Mel-Generalized Cepstrals (MGCs) to describe the speech signal which is defined as the inverse Fourier transform of the generalized logarithmic spectrum calculated on a warped frequency scale [6]. The smaller the MCD values, the smaller the spectral distance between two speech signals. MCD only require a simple outlier cleaning procedure, and they are less prone to errors in calculating than formants. Therefore, MCD is suitable for the analysis of large speech corpora.

Prosodic factors, such as stress or speech rate, have an influence on the spectral characteristics of vowels. Vowels under stress are more dispersed than unstressed vowels [7]. Also, vowel dispersion shows a positive correlation with decreased speech rate [8]. In the Smooth Signal Redundancy hypothesis (SSR) prosodic structure holds a mediating position between information density and acoustic redundancy [9]. Highly predictable words are usually unstressed, while unpredictable words are prominent. Conventions and rules of prominence, as well as predictability of linguistic units are language-specific. Also, the influence of information density on phonetic encoding density seems to be language-specific [4].

Vowel spectral characteristics are known prosodic correlates but they are also sensitive to ID. In contrast to previous studies, we aim to analyze the global spectral characteristics of German vowels in different ID conditions. We expect that same vowel identities differ in their spectral characteristics when they are (a) in contexts with different predictabilities at the local level (bi-phone surprisal) and at the word level, and (b) produced with different prosodic structure (speech rate, and primary lexical stress).

We hypothesize that vowels in non-contrasting ID contexts are more similar than vowels in contrasting ID contexts. Vowels in high ID condition are the least similar in their spectral characteristics compared to vowels in low ID condition. Regarding the relationship between vowel similarity and prosodic factors, we expect to find small distances between vowels in the same stress and speech rate conditions. Consequently, vowels in contrasting stress and speech rate conditions are more distant from each other than in non-contrasting conditions.

2. Method

2.1. Material

2.1.1. Speech corpus

For the extraction of MCD values the PhonDat2 corpus with 16 German native speakers (m = 10, f = 6) was used [10]. Speakers read a corpus of 200 different screen-prompted sentences from a train query task. They were asked to read carefully but fluently as if in a real-life train query scenario. Canonical transcription of the corpus including primary lexical stress information is provided as part of the corpus. Audio files were downsampled to 16 kHz and then filtered to 8 kHz cutoff frequency. Automatic phonetic-phonemic segmentation was done using MAUS [11]. For a subset of 120 files per speaker these forced-aligned segment boundaries were manually verified by a phonetic expert for the subsequent data analysis. Only vowels in content

Table 1: *Total number of MCD values per vowel identity.*

Vowel	**No. of items**		
/ə/	180155	/iː/	2751
/ɐ/	255957	/ɪ/	97902
/øː/	489	/oː/	4254
/œ/	12921	/ɔ/	30976
/aː/	103096	/uː/	205
/a/	91384	/ʊ/	105720
/eː/	10636	/yː/	518
/ɛ/	11548	/ʏ/	11817
Total	940676		

words were analyzed because content and function words show different behavior with regards to information density [12]. Table 1 lists the number of MCD values per vowel.

2.1.2. Language modeling corpus

Word and phoneme language models were based on the DeWaC corpus which was preprocessed and normalized using German-Festival [13]. The web-crawled corpus DeWaC contains 1.5 billion running words and about 8 millions types with a diverse range of genres from newspaper articles to chat messages. Text-internal criteria consist of removal of web-specific structures, such as HTML structures or long lists. DeWaC was split into training (80 %) and test corpus (20 %).

2.2. Data analysis

2.2.1. Speech data analysis

MGC representations and MCD distance metric were calculated using SPTK 3.7 [14] at the temporal midpoint of each vowel. Before the distance metric was estimated, the optimal feature vector size for the MGCs was calculated since speech sound classes differ in this feature [6]. The optimal feature vector size for a respective data set can be estimated by using the diagonal of covariance matrices. The variance of features at size 5, 12, 19, 24, 30 and 39 were compared. For the vowels in the current data set, vector size 30 had the lowest variance ($Var(m30) = 6.434372e - 18$). Further parameters for MGC extraction were $\alpha = 0.42$, $\gamma = 2$ and frame length of 512. In a second step, the Euclidean distance between vowel vectors for the same vowel identity in different ID conditions were extracted. All MCD values larger than 10 were identified as outliers and cleaned from the data.

2.2.2. Language model

The ID measure used in this study was *surprisal* which is defined as $S(unit_i) = -log_2 P(unit_i|context)$. This measure is frequently used in psycholinguistic studies. It is relevant for human processing difficulty of linguistic units at phoneme, word, and sentence level [15, 16]. Surprisal values were obtained from a bi-phone LM using SRILM [17] including word and sentence markers, and using Witten-Bell smoothing. Surprisal was log-transformed due to positive skewness. One outlier was cleaned from the data ($S(X|X\text{-}1) > 25$). Values were binned into three equally large groups for low, mid, and high surprisal. In total, there were six categories of surprisal context: high-high, mid-mid, low-low, high-mid, high-low, and mid-low. Unigram word probability (WP) was obtained similarly based on a word LM of DeWaC, and binned in the same way as surprisal values. This again led to six categories of unigram word probability, parallel to those for bi-phone surprisal.

2.2.3. Prosodic model

The prosodic model that accounted for variability in the spectral characteristics of vowels contained lexical stress information and speech rate per sentence. Information on lexical stress was based on the canonical transcription of PhonDat2. Vowels were marked as stressed when they were the nucleus of the stressed syllable of the lexical item. Only primary lexical stress was included resulting in a three-level factor with factor levels stressed-stressed, unstressed-unstressed, stressed-unstressed. Each factor level denoted a comparative condition, similar to the ID conditions. The speech rate for each sentence was calculated as phonemes per second excluding pauses using Praat [18], values were mean-centered, and then binned into three categories (slow, normal, fast) to make data comparisons more feasible. In order to account for additional effects of prosody on MCD, vowels in the first and last word of each sentence were excluded from the analysis.

3. Results

3.1. Descriptive statistics

We expected to find a hierarchy in MCD values for vowels in different ID conditions: Vowels in non-contrasting ID conditions were expected to have small distances from each other resulting in low MCD values, while vowels in contrasting ID conditions were assumed to have larger distances from each other in the spectral domain. For UNIGRAM WP we found such a hierarchy with non-contrasting conditions low-low (m = -1.83, sd = 0.53), high-high (m = -1.79, sd = 0.54) and mid-mid (m = -1.78, sd = 0.56) having lower values than the contrasting comparisons high-mid (m = -1.77, sd = 0.56), high-low (m = -1.77, sd = 0.54) and mid-low (m = -1.76, sd = 0.55). This hierarchy was not replicated for MCD vowels in different SURPRISAL conditions. Here, vowels in mid-mid SURPRISAL were the most similar (m = -1.88, sd = 0.54), followed by vowels in mid-low ID condition (m = -1.80, sd = 0.51), high-mid (m = -1.79, sd = 0.55), low-low (m = -1.76, sd = 0.53) and high-low condition (m = -1.75, sd = 0.53), while same vowel identities in the non-contrasting condition high-high SURPRISAL were the most distant from each other (m = -1.74, sd = 0.58) (see Figures 1 and 2).

3.2. Linear mixed-effects model

Because of the specific domain of the speech data, word and syllable frequencies of the PhonDat2 corpus were included as control factors. In that way, effects on the spectral vowel characteristics that were due to corpus-specific frequency distributions were identified. PhonDat2 was syllabified using the g2p tool from BAS [19]. Frequency values were binned into three categories (low, mid and high frequency), and put into six comparative factor levels, similarly to the ID values based on DeWaC.

There are only slight correlations between the predictor values with regard to the dependent variable MCD. WORD AND SYLLABLE FREQUENCY of PhonDat2 are positively correlated ($r = 0.18$) since both were extracted from the same data set. WORD FREQUENCY of PhonDat2 and UNIGRAM WP based on DeWaC, however, ere negatively correlated ($r = -0.11$) indicating the domain-specific word frequency distribution of the speech material. LMM were calculated using “lme4” [20] and “lmerTest” [21]. The backward model selection method was ap-

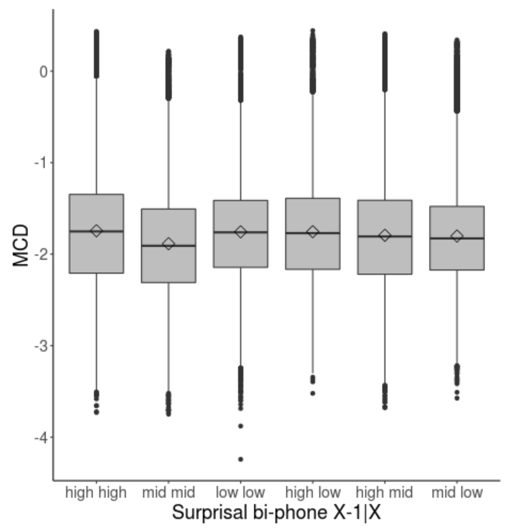

Figure 1: *MCD of German vowels in different* SURPRISAL *conditions.*

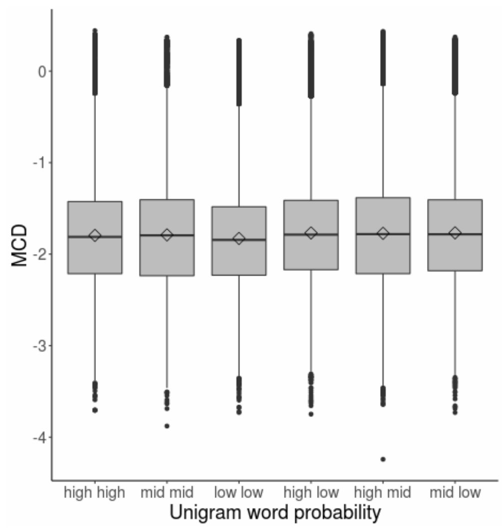

Figure 2: *MCD of German vowels in words with different* UNIGRAM WP *conditions.*

plied to identify the model that had the best fit for the data. The dependent variable MCD was log-transformed due to positive skewness. All categorical variables were treatment coded.

For the baseline LMM, the fixed effects SURPRISAL of the preceding bi-phone, STRESS, SPEECH RATE, UNIGRAM WP, PhonDat2 WORD FREQUENCY, and PhonDat2 SYLLABLE FREQUENCY, as well as GENDER were entered. The maximal random structure included random intercepts for SPEAKER and VOWEL IDENTITY, as well as random slopes for all fixed effects. Because of convergence errors the model was simplified, first removing random slopes. As the model converged, the predictor GENDER did not explain variance in the data and was therefore removed. Stepwise simplification resulted in a final model with random intercepts for SPEAKER and VOWEL IDENTITY. The coefficients, the $t-$test values and p-values are presented in Table 2. Reference level for all ID and corpus frequency factors was the comparative condition high-high. Reference level for the predictor value SPEECH RATE was the comparison between two vowels in sentences that were both produced at fast speech rate. For STRESS, reference level was the comparison between two vowels in syllables with primary lexical stress.

Results of the baseline LMM showed that for the predictors SURPRISAL, UNIGRAM WP, SPEECH RATE and STRESS all comparisons with their respective reference level led to significant results in explaining variability of MCD in German vowels. With regards to the control factor PhonDat2 SYLLABLE FREQUENCY, vowels in high-high syllable frequency condition were not significantly more similar than vowels in low-low condition. Vowels in the comparative condition PhonDat2 WORD FREQUENCY mid-low were not more distant from each other than vowels in high-high condition.

Regarding our hypotheses, post-hoc analysis using Tukey-tests was performed to identify differences between contrasting and non-contrasting conditions. Contrary to our hypothesis, non-contrasting comparative SURPRISAL conditions were significantly different from each other in their MCD. The same was true for non-contrasting UNIGRAM WP conditions, except for the comparison between mid-mid and low-low which was not significant (z = -0.64 , $p = 0.98$). In contrast to our expectation, we also found a non-significant differences between non-contrasting and contrasting ID conditions. MCD values in the condition high-high and high-mid SURPRISAL (z = 2.06, $p = 0.30$) did not differ from each other significantly. We found the same phenomena for the ID factor UNIGRAM WP. Comparing MCD values in high-high and mid-low UNIGRAM WP condition did not give a significant result (z = 0.54, $p = 0.99$) (see Figures 1 and 2).

For the prosodic factors, all STRESS conditions differed significantly from each other in post-hoc analysis. MCD values in stressed-stressed condition were significantly smaller than in stressed-unstressed (Coeff. = -0.06, z = -34.73, $p < 0.001$) and unstressed-unstressed condition (Coeff. = -0-04, z = -21.83, $p < 0.001$). Contrary to our hypothesis, MCD values in non-contrasting SPEECH RATE conditions differed from each other significantly according to LMM output and additional post-hoc analysis. Vowels in two slowly produced sentences were not less distant from each other than vowels in comparative conditions slow-fast (z = -1.73, $p = 0.50$) and slow-normal read speech (z = 0.35, $p = 0.99$). Also, there was no significant difference between MCD of vowels in normal-normal and normal-fast tempo (z = -0.53, $p = 0.99$).

The marginal pseudo-R^2 indicating how much variance is explained by the fixed factors showed that both ID effects explain 0.34 % of the MCD variance alone, control factors for domain-specific frequency distribution add 0.47% of the explained variance. When the prosodic model was added, explained variance increased by 0.14 %. The conditional pseudo-R^2 for the variance explained by both fixed and random effects equaled 12.94 % in the final model.

4. Discussion

This study aimed to investigate the global spectral characteristics of German vowels in different ID conditions. MCD values between same vowel identities in the same ID context and in contrasting ID conditions were compared. Confirming our hypothesis (a) SURPRISAL and UNIGRAM WP explained variability in the German MCD values: Vowels in contrasting and non-contrasting ID conditions were significantly different from each

Table 2: *Linear mixed-effects model for MCD. Baseline ID-prosody analysis.*

	Terms	Coeff.	t-value	p-Value
ID	Surprisal (h-l)	-0.03	-15.19	< 0.001
	Surprisal (h-m)	0.004	2.27	$= 0.02$
	Surprisal (l-l)	-0.06	-19.97	< 0.001
	Surprisal (m-l)	-0.03	-11.60	< 0.001
	Surprisal (m-m)	-0.04	-17.89	< 0.001
	Word (h-l)	0.01	6.25	< 0.001
	Word (h-m)	-0.01	-5.23	< 0.001
	Word (l-l)	-0.02	-7.00	< 0.001
	Word (m-l)	0.001	0.54	$= 0.59$
	Word (m-m)	-0.02	-7.69	< 0.001
Prosodic model	Stress (y-n)	0.02	12.98	< 0.001
	Stress (y-y)	-0.04	-21.83	< 0.001
	Speech rate (n-f)	0.01	5.77	< 0.001
	Speech rate (n-n)	0.009	4.13	< 0.001
	Speech rate (s-f)	0.02	9.95	< 0.001
	Speech rate (s-n)	0.02	7.75	< 0.001
	Speech rate (s-s)	0.02	7.14	< 0.001
Corpus frequency	Word (h-l)	0.04	16.39	< 0.001
	Word (h-m)	0.01	5.46	< 0.001
	Word (l-l)	-0.001	-0.74	< 0.001
	Word (m-l)	-0.009	-4.16	$= 0.09$
	Word (m-m)	-0.05	-18.96	< 0.001
	Syllable (h-l)	0.07	32.03	< 0.001
	Syllable (h-m)	0.05	20.91	< 0.001
	Syllable (l-l)	0.009	3.65	$= 0.08$
	Syllable (m-l)	0.02	9.28	< 0.001
	Syllable (m-m)	-0.01	-5.48	< 0.001

other. This observation, however, did not hold for all comparisons. Vowels in high-high SURPRISAL condition were not less distant from each other than vowels in contrasting conditions high-mid SURPRISAL. Same vowel identities in non-contrasting high UNIGRAM WP condition were not more similar than the same vowels in contrasting mid-low UNIGRAM WP condition.

Additionally, we hypothesized that there is a hierarchy in the distance metric modeled as a function of ID context: Smaller distances are supposedly found in same ID conditions, while larger distances are apparent between vowels in contrasting ID conditions. This hypothesis was confirmed for MCD in different UNIGRAM WP conditions, but not for different SURPRISAL conditions. Therefore, it seemed that UNIGRAM WP was the better ID measure to predict differences in MCD values for German vowels. However, non-contrasting and contrasting ID condition did not form homogeneous groups. As explained earlier, there were also significant differences between members of both categories. It should be noted that we found the same hierarchy for MCD values in different PhonDat2 WORD FREQUENCY conditions as for UNIGRAM WP: Vowels in non-contrasting conditions were less distant than in contrasting conditions with significant differences between all groups[1]. It followed that corpus-specific word frequency distribution also seemed to be an equally good predictor of spectral similarity between vowels.

As expected in hypothesis (b), the prosodic model explained variance of MCD of German vowels. Vowels in stressed syllables were less distant from each other than vowels in unstressed syllables, and when vowels in unstressed-stressed condition were compared. This finding possibly relates to larger variability in unstressed German vowels because unstressed syllables are produced with a higher degree of coarticulation [22]. Contrary to our expectations, non-contrasting SPEECH RATE conditions showed a clear hierarchy of MCD with lowest values for slow-slow, followed by normal-normal and than largest differences between vowels in sentences that were both produced at fast speech rate. Again, this result can be explained by larger variability in vowels at fast compared to slow speech rate [23]. However, the global sentence-based measurement of speech rate did not capture local speech rate deviations. These would possibly be a better predictor of locally measured MCD values.

Analysis of marginal pseudo-R^2 revealed that the ID conditions explained a larger part of the variance in the dependent variable than the prosodic model that was used here, contrary to previous studies [3]. Still, the spectral characteristics of vowels were only subtly influenced by ID, as expected [2]. The amount of explained variance highly depends on the fit of the predictor values for the dependent variable. It is likely that the prosodic model increases in its strength if additional factors were added, for instance realised prominence, phrasal accent or boundary strength.

Most variance in the MCD values was explained by random intercepts for SPEAKER and for VOWEL IDENTITY. This finding can be explained by vowel-inherent variability and markedness of vowels. For instance, /ə/ was not found in the high SURPRISAL condition, while vowels /øː, œ, yː, ʏ/ only stood in high SURPRISAL bi-phones. Also, investigation of the conditional modes of the random intercepts showed that the large vowel-inherent variability within /ə/ and /ɐ/ was reflected in overall larger MCD values in pairwise comparisons than for all other German vowels. Coefficients were 0.24 and 0.23 respectively.

The current analysis was based on read speech. In spontaneous speech, effects of information density on phonetic structure have been found to be more pronounced than in other speech registers [24] suggesting that the patterns found here might at least also be found in other registers.

5. Conclusions

The current study showed that German vowels differ significantly from each other when they stand in different ID contexts. However, information density was not a strong predictor of MCD values of German vowels. The prosodic model explained even less variance than ID which was possibly due to a weak model that was based on canonical stress and a global sentence-based speech rate measure. Unigram word probability and corpus-specific word frequency showed expected tendencies in MCD: Smaller distances were found between vowels in the same conditions compared to vowels in contrasting conditions.

6. Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of SFB 1102 'Information Density and Linguistic Encoding' at Saarland University.

[1] Results for the control factor were not reported in the Results section.

7. References

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.

[2] T. F. Jaeger and E. Buz, *Handbook of Psycholinguistic*. Oxford, United Kingdom: Wiley-Blackwell, 2017, ch. Signal reduction and linguistic encoding, accepted.

[3] M. Aylett and A. Turk, "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," *Journal of the Acoustical Society of America*, vol. 119, pp. 3048–3058, 2006.

[4] E. Schulz, Y. M. Oh, Z. Malisz, B. Andreeva, and B. Möbius, "Impact of prosodic structure and information density on vowel space size," in *Speech Prosody 2016*, Boston, 2016.

[5] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, "Intelligibility of normal speech. i. global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, vol. 20, pp. 255–272, 1996.

[6] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *International Conference on Spoken Language Processing*, vol. 3, Yokohama, Japan, 1994.

[7] D. R. van Bergem, "Acoustic vowel reduction as a function of sentence accent, word stress, and word class," *Speech Communication*, vol. 12, pp. 1–23, 1993.

[8] B. Weiss, "Rate dependent vowel reduction in German," in *Proceedings of the 12th SPECOM*, Moscow, 2007.

[9] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language*, vol. 47, no. 1, pp. 31–56, 2004.

[10] PHONDAT2 - PD2. (1995) University of Munich. Version 2.8.

[11] T. Kisler, R. U. D., F. Schiel, C. Draxler, B. Jackl, and N. Pörner, "Bas speech science web services - an update of current developments," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia, 2016.

[12] A. Bell, J. 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*, vol. 60, no. 1, pp. 92–111, 2009.

[13] B. Samlowski, "The syllable as a processing unit in speech production: evidence from frequency effects on coarticulation," Ph.D. dissertation, University of Bielefeld, Bielefeld, 2016.

[14] T. Kobayashi, K. Tokuda, and K. Koishida, *Speech Signal Processing Toolkit (SPTK)*, 2013, version 3.7.

[15] V. Demberg, A. B. Sayeed, P. J. Gorinski, and N. Engonopoulos, "Syntactic surprisal affects spoken word duration in conversational contexts," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 356–367.

[16] R. Levy and T. F. Jaeger, "Speakers optimize information density through syntactic reduction," *Advances in Neural Information Processing Systems*, vol. 19, pp. 849–856, 2007.

[17] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, Denver, 2002, pp. 901–904.

[18] P. Boersma and D. Weenink. (2017) Praat: doing phonetics by computer [computer program]. version 6.0.26. [Online]. Available: retrieved 1 January 2017 from http://www.praat.org/

[19] U. Reichel and T. Kisler, "Language-independent grapheme-phoneme conversion and word stress assignment as a web service," in *Elektronische Sprachverarbeitung. Studientexte zur Sprachkommunikation 71*, R. Hoffmann, Ed. Dresden: TUD-press, 2014, pp. 42–49.

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

[21] A. Kuznetsova, P. Bruun Brockhoff, and R. Haubo Bojesen Christensen, *lmerTest: Tests in Linear Mixed Effects Models*, 2016, r package version 2.0-33. [Online]. Available: http://CRAN.R-project.org/package=lmerTest

[22] C. Mooshammer and C. Geng, "Acoustic and articulatory manifestations of vowel reduction in German," *Journal of the International Phonetic Association*, vol. 38, no. 2, pp. 117–136, 2008.

[23] B. Lindblom, "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, vol. 35, pp. 1773–1781, 1963.

[24] I. Hanique and M. Ernestus, "Final /t/ reduction in Dutch past-participles: The role of word predictability and morphological decomposability." in *Interspeech*, 2011, pp. 2849–2852.