10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Towards automatic annotation of prosodic prominence levels in Austrian German

Julian Linke[1], *Anneliese Kelterer*[2], *Markus A. Dabrowski*[1], *Dina El Zarka*[2], *Barbara Schuppler*[1]

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

linke@tugraz.at, m.dabrowski@student.tugraz.at, anneliese.kelterer@uni-graz.at, dina.elzarka@uni-graz.at, b.schuppler@tugraz.at

Abstract

The creation of prosodic annotations is one of the most difficult and time-consuming aspects of creating a speech database. Generally, only the speech signal and manually created transcriptions are available in an early resource development stage. This paper presents a tool for annotating prosodic prominence at the word level, using exclusively acoustic features (96 f0-, intensity- and durational features). The best performance for separating prominent from non-prominent words in Austrian read speech was reached with a decision tree with the absolute word duration as the only feature. For distinguishing more prominence levels, a good performance was reached with a random forest model, similar to the best inter-annotator agreement. Furthermore, we analyzed in detail the feature ranking of the random forest to give us insights into the relative importance of the features contributing to prominence in Austrian German: Word duration $>$ f0 range, RMS range. The specific findings of this study will mainly be relevant for speech scientists and prosody researchers interested in German. Our methodological approach of analyzing prosodic prominence from a purely acoustic perspective at the word-level will also be interesting for researchers focusing on prosody in other languages.

Index Terms: prosodic prominence, prosodic features, Austrian German, decision trees, random forests

1. Introduction

Prosodic prominence is a complex phenomenon that can be studied from different perspectives [1]. Some linguistic unit is usually considered prosodically prominent if it is perceived as standing out from its environment [2]. However, what is perceived as prominent is influenced by a multitude of factors involving functional, structural and frequency criteria, and the rating task as much as the physical properties of the item that is perceived as being prominent [3, 4, 5, 6, 7]. In this paper, we focus on the correlates of prominence in the acoustic signal.

The aims of this paper are two-fold: The first aim is to develop a tool that facilitates the ongoing prosodic annotation process of GRASS, the first large scale database for Austrian German, containing both read and conversational speech [8]. So far, only a small portion of the speech material has been annotated manually. With the help of the tool presented here, the rest of the corpus will be annotated using a semi-automatic approach. The second aim is to establish the extent to which the acoustic features used by our models contribute to the perception of prosodic prominence in Austrian German (as given by the perception-based annotations).

1.1. Acoustic cues and perception of prosodic prominence

Different languages rely on different weightings of acoustic cues to create perceptual prominence [9], the most important of which are f0 variation, duration and intensity (e.g., [10, 11, 12, 13, 14]). With respect to f0, not only pitch height or excursion are relevant for prominence, but also the shape and alignment of pitch contours [15, 16, 14]. For English, Cole et al. [6] found that duration is a more important cue for prominence perception than intensity/loudness (e.g., [10, 17]).

For German, [18] and [19] reported that f0 was a more important correlate of prosodic prominence than syllable duration. Pitch-accent related variables outranked both acoustic f0 and duration features in [20]. On the other hand, [17] argued that word duration was more important for prominence perception than f0 and intensity, and [21] found that force-accent related parameters (i.e., duration and spectral emphasis) were more important for syllable prominence than pitch-accent related parameters (i.e., f0-curve features and overall intensity). For Austrian German, it has been found that duration and spectral tilt are strong acoustic cues to perceptual prominence, whereas a change in f0 within a syllable did not necessarily correlate with stronger perceptual prominence [22]. Concerning vowel quality, higher prominence only seems to affect F1, but not F2 and F3 [22].

1.2. Automatic prosodic annotation tools

Several automatic prosodic annotation tools have been built and distributed. Some of them combine acoustic, lexical and syntactic features (e.g., [23] for American English, [24, 25] for French), others use lexical and syntactic information alone (e.g., [26] for Dutch). Arnold et al. [18] used GAMs and random forests to model prosodic prominence in German, with the aim of analyzing and comparing the contribution of acoustic, linguistic and contextual information. Like [18], we aim at using random forest models to learn more about the contribution of the features to prosodic prominence perception. Since we additionally aim at building a tool that can be incorporated into the annotation process of a not-yet annotated database, a requirement for the tool is the use of acoustic features alone.

For American English, Tamburini and Caini [27] proposed a tool that classifies whether a syllable is prominent or not. The prediction was based on the speech waveforms only, with no higher level linguistic information available to the tool. For German, only a few prosodic annotation tools have been built that use acoustic features alone. N. Braunschweiler [28], for instance, proposed ProsAlign, a system that automatically produces GToBI labels. The tool covers 56% of the manually established labels and can thus be integrated in a semi-automatic annotation procedure. Since the development of ProsAlign,

10.21437/SpeechProsody.2020-204

however, other prosodic annotation systems than GToBI have been developed for German (e.g., KIM, DIMA) [29]. The tool by Tamburini and Wagner [21] annotates prominence as a continuous, rather than a categorical parameter. Their analysis led to the conclusion that force accents are a more reliable cue to prominence than pitch accents in German. For Austrian German, no tool is available at this point.

2. Materials and Methods

2.1. GRASS corpus

This study is based on read speech from GRASS, the *Graz Corpus of Read and Spontaneous Speech* [8, 30]. GRASS comes with automatically created segmentations using MAUS [31], which were corrected manually. GRASS was manually annotated prosodically, using the same criteria as the Kiel corpus [32], with prominence levels 0 (no prominence), 1 (weak prominence), 2 (strong prominence) and 3 (emphatic prominence). Three phonetically trained transcribers created the prosodic annotations in the following way: one transcriber created the first version of the annotation, which was subsequently corrected by the other transcribers. This procedure reached a high inter-annotator agreement (Cohen's kappa: 0.81, 0.76, 0.63, calculated on 269 word tokens from 47 utterances). Prominence classification in Section 3 is based on a training set of 197 utterances (2919 word tokens) from two narratives, read by 10 male and 9 female speakers. The test set for the classification experiments consists of 47 utterances (269 word tokens) annotated by all three annotators. Prominence ratings for the test set were assigned by majority decision.

2.2. Acoustic feature extraction

For each word, we extracted 96 features based on the fundamental frequency f0, the sound intensity (RMS) and durational characteristics. f0 was calculated with the library *AMFM decompy* [33]. This package contains an implementation of the pitch detection algorithm *YAAPT* [34]. Sound intensity was calculated directly from the waveform by calculating the root mean square. For f0 and RMS, and their respective first and second derivatives, 10 measurements were extracted: maximum, minimum, range, relative position of maximum and minimum in the word, mean, median, first and third quartile and standard deviation (60 features). For the basic f0 and RMS curves, we extracted 12 measurements: left and right slope of the maximum and minimum, absolute and relative onset and offset within the word, as well as maximum, minimum, range and mean relative to the utterance (24 features). We employed the peak detection algorithm [35] and *numpy* [36] for the statistial features. The 12 durational features were: word duration, total speech rate (phrase), local speech rate (word), and 9 relative speech rate measures. The local speech rate is estimated as the ratio of number of segments to word length.

2.3. Classification methods and testing

Two classification methods were implemented in Python with the *scikit learn* toolkit (version 0.21.3.) [37]. First, we modelled a decision tree (DT) with only two classes and the Gini Impurity (I_G) as an impurity measure. For the purpose of this classification task, prominence levels 1, 2 and 3 were combined in the class ***Pr*** (prominent, 1669 tokens). The associated null class is called ***NPr*** (non-prominent, 1250 tokens). This binary classification task has the aim to test how well prominence detection is

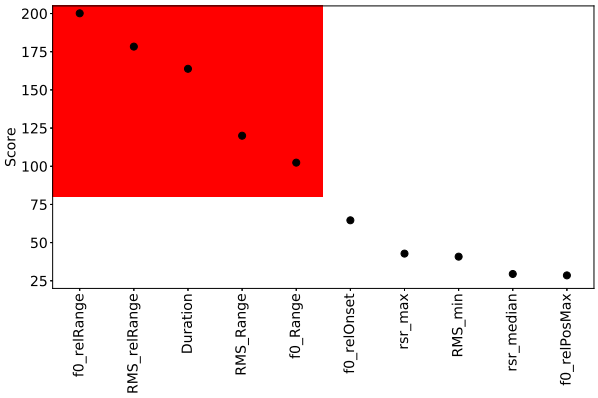

Figure 1: *10 highest χ^2-scores in case of a classification task with 3 classes. The selected features are located in the red marked area. A selection of the first 3 features did not result in a satisfying performance. Adding 2 more features indicated a more reliable classification performance, which can be compared to the ranking of a RF-model with all 97 features included (cf., Section 3).*

at all possible with the given data set compared to prior studies. Parameterization was done by comparing the results of different DT-models with varying tree depths fitted with the training set. A comparison of different DT-topologies showed that a depth of 1, which obviously leads to a highly simplified model, was sufficient to distinguish between the classes ***NPr*** and ***Pr***. Hence, including higher depths resulted in more complicated models with no improvement in the respective F1-scores.

Second, a random forest (RF) with 1000 decision trees was used for a classification task with three classes: ***0*** (no prominence, 1250 tokens), ***1*** (weak prominence, 726 tokens), and ***2*** (strong prominence, 943 tokens; prominence ratings 2 and 3 were subsumed; cf., Section 2.1). The impurity measure of each decision tree was the Gini Impurity and the depth of each decision tree was maximal, resulting in pure leaves or leaves with less than 2 samples. Feature selection based on χ^2-statistics of each attribute and comparisons of different feature sets (full set or sets with 1-20 features with best χ^2-scores) fitted to unique RF-models showed that a set of 5 features was sufficient to solve the classification task (Fig. 1). Other studies have shown that RF-models have good prediction quality and they can cope with a feature space including many highly correlated features [38]. Moreover, the feature importances of the RF-model provide a ranking of the respective features allowing both a better linguistic interpretability of the selected features (e.g., [18, 39]) and a link to the used χ^2 feature selection algorithm. In both classification experiments, a set containing 10% of each class of the training set was used for validation. Methods are evaluated by measuring the F1-score and by presenting the respective confusion matrices.

3. Results

3.1. Results of Decision tree (2 prominence classes)

In the binary classification task, the feature *duration* in the root node of a decision tree obtained a sufficient separability, which led to a highly simplified model to distinguish between 2 prominence levels (***Pr*** vs. ***NPr***). If the condition $duration \leq 0.25s$ was fulfilled, the observation was classified as ***NPr***. Confusion matrices of the validation and the test set (Fig. 2) showed

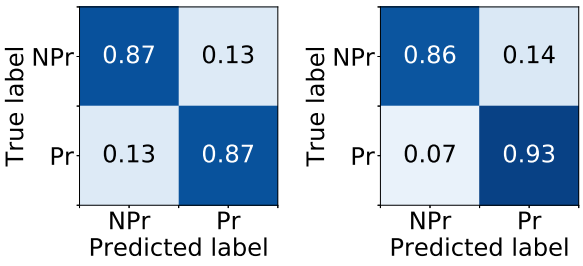

Figure 2: *Confusion matrices showing the respective Recalls of the 2 classes in the main diagonal for the validation set (left) and the test set (right).*

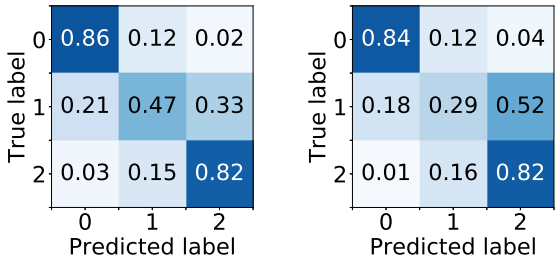

Figure 3: *Confusion matrices showing the respective Recalls of the 3 classes in the main diagonal for the validation set (left) and the test set (right).*

that non-prominent levels had similar Recalls. Prominence was recognized better in the test set (Recall $= 0.93$) and the corresponding F1-score was 0.92. In both sets, non-prominent words had a F1 > 0.85.

3.2. Results of Random Forest (3 prominence classes)

In the second classification task with 3 prominence levels (cf., Section 2.3), a set of 5 features (Fig. 1) was chosen. The prominence rating of the final RF-model (averaged impurity decrease of an ensemble of 1000 decision trees) showed that the feature *duration* was rated as the most important feature (Fig. 4), followed by the features referring to the f0 range (*f0_relRange* and *f0_range*) and the RMS range (*RMS_relRange* and *RMS_range*). Since the relative ranges of f0 and RMS were calculated by relating the f0 or RMS range of the word to the range of the respective utterance, there was a correlation between the features *f0_range* and *f0_relRange* ($r = 0.86$ (2909), $p < .0001$) and *RMS_range* and *RMS_relRange* ($r = 0.96$ (2899), $p < .0001$) [40].

Figure 3 shows the confusion matrices of the classification with three classes for the test set and validation set. When comparing the Recalls (corresponding to the main diagonal of the confusion matrices) of the two sets, similar results can be seen for classes ***0*** and ***2*** (Recall > 0.82 in both cases). However, for class ***1***, a Recall of only 0.47 was measured for the validation set, while 33% of class ***1*** was predicted as class ***2***. In the test set, 19% more observations of class ***1*** were predicted as class ***2*** in comparison to the validation set. This uncertainty of classifying prominence level ***1*** was also represented in the corresponding F1-scores of the validation set (F1 $= 0.51$) and the test set (F1 $= 0.34$). No prominence was recognized in both sets with F1 > 0.85. In contrast, F1-scores of class ***2*** were higher in the validation set (F1 $= 0.81$) than in the test set (F1 $= 0.74$).

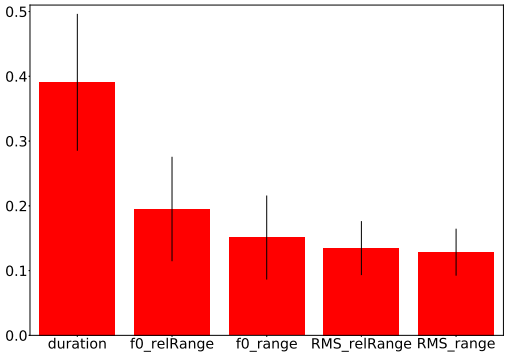

Figure 4: *Feature Ranking of the fitted RF-model corresponding to the averaged impurity decrease computed with 1000 decision trees. Black lines indicate the standard deviations referring to the impurity measurements of each tree of the forest.*

4. Discussion

4.1. Classification performance

The binary classification task indicates a good separability of prominence. A preliminary decision about word prominence could help in the annotation process by anticipating the distinction between no prominence and prominence, so the annotators can focus on a simpler manual binary decision task (between weak and strong prominence) instead of the more complex three-way decision task. Recalls of classes ***0*** and ***2*** in the second classification task with three prominence levels show very good results in both sets. Results of class ***1*** in the confusions matrices, however, indicate more variation in the production or an uncertainty in the annotation of weak prominence. One reason why the recognition performance of class ***1*** was poorer in the test than in the validation set could be that acoustic cues are weighted differently in the two data sets (cf., Section 4.2).

4.2. Contribution of acoustic features to classification and perception

In both classification methods, word duration was the most important feature for distinguishing prominence levels. Figure 4 shows that prominence was represented by the classical triad of prosodic features [41] in the RF: duration, two f0 features and two intensity features. Our experiment showed that word duration was a more important feature than f0 range, which was, in turn, more important than RMS range. These results are in line with the findings by [17]. Similarly, [21] found that force-accent parameters were more important for prominence in German than pitch-accent parameters. However, other studies of prominence in German found that f0 related parameters were more important [19, 18]. Due to methodological differences, the results of the different studies are however not fully comparable (cf., [17]). Many studies on word prominence use acoustic measures that relate to the stressed vowel or syllable (e.g., [6, 20]), while we investigated word-level features only. Thus, we cannot conclude from the importance of word duration in our data that force-accent parameters are more important than pitch-accent parameters (cf., [21]) as the stressed syllable is the constituent a force/pitch accent is associated with. Remarkably, the conclusion that duration was more important for perceived word prominence than f0 features was also drawn by one other study that investigated word prominence by measuring word-

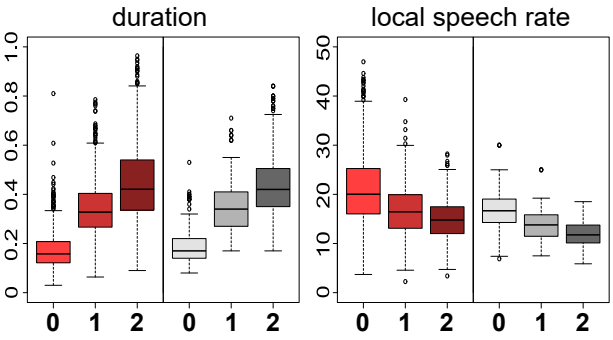

Figure 5: *Boxplots of prominence ratings 0, 1 and 2 of duration (left) and local speech rate (right), for the training set (red) and the test set (grey).*

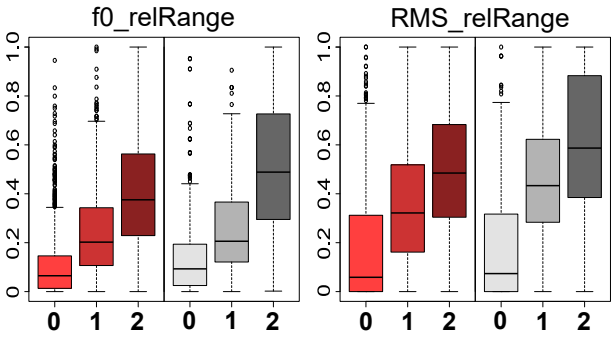

Figure 6: *Boxplots of prominence ratings 0, 1 and 2 of f0_relRange and RMS_relRange, for the training set (red) and the test set (grey).*

level acoustic features [17]. An advantage of our approach over studies measuring acoustic features in the stressed syllable is that it captures f0 excursions related to late peaks which are often realized outside the stressed syllable. In addition, word-level duration does not only capture whether the prominent syllable is shortened or lengthened, but also whether reductions (e.g., segment deletions) shorten non-prominent words as a whole.

The five features in the RF all had higher values the higher the prominence level was (Fig. 5 and 6). For prominence level ***0***, *duration* values are located in the lower range and are clearly distinct from those of ***1*** and ***2***, while there is more overlap between ***1*** and ***2*** (Fig. 5). Less scattering of ***1*** and ***2*** towards lower values in the test set (Fig. 5) also explains why the recall of ***Pr*** was higher in the test set than the validation set (Fig. 2). One reason for non-prominent words being shorter is that 94% of them were function words, which are generally shorter in terms of syllables as well as duration. Less prominent words also have a shorter duration because the speech rate in less prominent words is higher (Fig. 5). A mixed effects model [42] with local speech rate as dependent variable, prominence rating as independent variable and word as random variable showed that the local speech rate is significantly higher for class ***0*** than class ***1*** (Est. = -3.39, t = -11.39, $p < .001$) and class ***2*** (Est. = -5.18, t = 15.78, $p < .001$). Thus, words with the same number of phones are produced faster in less prominent position.

Since the range and the relative range of f0 and RMS are correlated and show similar distributions, only the relative ranges are discussed here. Figure 6 shows that *f0_relRange* increases with prominence in both data sets. The higher f0 excursion for class ***2*** in the test set could explain the better Recall of ***Pr*** in this set (Fig. 2). RMS relative range also increases with prominence in both data sets. However, for *RMS_relRange*, the distribution of class ***1*** in the test set is similar to the distribution of class ***2*** in the training set. This difference in *RMS_relRange* between the two data sets could explain why level ***1*** was classified more often as ***2*** in the test set (Fig. 3). This could be because the test set includes isolated short sentences (cf., Section 2.1), which in turn might result in a different reading behavior characterized by a different weighting of the acoustic features involved in expressing prominence.

5. Conclusions

The aim of this paper was to build an annotation tool for prosodic prominence with as little pre-processing effort as possible. Therefore, the models rely on acoustic features extracted from the word and its automatically created word and phone-level segmentations, as this information is usually available first in the resource development process. We thus chose not to use any other linguistic information such as the position of the prominent syllable in the word. Based on manual prominence annotations of a small part of GRASS, we explored the combinations of different acoustic feature sets and classification methods. Our results show that both classification methods can distinguish between prominent and non-prominent words. In a binary classification, 93% of all prominent words were classified as prominent in the test set. Results with three prominence levels indicate that strong prominence is recognized well (Recall $= 82\%$), but weak prominence tends to be confused with strong prominence.

In our analysis of the contribution of acoustic features to prosodic prominence in Austrian German, word duration was the most important feature, followed by f0 range and RMS range. These results are in line with one other study of word prominence in German, but deviate from most other studies in which f0 features ranked higher than other acoustic features. This discrepancy could be due to different methodologies, in particular concerning the domain of acoustic feature extraction (i.e., syllable vs. word).

The presented classifiers will be used in two ways in the future: (1) as part of a semi-automatic annotation process for the rest of the GRASS corpus to yield faster and more consistent annotations; (2) as part of a prosody-dependent ASR system. In such a system, prosodic prominence levels will be related to pronunciation in order to improve word recognition by priming the system for the relationship between prominence levels and the degree of segmental reduction. For these two purposes, the performances reached are sufficient. Whereas the specific findings of this study will mainly be relevant for speech scientists and prosody researchers interested in German, our methodological approach of analyzing prosodic prominence from a purely acoustic perspective at the word-level will also be interesting for researchers investigating the prosody of other languages.

6. Acknowledgements

The work by Anneliese Kelterer was funded by grant P-32700-N and the work by Markus A. Dabrowski and Barbara Schuppler by grant V-638-N33, both from the Austrian Science Fund. We would like to thank the transcribers David Ertl and Katerina Petrevska for their efforts.

7. References

[1] B. Wagner, A. Origlia, C. Avesani, G. Christodoulides, F. Cutugno, and M. D'Imperio et al., "Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proc. of ICPhS*, 2015.

[2] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *J. Acoust. Soc. Am*, vol. 89, no. 4, pp. 1768–1776, 1991.

[3] J. Cole, I. Hualde, C. Smith, T. Mahrt, and R. Napoleao de Souza, "Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish," *J. Phon*, vol. 75, pp. 113–147, 2019.

[4] R. Turnbull, A. Royer, K. Ito, and S. R. Speer, "Prominence perception is dependent on phonology, semantics, and awareness of discourse," *Lang. Cogn. Neurosci.*, vol. 32, no. 8, pp. 1017–1033, 2017.

[5] J. Bishop, "Information structural expectations in the perception of prosodic prominence," in *Prosody and Meaning*, G. Elordieta and P. Prieto, Eds. Berlin: Mouton de Gruyter, 2012, p. 239–270.

[6] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, no. 1, p. 425–452, 2010.

[7] P. Wagner, "Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates," in *Proc. of Interspeech*, 2005, pp. 2381–2384.

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

[9] M. E. Beckman, "Stress and non-stress accent," *Netherlands Phonetic Archive*, vol. 7, 1986.

[10] A. E. Turk and J. R. Sawusch, "The processing of duration and intensity cues to prominence," *J. Acoust. Soc. Am*, vol. 99, no. 6, pp. 3782–3790, 1996.

[11] J. Terken and D. J. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and Experiment, studies presented to Gösta Bruce*. Dordrecht: Kluwer Academic Publishers, 2000, pp. 89–127.

[12] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: fundamental frequency lends little," *J. Acoust. Soc. Am*, vol. 118, no. 2, pp. 1038–1038, 2005.

[13] H. Mixdorff, C. Cossio-Mercado, A. Hönemann, J. Gurlekian, D. Evin, and H. Torres, "Acoustic correlates of perceived syllable prominence in german," in *Proc. of Interspeech*, 2015, pp. 51–55.

[14] S. Baumann, O. Niebuhr, and B. Schroeter, "Acoustic cues to perceived prominence levels – evidence from German spontaneous speech," in *Proc. of Speech Prosody*, 2016.

[15] K. J. Kohler and R. Gartenberg, "The perception of accents: F0 peak height versus F0 peak position," *AIPUK*, vol. 25, pp. 219–294, 1991.

[16] O. Niebuhr, "On the phonetics of intensifying emphasis in german," *Phonetica*, vol. 67, pp. 1–29, 2010.

[17] D. Arnold, B. Möbius, and P. Wagner, "Comparing word and syllable prominence rated by naive listeners," in *Proc. of Interspeech*, 2012, pp. 1877–1880.

[18] D. Arnold, P. Wagner, and R. H. Baayen, "Using generalized additive models and random forests to model prosodic prominence in German," in *Proc. of Interspeech*, 2013, pp. 272–276.

[19] O. Niebuhr and J. Winkler, "The relative cueing power of f0 and duration in german prominence perception," in *Proc. of Interspeech*, 2017, pp. 611–615.

[20] S. Baumann and B. Winter, "What makes a word prominent? predicting untrained german listeners' perceptual judgements," *J. Phon*, vol. 70, pp. 20–38, 2018.

[21] F. Tamburini and P. Wagner, "On automatic prominence detection for German," in *Proc. of Interspeech*, 2007, pp. 1809–1812.

[22] D. El Zarka, B. Schuppler, C. Lozo, W. Eibler, and P. Wurzwallner, "Acoustic correlates of stress and accent in Standard Austrian German," in *Phonetik in und über Österreich, Veröffentlichungen zur Linguistk und Kommunikationsforschung: 31*, S. Moosmüller, C. Schmid, and M. Sellner, Eds. Vienna: ÖAW Austrian Academy of Sciences Press, 2017.

[23] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. Audio Speech Lang Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[24] M. Avanzi, A. Lacheret-Dujour, and B. Victorri, "ANALOR. A tool for semi-automatic annotation of French prosodic structure," in *Proc. of Speech Prosody*, 2008, pp. 119–122.

[25] G. Christodoulides, M. Avanzi, and A. C. Simon, "Automatic labelling of prosodic prominence, phrasing and disfluencies in French speech by simulating the perception of naïve and expert listeners," in *Proc. of Interspeech*, 2017, pp. 3936–3940.

[26] E. Marsi, M. Reynaert, A. van den Bosch, W. Daelemans, and V. Hoste, "Learning to predict pitch accents and prosodic boundaries in Dutch," in *Proc. of ACL*, 2003, pp. 489–496.

[27] F. Tamburini and C. Caini, "Automatic annotation of speech corpora for prosodic prominence," in *Proc. of LREC*, 2004, pp. 53–58.

[28] N. Braunschweiler, "ProsAlign - The Automatic Prosodic Aligner," in *Proc. of ICPhS*, 2003, pp. 3093–3096.

[29] F. Kügler, S. Baumann, B. Andreeva, B. Braun, M. Grice, J. Neitsch, O. Niebuhr, J. Peters, C. T. Röhr, A. Schweitzer, and P. Wagner, "Annotation of German intonation: DIMA compared with other annotation systems," in *Proc. of ICPhS*, 2019, p. No. 181.

[30] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Speech Communication*, vol. 94, no. C, pp. 62–74, 2017.

[31] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer, Speech and Language*, vol. 45, no. C, pp. 326–347, 2017.

[32] IPDS, "CD-ROM: The Kiel Corpus of Spontaneous Speech, vol i- vol iii," Christian-Albrechts Universität zu Kiel, 1997.

[33] B. J. B. Schmitt, "Amfm decompy documentation 1.0.8," http://bjbschmitt.github.io/AMFM_decompy/, 2018.

[34] S. Zahorian and H. Hu, *A spectral/temporal method for robust fundamental frequency tracking*, 07 2008.

[35] M. Duarte and R. Watanabe, "Notes on scientific computing for biomechanics and motor control," https://github.com/BMClab/BMC, 2018.

[36] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: A structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[37] 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.

[38] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, 2008. [Online]. Available: http://www.biomedcentral.com.proxy.lib.uiowa.edu/1471-2105/9/307/abstract

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

[40] W. Kirch, "Pearson's correlation coefficient," in *Encyclopedia of Public Health*. Dordrecht: Springer, 2008.

[41] I. Lehiste, *Suprasegmentals*. Cambridge: MIT Press, 1970.

[42] 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.