



Homophone Disambiguation Profits from Durational Information

Barbara Schuppler¹, Emil Berger¹, Xenia Kogler¹, Franz Pernkopf¹

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

b.schuppler@tugraz.at, berger@student.tugraz.at, xenia.kogler@student.tugraz.at,
pernkopf@tugraz.at

Abstract

Given the high degree of segmental reduction in conversational speech, a large number of words become homophoneous that in read speech are not. For instance, the tokens considered in this study *ah*, *ach*, *auch*, *eine* and *er* may all be reduced to [a] in conversational Austrian German. Homophones pose a serious problem for automatic speech recognition (ASR), where homophone disambiguation is typically solved using lexical context. In contrast, we propose two approaches to disambiguate homophones on the basis of prosodic and spectral features. First, we build a Random Forest classifier with a large set of acoustic features, which reaches good performance given the small data size, and allows us to gain insight into how these homophones are distinct with respect to phonetic detail. Since for the extraction of the features annotations are required, this approach would not be practical for the integration into an ASR system. We thus explored a second, convolutional neural network (CNN) based approach. The performance of this approach is on par with the one based on Random Forest, and the results indicate a high potential of this approach to facilitate homophone disambiguation when combined with a stochastic language model as part of an ASR system.

Index Terms: homophone disambiguation, prosodic features, Random Forest, CNN, conversational speech, Austrian German

1. Introduction

Homophones are words that have the same canonical pronunciation but different meanings. How many homophones there are in a corpus is speaking style dependent. Given the high degree of segmental reduction in conversational speech, a large number of words become homophoneous that in read speech are typically realized with different canonical pronunciations. Whereas in face-to-face conversation, homophones do not tend to cause misunderstandings, they pose a serious problem for automatic speech recognition (ASR) of conversational speech. Machines classify homophones using the semantic context and the position of the token in the phrase. In spontaneous conversations, however, speakers frequently produce utterances that are ungrammatical, include repetitions or have incomplete syntactic structures, which makes homophone disambiguation from lexical context less reliable for conversational than for less spontaneous speaking styles. The aims of this study are twofold: (1) to analyze which prosodic and spectral cues distinguish homophones resulting from segmental reduction in conversational Austrian German and (2) to investigate methods for automatic homophone disambiguation based on acoustic information that are suitable for the integration into an ASR system.

From studies in the field of human speech recognition, we know that humans use various contextual indications to make ambiguous input disambiguous. Syntactic structure, gestures or statistical aspects of linguistic context can be used to differentiate homophones [1]. In addition, studies from the field of

phonetics have shown that homophones, despite being per definition produced with the same sequence of phone-segments, differ with respect to their prosodic and phonetic detail. This has been reported for homophones of words with different lexical meaning [2], with different communicative functions [3], and for pronouns of different grammatical functions [4, 5]. Importantly, it has been found that phonetic detail is also used by listeners to distinguish between words that become homophoneous due to reduction processes (e.g., [6]).

The question arises whether prosodic and phonetic detail can also be used for automatic homophone classification. Nemoto et al. [7] achieved classification accuracies of around 70% for the binary task of distinguishing the French words *et* vs. *est*, using, among others, Bayesian classifiers and SVMs with a set of 41 prosodic features. By means of a Random Forest classification using a large set of 193 acoustic and lexical features, Schuppler et al. [5] achieved a performance of 74% F_1 (92% accuracy) for distinguishing the three different grammatical functions *article*, *relative pronoun* and *demonstrative pronoun* of homophoneous tokens. What both studies have in common is that the classification methods rely on acoustic feature sets that are partly derived from manual annotations and require prior segmentation of the word and its context.

In this paper, we investigate two different approaches to automatically distinguish five German tokens *ah* 'eh', *ach* 'oh', *auch* 'also', *eine* 'a' and *er* 'he', which in conversational Austrian German all tend to be reduced to [a]. First, we use a Random Forest together with a large set of features related to the prosody and phonetic detail of the target word and of its context. Random Forests deal well with correlating features, high-order interactions between them, and with a combination of a small sample size and a large number of features. For these reasons, in the last decade Random Forests have also been used for studies in the fields of phonetics and linguistics (e.g., [8, 9]). The Random Forest based experiment gives us information about how the homophoneous tokens differ with respect to phonetic detail, where the durational features of the target word and its context show to be highly important. This Random Forest based approach, however, would not be suitable for the integration into an ASR system, as manual orthographic annotations are required for the segmentation and the subsequent extraction of the features. We therefore explore a second approach, where the spectrogram and the prosody of the token are treated as a signal, and classification is done via a convolutional neural network (CNN). CNNs have recently reached high popularity in the speech technology community for various applications, for instance in the field of speech separation [10, 11, 12], keyword spotting [13] and ASR in general [14, 15]. Given the challenge of a small dataset, we apply *MixUp Augmentation* [16] and investigate whether CNN performance improves when combining the spectrogram with durational information, as this resulted to be of high impact for the Random Forest classification.

2. Materials

All experiments of this study are based on the conversational speech component of the Graz corpus of Read and Spontaneous Speech (GRASS) [17, 18]. One hour long conversations between two speakers were recorded in a sound-proof room, resulting in 19 hours of high quality recordings of speech from 38 speakers (19f, 19m) from eastern Austria. The speaker pairs were mixed or gender-homogeneous and were familiar with each other (friends, colleagues, etc). The conversations were not supervised by an experimenter and the speakers were not given any instruction about the topic of the conversation, resulting in spontaneous, casual conversations. The recordings were subsequently manually transcribed orthographically, containing a total of 198,129 word tokens from 13,231 different word types (i.e, lexicon entries).

Using these manually created orthographic transcriptions, the corpus was automatically segmented into words and phones by means of a Forced Alignment with a Kaldi-based ASR system [19]. For this purpose, a pronunciation lexicon with pronunciation variants was used. These comprised automatically created variants by applying 36 phonological and reduction rules to the canonical pronunciation of the words and additional manually created variants, reflecting typical Austrian pronunciation for specific word types [20, 21]. Based on these word-level and phone-level segmentation, a total of 4175 tokens pronounced as single segment [a] were extracted, stemming from the word types *ach* 'oh' (222), *ah* 'eh' (765), *auch* 'also' (1408), *eine* 'a' (580), and *er* 'he' (1200).

3. Which acoustic features contribute most to disambiguate homophones?

3.1. Acoustic features

We extracted a total of 128 features related to fundamental frequency (F0), intensity (RMS), duration and spectral characteristics. F0 was extracted using *pYAAPT* (Yet Another Algorithm for Pitch Tracking)[22], for the calculation of the peaks the package *detecta* [23] was used. Besides calculating the minimum, mean, etc. of the F0, we also included positional F0 features (e.g., relative position of the F0 onset). The same calculations were made from the RMS trajectory, as earlier implemented by [24]. The 13 durational features included (among others) the duration of a preceding silence (if present, the duration of the target word itself, the local speaking rate (lspr) (i.e., the number of the segments per second), and the total speaking rate (tspr) (i.e., the number of segments produced in the entire utterance divided by its duration). The spectral features were all extracted using *Parselmouth* [25] and included the means and medians of the formants F1-F3, cepstral peak prominence (CPP), the relative strengths of the first (H1) and second harmonic (H1-H2), jitter and shimmer.

3.2. Random Forest

In order to analyze which acoustic features contribute to the distinction of the five word types *ach* 'oh', *ah* 'eh', *auch* 'also', *eine* 'a', and *er* 'he', we used scikit-learn [26] to build a Random Forest (RF) classifier with 500 estimators, a maximum depth of 18, a minimum samples split of 7, the square root as maximum number of features considered for splitting a node and Gini impurity measure for both classification and determining feature importance. The dataset was split into 70% for training, 20% for development, and 10% for evaluation.

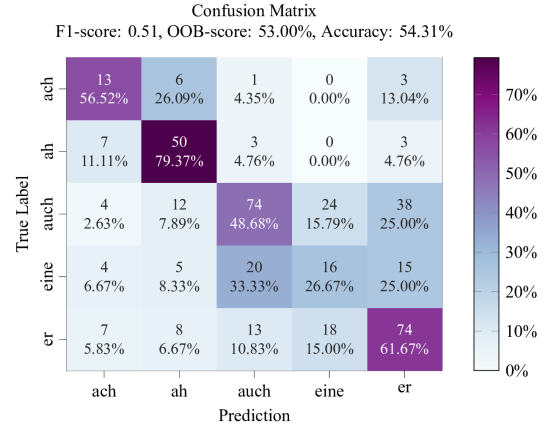


Figure 1: Confusion matrix for the Random Forest Classifier using the reduced feature set.

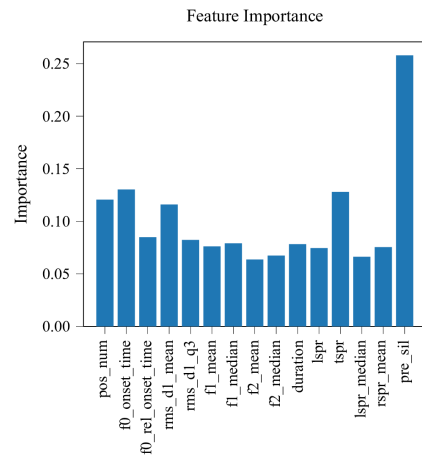


Figure 2: Feature importance as given by Shapley values for the Random Forest Classifier using the reduced feature set.

The overall importance of each feature was estimated using the package SHAP [27]. Shapley values are calculated by taking the average of each feature's contribution to each coalition of features., where high Shapley values indicate high feature importance [28]. Importances also take interactions and collinearities into account and allow for interpretations of the importances of the features relative to each other.

3.3. Results and Discussion

We first trained the RF with 128 features, achieving an overall accuracy of 57% and an F1-score of 0.55. On the basis of the Shapley values, we determined which features contributed most to the classification, and used all features (15) that reached a feature importance of > 0.025 for the subsequent retraining of the classifier. The classifier with the reduced feature set achieved an overall accuracy of 54.31% and an F1-score of 0.51. Figure 1 shows the respective confusion matrix. Reducing the features to only 12% of the original feature set avoids overfitting and still produced reasonably good results. The precision, recall and F1 score consistently is best for *ah*, by far lowest for *eine* and relatively low for *auch*. The reasons for why this is so can be found from analyzing the feature importances.

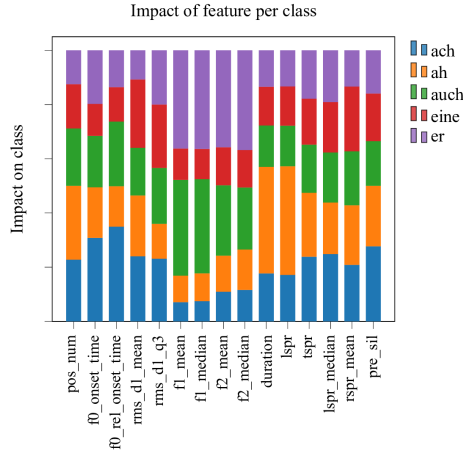


Figure 3: *Impact of the most important features for each class for the Random Forest Classifier using the reduced feature set.*

Figure 2 shows the remaining 15 features and their importance. The duration of the preceding silence reached the by far strongest impact on disambiguating the homophones considered here, followed by other durational features (total, local and relative speaking rate), two F0 and two RMS related features, and the mean and median of the first and second formant. In relation to the whole set of 128 features, relatively few F0 and RMS related features remained, and the durational features showed to be highly important for the task. None of the features related to voice quality (e.g. CPP) ranked upon the 15 most important features, indicating that these word types are equally weakly or strongly modal. Given these feature impacts, it is thus not surprising, when looking at the confusion matrix (Figure 1), that *eine* is mostly confused with those word types that also primarily occur in phrase-medial position (i.e., *auch* and *er*), where a preceding silence is just as unlikely as for *eine*. In line with this observation, the two word types *ah* and *ach* that are both likely to occur in phrase-initial position, are also most often confused with each other.

Figure 3 shows the impact of the features on the prediction of the classes. The F1 and F2 related features have the highest impact on the classes *er* (with the canonical pronunciation [ɛʁ]) and *auch* (canonical:[aʊχ]) compared to the other classes. This indicates that traces of the diphthong might still be present for these tokens, despite being classified with the Kaldi-based forced alignment as monophthong [a]. The durational features had the highest impact on the prediction of the class *ah*, whereas the F0-features were most important for class *ach*.

Having analyzed which features have high impact on the disambiguation of the five homophones *ach*, *ah*, *auch*, *eine* and *er* in Austrian German, rises the question of how well these findings generalize to other types of homophones in German, and what is more, in other languages. For spontaneous Northern German, Schuppler and Schrank [5] found that durational and F0 related features played a mayor role, similarly as we report here for Austrian German. A detailed analysis on the impact of specific features, however, was not provided in [5]. For spontaneous French, Nemoto et al. [7] analyzed ASR errors caused by the homophones *et* and *est* and found that the most important acoustic cues to distinguish these words were phoneme duration, the difference between inter-phonemic duration (which correlates to our measure of relative speechrate), F0

and the pause preceding the target word. These findings from French are highly similar to ours and suggest that the use of these durational and prosodic features may also help to distinguish homophones in other (at least Germanic and Romance) languages.

4. Does homophone disambiguation with CNNs profit from information on the duration of the preceding silence?

What the approaches mentioned in the previous section (those from literature, and our own) have in common is that they require the extraction of acoustic features which rely on the segmentation of the whole phrase. As such an approach is not well incorporable into an ASR system, we conducted a second experiment based directly on the spectrogram of the recorded audio. Since the duration of the preceding silence as well as the duration of the token resulted to be crucial for homophone disambiguation, we investigate whether an encoding of this durational information improves CNN-based classification performance.

4.1. Data Preprocessing

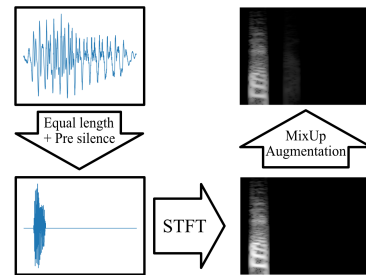


Figure 4: *Preprocessing of recorded words.*

We extracted log-magnitude mel spectrograms using the Short-time Fourier Transform (STFT). We used a Hanning window of size 4096. The sampling rate was 48kHz. The selected hop size of 256 provides a sufficient time-resolution. The information about the preceding silence was incorporated to the spectrogram by adding silence in the respective length in front of the audio signal¹. In order to account for the different durations of the tokens, we additionally performed zero padding of the spectrogram, resulting in tokens of equal size (see Figure 4). As shown in Figure 5(A), the token count was very unbalanced. Therefore, *MixUp Augmentation* was used to oversample under-represented classes. The idea is to mix up the features and labels of a sample with a randomly chosen sample from another class [16]:

$$x_{mixed} = (1 - \lambda) \cdot x_{original} + \lambda \cdot x_{random}$$

$$y_{mixed} = (1 - \lambda) \cdot y_{original} + \lambda \cdot y_{random},$$

where x represents the features of a sample, y denotes the corresponding one-hot-encoded label and λ is the mixing weight. For our experiments, a λ of 0.2 worked best. Figure 5(B) shows the increase of samples for each token. Due to the oversampling, the ratio of training, development and evaluation data changed to 79.86%, 12.99% and 7.15% respectively. The evaluation data set had the same size as for the RF experiment.

¹When integrated into an ASR system, the duration of the silence could be retrieved from a speech-silence detector.

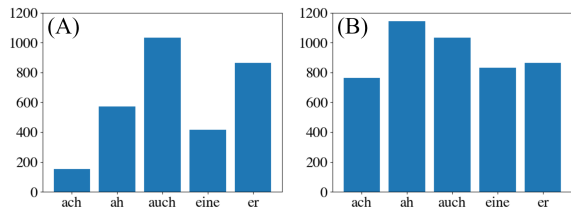


Figure 5: (A) Original token distribution; (B) Distribution of oversampled training data.

4.2. CNN Model

Figure 6 shows the structure of the CNN model. The first layer performs a 2D-convolution with a kernel of size 3x3 on the spectrogram. Then max-pooling of size 3x3, a ReLU activation, batch-normalization and dropout at a value of 0.17 is performed. This structure of 2D-convolutions, pooling, ReLU activations and batch-normalization is replicated. After the final dropout (0.15), a fully connected layer with a softmax activation function provides the classification results. As data is limited, a small-scale CNN is necessary. In total, we have 13589 parameters. We trained the network using the Adam optimizer [29] with a batch size of 2048 for 150 epochs. Furthermore, weight decay regularization was used.

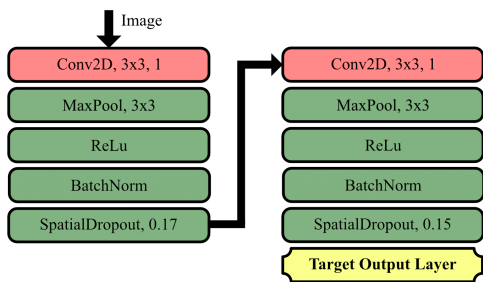


Figure 6: CNN Architecture.

4.3. Results and Discussion

Class	CNN	CNN
	with pre-silence	without pre-silence
ach	0.41	0.33
ah	0.63	0.46
auch	0.61	0.56
eine	0.33	0.23
er	0.61	0.55
overall accuracy	57 %	48 %

Table 1: F1-scores and overall accuracy for the different classes with a CNN, trained with and without pre-silence.

Table 1 shows the F1-scores for each word and the overall accuracy of the CNN, where performance with preceding silence was better in every class, and overall better by 0.09 than without pre-silence. This result supports the findings from the RF experiment with respect to the importance of the preceding silence for the disambiguation of the considered homophones. For the CNN containing preceding silence, we achieved an overall accuracy of 57 %, with the best class *ah* reaching an

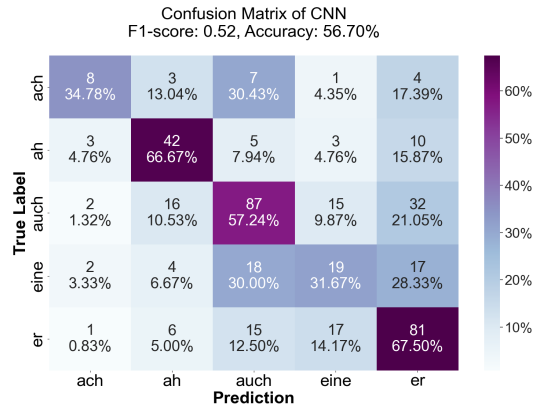


Figure 7: Confusion matrix of the CNN with preceding silence.

F1-score of 0.63 and the worst class *eine* at a score of 0.33. Figure 7 shows the confusion matrix for the CNN using preceding silence. Compared to the prediction accuracy of the RF (see Figure 1), the CNN shows similar confusion patterns across all classes. From the RF analysis, we have learnt that whereas the classes *ach* and *ah* are mainly distinguished by durational and prosodic features, for the classes *auch* and *er* also F1 and F2 play a significant role. The fact that with the CNN approach, the accuracies for *ach* and *ah* are lower than with the RF, reflects that durational information is captured only to a limited extend compared to the large set of acoustic features used for the RF. In line with this observation, for those classes where spectral information is of high impact (i.e., *auch* and *er*), the accuracies with the CNN are higher than those of the RF.

5. Conclusions

The aims of this paper were (1) to analyze which prosodic and spectral features distinguish homophones resulting from segmental reduction in conversational Austrian German, and (2) to investigate methods for automatic homophone disambiguation that are suitable for the integration into an ASR system. In our analysis based on a set of 128 features and a Random Forest classifier, we found that the overall most important features are the duration of a silence preceding the target word as well as its durational aspects (local, global and relative speechrate).

In a second experiment, we showed that information on pause and token duration is also beneficial for a CNN based approach towards homophone disambiguation. Classification performance with pre-silence was 0.09 F1-score better than without pre-silence. In future, we plan to expand our study to other types of homophones resulting from segmental reduction in conversational speech. For instance, a large number of function words are reduced to [s] and [tə] in Austrian German, causing enormous WERs. Since our results indicate a high potential of the CNN based approach to facilitate homophone disambiguation, we plan to combine this approach with a stochastic language model as part of the Kaldi-based ASR system for conversational Austrian German currently under development at our department [30, 31].

6. Acknowledgements

B. Schuppler was partly funded by grant V-638-N33 from the Austrian Science Fund (FWF). We thank Saskia Wepner for her support with the Kaldi-based forced alignment of GRASS.

7. References

- [1] S. Trott and B. Bergen, “Why do human languages have homophones?” *Cognition*, vol. 205, p. 104449, 2020.
- [2] S. Gahl, ““time” and “thyme” are not homophones: the effect of lemma frequency on word duration in spontaneous speech,” *Language*, vol. 84, no. 3, pp. 474–496, 2008.
- [3] J. Volín, L. Weingartová, and O. Niebuhr, “Between recognition and resignation – the prosodic forms and communicative functions of the Czech confirmation tag “jasně,”” in *Proceedings of Speech Prosody 7*, 2014, pp. 115–119.
- [4] B. Samlowski, P. Wagner, and B. Möbius, “Effects of lexical class and lemma frequency on German homographs,” in *Proceedings of INTERSPEECH*, 2013, pp. 597–601.
- [5] 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, 01 2018.
- [6] O. Niebuhr and K. J. Kohler, “Perception of phonetic detail in the identification of highly reduced words,” *Journal of Phonetics*, vol. 39, pp. 319–329, 2011.
- [7] R. Nemoto, I. Vasilescu, and M. Adda-Decker, “Speech errors on frequently observed homophones in French: Perceptual evaluation vs. automatic classification,” in *Proceedings of LREC’08*, 05 2008, pp. 2189 – 2195.
- [8] S. Tagliamonte and R. Baayen, “Models, forests and trees of York English: Was/were variation as a case study for statistical practice,” *Language Variation and Change*, vol. 24, no. 2, pp. 132–178, 2012.
- [9] S. Baumann and B. Winter, “What makes a word prominent? predicting untrained german listeners’ perceptual judgments,” *Journal of Phonetics*, vol. 70, pp. 20–38, 2018.
- [10] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7092–7096.
- [11] M. Zöhrer, R. Peharz, and F. Pernkopf, “Representation learning for single-channel source separation and bandwidth extension,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2398–2409, 2015.
- [12] D. S. Williamson and D. Wang, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.
- [13] D. Peter, W. Roth, and F. Pernkopf, “Resource-efficient dnns for keyword spotting using neural architecture search and quantization,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 9273–9279.
- [14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [15] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.
- [16] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: <http://arxiv.org/abs/1710.09412>
- [17] 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.
- [18] B. Schuppler, M. Hagmüller, and A. Zahrer, “A corpus of read and conversational Austrian German,” *Speech Communication*, vol. 94, pp. 62–74, 09 2017.
- [19] S. Wasserfall, *Automatic speech segmentation using Kaldi*. Master Thesis from Graz University of Technology, 2020.
- [20] B. Schuppler, M. Adda-Decker, and J. A. Morales-Cordovilla, “Pronunciation variation in read and conversational Austrian German,” in *Proceedings of INTERSPEECH*, 2014, pp. 1453–1457.
- [21] B. Schuppler, S. Grill, A. Menrath, and J. A. Morales-Cordovilla, “Automatic phonetic transcription in two steps: forced alignment and burst detection,” in *Statistical Language and Speech Processing. SLSP 2014. Lecture Notes in Artificial Intelligence*, L. Besacier, A. Dediu, and C. Martín-Vide, Eds. Springer, 2014, vol. 8791, pp. 132–143.
- [22] B. J. B. Schmitt, “AMFM decomp documentation 1.0.11,” in *Version 1.0.11*, retrieved 14 April 2021. http://bjbschmitt.github.io/AMFM_decomp/index.html, 2021.
- [23] M. Duarte, “detecta: A Python module to detect events in data,” *Zenodo*, vol. v0.0.5, 2021.
- [24] M. A. Dabrowski, *Prosodische Prominenz - Berechnung von akustischen Merkmalen zur Erkennung von prosodischer Prominenz in gesprochener Sprache*. Bachelor Thesis from Graz University of Technology, 2020, Graz, Austria.
- [25] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [27] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- [28] P. Rathi. A novel approach to feature importance — shapley additive explanations. [Online]. Available: <https://towardsdatascience.com/a-novel-approach-to-feature-importance-shapley-additive-explanations-d18af30fc21b>
- [29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR (Poster)*, 2015. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [30] S. Wepner, B. Schuppler, and G. Kubin, “How prosody affects ASR performance in conversational Austrian German,” in *Proceedings of Speech Prosody 2022*, 2022, pp. 195 –199.
- [31] J. Linke, P. N. Garner, G. Kubin, and B. Schuppler, “Conversational speech recognition needs data? Experiments with Austrian German,” in *Proceedings of LREC*, 2022, pp. 4684 — 4691.