Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

How prosody affects ASR performance in conversational Austrian German

Saskia Wepner[1], *Barbara Schuppler*[1], *Gernot Kubin*[1]

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

wepner@tugraz.at, b.schuppler@tugraz.at, gernot.kubin@tugraz.at

Abstract

Currently available Automatic Speech Recognition (ASR) systems achieve good word error rates (WER) for read speech ($2-10\%$), but not for conversational speech ($20-40\%$), a speaking style especially relevant for dialogue systems, as they become more conversational and interactional. Here, we analyse how prosody affects WER in a Kaldi-based speech recognition system for a corpus of conversational Austrian German. This analysis is a step towards improving ASR systems and increasing our knowledge about which aspects are relevant to consider for ASR of conversational speech. For this purpose, we compare a typical language model (LM) with an oracle LM trained on the utterances from the whole corpus, thus knowing each possible N-gram in advance. We find that short, deaccented words have the lowest recognition accuracy, which also cannot be compensated for by the oracle LM. Despite our overall high WERs, the highly prominent words were recognised significantly better. Our findings suggest that reporting global WERs for an ASR system of conversational speech does not predict its usefulness in dialogue systems. Given the role of prominent words in carrying meaning and function in conversation, our analysis is relevant for researchers developing automatic speech understanding systems.

Index Terms: Prosodic prominence, Automatic Speech Recognition, Kaldi, conversational speech, Austrian German

1. Introduction

Currently available ASR systems perform well for read speech (100–90% accuracy), but not for conversational speech (80–60%) [1, 2]. Highly accurate ASR systems for conversational speech are especially relevant for dialogue systems, as they become more conversational and interactional rather than solely transactional [1]. Thus, an increasing number of studies have investigated the differences between these speaking styles in order to improve ASR performance for conversational speech.

On the search for solutions on how to improve ASR for conversational speech, studies analysed the potential role of prosody for ASR performance. Goldwater et al. [3] compared two ASR systems [4, 5] for English telephone conversations. They found that word errors occur with word tokens that have extremely low intensity or extremely high F0 values. Regarding duration, they detected more errors with lower durations. Further, turn-initial words had higher Word Error Rates (WERs) than medial or turn-final tokens in Goldwater's study where they also reported, high variation in recognition performance depending on the speaker.

In line with these studies, we present an analysis on how prosodic structures affect WER based on a corpus of casual conversations among friends speaking Austrian German. The aims of our analysis is not only to show how prosody affects ASR performance, but to find the potential role of the language model (LM) in compensating difficulties originating from prosodic variation. For this purpose, we compare recognition results with two different LMs: 1) with a standard LM based on training data only (excluding development and evaluation data, hence named "LMnormal") and 2) with an 'oracle' LM trained on all the data (including development and evaluation data, hence named "LMoracle").

Attempts to incorporate prosodic information into speech recognition systems have been made in acoustic models (AMs), pronunciation lexicons, and LMs. ASR systems have been successfully improved by building prosody-dependent AMs (e.g., [6, 7, 8, 9]). Chen et al. [10], for instance, explicitly incorporated the well-known process of phrase-final lengthening into their AMs. These prosody-dependent AMs capture variation in terms of the spectral characteristics of phones and their durations. However, pronunciation variants resulting from multiple segment deletions and substitutions, such as the pronunciation of 'yesterday' as [ˈjɛʃɐɪ], cannot be captured with AMs alone.

Despite findings showing that segment deletions also depend on the prosodic status of a word, research towards building prosody-dependent pronunciation models is limited. [11] and [12] came to promising results, but never tested their prosody-dependent models in an ASR task. For phone recognition, [7] achieved improvements by combining pronunciation models using F0, duration and energy and AMs using information about syllable position and stress.

A decade ago, new statistical approaches were introduced that allow incorporating linguistic knowledge into language models. Huang and Renals [13], for instance, showed that syllable-based prosodic features help to reduce LM perplexity and to marginally reduce WER. Chan [14] improved WER by using a prosody-dependent LM based on maximum-entropy Markov models (MEMMs). Chien and Chueh [15] reported that WERs decrease if AM and LM are modeled together. Chen et al. [8] decreased WER by 11% by incorporating information on phrase boundaries and pitch accents into the AM and LM of their recogniser. Su and Jelinek [16] also improved their random forest LM with information on prosodic breaks. These studies use a symbolic representation of prosody. In contrast, [17] used prosodic features directly and yielded significant perplexity reductions of LMs. In our own experiments on homophone classification in spontaneous German, we achieved significant improvements by having the language model learn from both lexical and prosodic features simultaneously [18].

2. Materials

The Graz Corpus of Read and Spontaneous Speech (GRASS) [19, 20] contains about 30h of Austrian German read and conversational speech, collected from 38 Austrian speakers (19f, 19m). As language use in conversational speech varies strongly with educational level, social background and dialect region, we selected speakers who were born in the same broad dialect region (Eastern Austria), had been living in an urban area for years and had a higher education degree. For the conversa-

10.21437/SpeechProsody.2022-40

tional speech component, 19 pairs of speakers who have known each other for several years were recorded for one hour each without interruption in order to encourage a fluent, casual conversation. There was no restriction in terms of chosen topic or speaking behaviour, leading to the use of authentic, partly dialectal pronunciation with its typical characteristics, such as frequent occurrence of overlapping speech, laughter, or the use of swear words [20]. Despite the speakers' awareness of being recorded, most of them appeared to completely disregard the studio recording setting after a period of five to ten minutes, entering into a completely casual, face-to-face conversation.

Prosodic annotations were created following the Kiel Intonation Model (KIM [21]), and included annotations for prosodic boundaries, sentence accent (i.e., prominence levels 0, 1, 2, 3) and pitch contour annotations for words of prominence level > 0 (i.e., distinguishing early, medial and late peak, early and late valley and flat pitch contours). Whereas decisions concerning prosodic boundaries and prominence levels were taken on a purely perceptual basis, for the decision on pitch contour labels, spectrogram and F0 contour were inspected as well. To guarantee a high annotation quality, conversations were first annotated by one trained annotator and then checked by another annotator.

For this study, we excluded the 35 tokens with prominence level 3, and tokens with the label indicating that no decision on the label could be made among the annotators. This resulted in a total of 4169 prosodically annotated word tokens.

3. Kaldi-based ASR system

We set up a Kaldi [22] recipe that had been optimised for the GRASS's conversational component with the parameters that are described in the subsections below. The data was split into training (80%), development (10%) and evaluation set (10%). We performed cross-validation by splitting the data into different sets for each evaluation run. In every split, the system was optimised on the development set and then tested with the evaluation set. The speakers of evaluation and development set were *not* part of the training set in any of the splits. For the language models, we did include utterances of both development and evaluation set in the training one of the two language models (see Subsection 3.3).

3.1. Acoustic models

We extracted 13 mel-frequency cepstrum coefficients (MFCCs) and performed cepstral mean and variance normalisation (CVMN) which is a standard technique to obtain more noise-robust features [23]. With a frame length of 20ms and a frame shift of 12.5ms, we trained subspace Gaussian Mixture Models (sGMM) with feature space Maximum Likelihood Linear Regression (fMLLR) and a final discriminative training on boosted maximum mutual information (bMMI, boost 0.1).

3.2. Pronunciation lexicon

For the German words in the corpus, we generated pronunciation variants with a set of 34 rules that account for pronunciation processes that are typical in spontaneous German (e.g., schwa deletion) and rules specific for the Austrian German variety (e.g., monophthongation) [24]. Together with foreign language words and broken words, this resulted in a lexicon with 83979 entries for 13701 word types (one word type may have several pronunciations). Since an automatic generation of pronunciation variants does produce some variants that are very unlikely to be actually realised by the speakers, the lexicon contains much more variants than needed. E.g., long words with four or more syllables, such as prefix verbs or compounds which are quite common in German, may have more than 30 variants. The pronunciation lexicon included words of the development and evaluation sets, thus there are no out-of-vocabulary words.

3.3. Language models

As a statistical Language Model (LM), we trained an SRILM [25] 4-gram (referred to as *LMnormal*). Due to the small size of our data set coupled with its diversity, it is not surprising that we are dealing with relatively high WERs (see Section 3.4). Since this makes it difficult to trace misrecognitions back to the acoustic properties of misrecognised tokens, we trained a second LM to which we already revealed the utterances from the development and evaluation sets yielding an LM that only suggests best possible N-grams. We call this second model *LMoracle*, a terminology borrowed from the literature in the fields of speech enhancement and speaker separation [26]. Obviously, our ASR setup produces much lower WERs with *LMoracle* (see Table 1).

3.4. Recognition results

Figure 1 shows that recognition performance did strongly increase with *LMoracle*, which is expected since this LM is trained on the whole corpus data and not the training set only. The mean WER per utterance with *LMnormal* is 53.43% and 33.12% with *LMoracle*. Still we saw a strong speaker dependency for both LMs. On average, *LMnormal* recognised words for both sexes equally well (female: 53.33%, male: 53.89%), where *LMoracle* deals slightly better with female speakers (29.61%) than with male speakers (33.08%). The greatest decrease in WER was found for speakers 029F (32.20%) and 027F (31.37%), the lowest for 031F (14.35%) and 012M (16.73%).

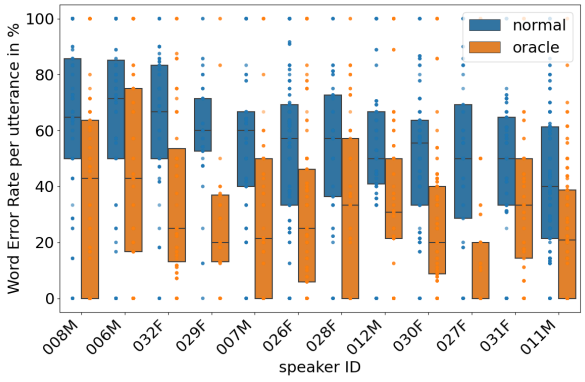

Figure 1: *Comparison of the WERs per utterance of the two LMs, speakers are sorted by descending mean WER with LMnormal.*

4. Analysis of prosodic effects on WER

4.1. Statistical analysis

In order to test which prosodic characteristics affect WER in our ASR system, we built mixed-effects logistic regression models with *Correct* (false, true) as dependent variable, indicating whether a token was recognised correctly or not. We included the independent variables *LogFrequency* (i.e., the natural loga-

rithm of the frequency of the word in the whole GRASS corpus, $\text{range} : 0 \ldots 8.95$), *Duration* (i.e., the absolute duration of the word in ms, $\text{range} : 27\text{ms} \ldots 13200\text{ms}$), and whether it occurred in *Overlap* (false, true) with the interlocutor, as well as the random variables *Speaker* and *Word*. For the model on the complete set of prosodically annotated data ($N = 4169$), we included the (categorical) prosodic variables *Position*, which indicated whether a word occurred in initial (I), medial (M) or final (F) position within the prosodic phrase and prosodic *Prominence* (on a discrete scale from 0 to 2). We built separate models on the subset of tokens with prominence level 1 or 2, to which pitch contour labels were assigned ($N = 2192$), including either the independent variable *Pitch_Type* (flat, peak, valley) or *Pitch_Contour* (flat, medial peak, early peak, late peak, early valley, late valley).

Models were built using the `glmer()` function of the `lme4` package in *R* [27]. Predictors and interactions were reduced by stepwise backward selection, and they were removed as models would still significantly improve as given by their AIC value and their degree of freedom. Random variables were only kept in the model if they improved the model as given by model comparison using the `anova()` function [28, 29].

4.2. How prosodic prominence and position within the phrase affect WER

First, we built mixed-effects logistic regression models on the complete data set, separately for *LMoracle* and *LMnormal* (cf. Table 1). With both LMs, words that were recognised correctly tended to have significantly longer duration and higher frequency than misrecognised words. The significant interaction term between *Duration* and *LogFrequency*, however, is in opposite directions for *LMoracle* and *LMnormal*, indicating that with *LMoracle* the positive effect of frequency on recognition likelihood tends to be lower for long words whereas with *LMnormal* it tends to be even larger for long words. For both LMs, words in (prosodic) phrase initial positions were recognised significantly worse than words in phrase medial or final position, and for *LMoracle* this effect is even larger for long words. Unsurprisingly, words in initial position which were produced in overlap with the interlocutor, tended to be recognised even less frequently (i.e., as indicated by the significant interaction term between *Position* and *Overlap*, cf. Table 1).

For both LMs and with high significance, the prominent words were recognised better than deaccented words: for *LMnormal*, 45.7% of words with prominence level 0 were recognised correctly, 53.2% with level 1 and 50.5% with level 2. For *LMoracle*, the gain in accuracy was even higher for words of prominence level 2 (62.8% (0), 71.7% (1) and 83.0% (2)), with a highly significant improvement for words with prominence level 2 from those of level 1 ($\beta = 0.49, z = 3.67, p < 0.001$). The analysis of word durations shows a similar effect: For prominence level 2, the recognised words in *LMnormal* are on average 29.4ms shorter than those in *LMoracle*, whereas misrecognised words in *LMnormal* are 62.3ms longer than those in *LMoracle*. Thus, the additional knowledge available to *LMoracle* can compensate for prominent words which often show longer duration, but not for deaccented words, which are typically high-frequency and short function words.

By examining word types (i.e., different kinds of words) within the prominence levels, we found $N = 304$ different word types within prominence level 0, $N = 514$ types within level 1 and $N = 636$ types within level 2, which meets our expectation for a larger variety of word types towards more prominent words. Furthermore, in the deaccented words, we could find most of the word types that occur frequently as homophones in conversational speech; for instance, *das, dass, es, ist, sie, so* are commonly highly reduced and thus often realised as a single phone [s], or *dann, den, in, nein, und* are frequently realised as [n]. The possible homophones occurred most frequently in prominence level 0 ($N = 314$), while only $N = 70$ could be found in prominence level 1 and $N = 12$ in level 2.

Table 1: *Regression models predicting WERs for LMoracle ($AIC = 4611.2$) and LMnormal ($AIC = 4914.7$) in the evaluation set ($N = 4169$), including the significantly contributing random intercepts Speaker and Word; dash (–) refers to variables not part of the model; estimator β, z-value and significance level (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, n.s.: $p > 0.05$) are given according to e.g., [28].*

	LMoracle			**LMnormal**		
Predictor	β	z-value		β	z-value	
Intercept	-1.66	-3.20	**	-4.62	-8.64	***
LogFrequency	0.24	3.87	***	0.48	6.65	***
Position (I)	-2.45	-4.41	***	-0.87	-2.61	**
Position (M)	-0.49	-1.08	n.s.	-0.10	-0.47	n.s.
Prominence (1)	0.82	3.31	***	1.67	3.11	**
Prominence (2)	1.48	4.02	***	2.61	4.76	***
Duration	6.96	4.43	***	3.96	2.79	**
Overlap (T)	-0.46	-1.85	n.s.	–	–	
LogFrequ.: Dur.	-0.43	-2.77	**	0.39	2.42	*
LogFrequ.: Pos. (I)	0.27	4.73	***	0.17	3.36	***
LogFrequ.: Pos. (M)	0.03	0.59	n.s.	0.02	0.75	n.s.
LogFrequ.: Overl. (T)	0.09	2.63	**	–	–	
Prom. (1): Dur.	-2.90	-2.48	*	-1.14	-0.92	n.s.
Prom. (2): Dur.	-3.31	-2.44	*	-2.99	-2.30	*
Prom. (1): LogFrequ.	–	–		-0.15	-2.56	*
Prom. (2): LogFrequ.	–	–		-0.14	-2.08	*
Pos. (I): Duration	2.94	2.78	**	–	–	
Pos. (M): Duration	2.55	2.74	**	–	–	
Pos. (I): Overl. (T)	-0.42	-1.71	n.s.	–	–	
Pos. (M): Overl. (T)	-0.32	-1.57	n.s.	–	–	

Comparing the recognition results of *LMoracle* vs. *LMnormal*, *Overlap* only showed (marginally significant) deteriorating effects with *LMoracle*. Drawing conclusions about this effect, requires further investigation in the future.

4.3. How pitch contour affects WER

In order to investigate whether word recognition is affected by the word's pitch contour (i.e., early, medial, late peaks, early and late valleys and flat contours), we built mixed effects logistic regression models on the subset of the tokens having prominence levels 1 and 2 ($N = 2190$), i.e., those with pitch contour annotations. We did so separately for *LMoracle* and *LMnormal*. These models did not reveal any significant effects of *Pitch_Contour* on WER. Only with *LMnormal*, we found a marginally significant interaction between *Pitch_Contour* and *Overlap*, indicating that words with a peak tend to be recognised worse than words with a flat pitch, and this effect is significantly higher for words produced in overlap with the interlocutor ($\beta = -0.48, z = -1.79, p < 0.1$).

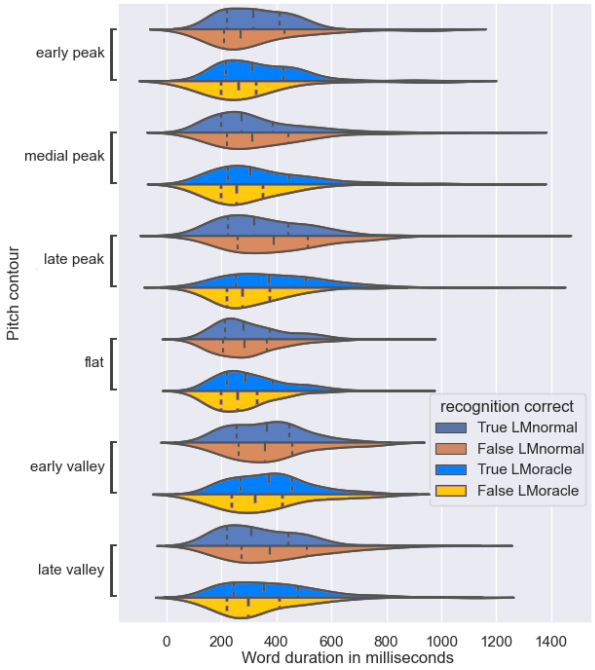

Figure 2: *Pitch contour category of (mis-)recognised tokens by duration for the two Language Models.*

Comparing the performances of *LMnormal* and *LMoracle* for the different pitch contour categories regarding duration, the two systems show considerable differences with respect to how well they deal with each pitch contour category. Whereas *LMnormal* recognised less than half of the late peak tokens (49.08%) correctly, *LMoracle* yielded the best recognition (80.28%) for this pitch contour (which is also the largest difference of the LMs' recognition performance with 31.19%). The average duration of misrecognised late peak tokens is 100.17ms lower for *LMoracle* than for *LMnormal*; late peak tokens recognised correctly were on average 49.79ms longer for *LMoracle* than those recognised correctly by *LMnormal*. Worst recognition is achieved for medial peak with *LMoracle* (73.81%) and for late valley with *LMnormal* (50.52%). As with prominence levels, we found that *LMoracle* misrecognised less of the words with longer duration, indicated by a larger shift of the mean towards lower durations in *LMoracle* than in *LMnormal* for all pitch contour categories (see Figure 2 and Table 2).

We analysed the ratio of word types to word tokens within the pitch categories, in order to investigate whether the underlying effect of different recognition performance between these categories potentially stems from their frequency. For flat contours, the ratio of word types to word tokens is 0.52, meaning that though this category comprises most word types, concerned tokens are rather frequent. Late valleys and all peak contours have comparable ratios (i.e., late valley: 0.76, early peak: 0.74, medial peak: 0.72, late peak: 0.73), and early valleys have a ratio of 0.81 types to tokens. Thus, neither the absolute WER between the categories, nor the differences between the two LMs seems to be an effect of word frequency.

5. Conclusion

We analysed how the prosodic properties of a word affect its recognition with a Kaldi-based ASR system. We compared two language models; one that was trained on the utterances in the training data (*LMnormal*) and one that was trained on utterances of the whole corpus, thus knowing each possible N-gram in advance (*LMoracle*). Overall, the WER for *LMnormal* was 53.43% and for *LMoracle* 33.12%, indicating that though its rather small size, our corpus contains a broad variation which stems from both high speaker variation and the inherently difficult nature of our casual, conversational speech material. In line with what Goldwater et al. [3] reported for conversational telephone speech in American English, the variation observed among speakers in our data was very large with both LMs.

Our analysis showed that words with longer durations were recognised more often than shorter words. We found that recognition performance of long words scaled with their frequency for both LMs, but for *LMoracle* this effect was smaller than for *LMnormal*, which is expected since *LMoracle* should be able to compensate for infrequent words as it was trained on all the words in our corpus. Words that occur in phrase-initial position were more often misrecognised than words in other positions of the prosodic phrase; Goldwater et al. found the same effect for turn-initial words. Prosodically prominent words could be recognised better than deaccented words with both LMs. Though we did not find significant effects for the different pitch contours, we saw the largest difference in recognition between the two LMs for late peaks and late valleys.

Among the miscrecognised deaccented words, a large portion stemmed from word types which frequently become homophones in conversational speech, as they are reduced to a single phone. The confusion of such homophones that occur in similar syntactic positions could not be resolved with the current AMs, nor with either of the LMs. In a pilot study [30], we obtained first promising results in classifying homophoneous word types reduced to [aː] on the basis of their acoustic prosodic features using Random Forests. In the future, we aim at expanding those prosody-based classification experiments to other types of homophones and integrating this knowledge into our LM.

Table 2: *Difference in duration of (mis-)recognised words for the two LMs ($\bar{d}_{\mathrm{LMnormal},i} - \bar{d}_{\mathrm{LMoracle},i}$, where i indicates the different pitch categories); WER with LMnormal and WER decrease with LMoracle.*

Pitch contour category		Recogn. correct	Mean duration *LMnorm. – LMor.*	WER w. *LMnorm.*	WER decr. with *LMor.*
peak	early	true	−12.01ms	43.41%	20.93%
		false	22.64ms		
	medial	true	−36.01ms	49.32%	23.13%
		false	57.58ms		
	late	true	−49.79ms	50.92%	31.19%
		false	100.17ms		
flat		true	−7.30ms	46.59%	21.00%
		false	25.93ms		
valley	early	true	−13.07ms	48.84%	26.74%
		false	24.70ms		
	late	true	−42.70ms	50.52%	30.15%
		false	70.30ms		

6. Acknowledgements

S. Wepner was funded by grant P-32700-NB, and B. Schuppler by grant V-638-N33, both from the Austrian Science Fund (FWF). We thank the annotators David Ertl, Katerina Petrevska, Nina Richter and Nikolaus Tlapak for their efforts.

7. References

[1] T. Baumann, C. Kennington, J. Hough, and D. Schlangen, "Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there," in *Proceedings of IWSDS*, 2016, pp. 1–12.

[2] K. Li, H. Xu, Y. Wang, D. Povey, and S. Khudanpur, "Recurrent neural network language model adaptation for conversational speech recognition," in *Proc. INTERSPEECH*, 2018, pp. 3373–3377.

[3] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[4] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei *et al.*, "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1729–1744, 2006.

[5] G. Evermann, H. Chan, M. Gales, B. Jia, X. Liu, D. Mrva, K. Sim, L. Wang, and P. Woodland, "Development of the 2004 CU-HTK english CTS systems using more than two thousand hours of data," 2004.

[6] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech understanding: An overview of recent research at SRI," in *Proceedings of ISCA Workshop on Speech Recognition and Understanding*, 2001, pp. 13–16.

[7] M. Ostendorf, I. Shaffran, and R. Bates, "Prosody models for conversational speech recognition," in *Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, 2003, pp. 147–154.

[8] K. Chen, M. Hasegawa-Johnson, A. Cohen, s. Borys, S. S. Kim, J. Cole, and J. Y. Choi, "Prosody dependent speech recognition on radio news corpus of American English," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 232–245, 2006.

[9] J.-T. Huang, P.-S. Huang, Y. Mo, M. Hasegawa-Johnson, and J. Cole, "Prosody-dependent acoustic modeling using variable-parameter hidden Markov models," in *Proceedings of Speech Prosody*, 2010.

[10] K. Chen, S. Borys, M. Hasegawa-Johnson, and J. Cole, "Prosody dependent speech recognition with explicit duration modelling at intonational phrase boundaries," in *Proceedings of Eurospeech*, 2003, pp. 393–396.

[11] R. A. Bates, "Speaker dynamics as a source of pronunciation variability for continuous speech recognition models," Ph.D. dissertation, University of Washington, 2003.

[12] A. Rosenberg, "Using prominence and phrasing predictions to improve weighted dictionary pronunciation models," in *Proceedings of Interspeech*, 2012, pp. 2410–2413.

[13] S. Huang and S. Renals, "Using prosodic features in language models for meetings," in *Lecture Notes in Computer Science. Proceedings of the 4th international conference on Machine Learning for Multimodal Interaction*. Springer Berlin Heidelberg, 2007, pp. 192–203.

[14] O. Chan, "Prosodic features for a maximum entropy language model," Ph.D. dissertation, The University of Western Australia, Perth, July 2008.

[15] J.-T. Chien and C.-H. Chueh, "Joint acoustic and language modeling for speech recognition," *Speech Communication*, vol. 52, pp. 223–235, 2010.

[16] Y. Su and F. Jelinek, "Exploiting prosodic breaks in language modeling with Random Forests," in *Proceedings of ICPhS*, 2008, pp. 91–94.

[17] N. G. Ward, A. Vega, and T. Baumann, "Prosodic and temporal features for language modeling for dialog," *Speech Communication*, vol. 54, no. 2, pp. 161–174, 2012.

[18] B. Schuppler and T. Schrank, "On the use of acoustic features for automatic disambiguation of homophones in spontaneous German," *Computer, Speech & Language*, vol. 52, pp. 209–224, 2018.

[19] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: the Graz corpus of Read And Spontaneous Speech," in *Proceedings of LREC*, 2014, pp. 1465–1470.

[20] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Accepted for publication in Speech Communication*, Accepted.

[21] K. J. Kohler, *Paradigms in experimental prosodic analysis: From measurement to function*. De Gruyter, 2006, pp. 123–152.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[23] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[24] B. Schuppler, M. Adda-Decker, and J. A. Morales-Cordovilla, "Pronunciation variation in read and conversational Austrian German," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[25] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[26] P. Mowlaee and R. Saeidi, "Time-frequency constraints for phase estimation in single-channel speech enhancement," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 337–341.

[27] 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.

[28] R. H. Baayen, *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge University Press, 2008.

[29] N. Levshina, *How to do Linguistics with R. Data exploration and statistical analysis*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015.

[30] X. Kogler, "Classification of homophones in conversational Austrian German via Random Forest," Bachelor Thesis, Graz University of Technology, 2021.