USING WORD-LEVEL FEATURES FOR PROSODIC PROMINENCE DETECTION IN CONVERSATIONAL SPEECH

Julian Linke[1], Gernot Kubin[1], Barbara Schuppler[1]

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology
[1]{linke, gernot.kubin, b.schuppler}@tugraz.at

ABSTRACT

This paper focuses on the automatic detection of prominent words in conversational speech. Most tools for prominence detection rely on prosodic features extracted at a syllable- or phone level and their accuracy thus strongly depends on the quality of the given phone-level segmentation. Given the high degree of pronunciation variation in conversational speech, automatic phonetic segmentation is not accurate enough to detect prominence reliably. Here we explore different approaches to prominence detection that require merely a prior word-level segmentation. The first experiment shows that by using word-level prosodic features cross-validation accuracies of $88\% \pm 4\%$ can be reached, and that word duration is the most important feature. The second experiment introduces entropy-based fundamental frequency and intensity features for prominence detection. Our findings suggest that entropy-based, word-level features can provide a robust approach to detecting prominent words in conversational speech.

Keywords: prosodic prominence, conversational speech, entropy, Austrian German

1. INTRODUCTION

Many tools to detect prosodic prominence (e.g., ProsAlign [1], Tamburini and Wagner [2] or Mishra et al. [3]) operate with features which rely on syllabic structure and/or phonetic annotations. These tools are motivated by linguistic studies that found that certain syllable-level characteristics of fundamental frequency (F0), intensity (RMS) and duration (e.g., speech rate, stressed syllable duration) are cues to perceived prominence, where the conclusions drawn about these cues' importance are not uniform: [4] showed that vowel duration is a more important cue for prominence perception than RMS, whereas [5, 6] found F0 to be a more important cue for prominence than syllable duration. [7] reported that RMS is a more important cue of perceived prominence than other continuous-valued prosodic variables [7]. In the mentioned studies, extracted durational features are related to vowel or syllable durations (e.g., [4, 7]). In contrast, [8] found word duration to be the strongest cue to prominence.

Given the high degree of reduction and overlapping sounds in conversational speech [9], the quality of automatically created phone segmentations has a lower accuracy than for read speech. [10] showed that, whereas for read speech, automatic phonetic annotations are good enough for a subsequent automatic prosodic boundary detection, for conversational speech, manual correction is required. In order to avoid the necessity of manual correction, we explore the use of word-level prosodic features, as the automatic detection of word boundaries showed to have a higher precision for spontaneous speaking styles than of phone boundaries.

So far, most studies on prosodic prominence analyzing F0/RMS contours considered features related to specific characteristics of those curves (i.e., mean, maximum, etc.). To the best of our knowledge, it has not yet been investigated whether entropy-based F0/RMS features, which directly relate to their distribution distinguishing prominence levels. In general, entropy-based features have broadly been used in speech science: For instance, [11] use relative entropy to measure the distance between two speech spectral distributions in concatenative synthesis applications whereas [12] showed that spectral entropy features which interpret the spectrum as a probability mass function, improved the performance of an automatic speech recognition (ASR) system. A study on voice signal characterization tested entropy measures coming from raw audio signals in order to extend voice analysis methods [13]. With respect to prosody of emotional expressions, it has been shown that the use of features capturing F0/RMS variability by calculating entropy from F0/RMS-curves helps to distinguish between arousal conditions in a free-speech setting [14].

This paper presents two approaches for automatic prominence detection in conversational Austrian German. The first experiment uses traditional F0/RMS features (i.e., maximum, mean, etc.) extracted a the word level and durational features (i.e.,

word duration and speech rate variation). The second experiment explores a new approach, i.e., the use of entropy-based F0/RMS features, which implicitly encode distribution information and thus do not require phone-level information.

2. MATERIALS AND METHODS

2.1. GRASS Corpus

The *Graz Corpus of Read and Spontaneous Speech* (GRASS) [15] contains Austrian German conversational speech from 38 Austrian speakers, containing a total of approx. 20h of speech. Word- and phone level segmentations were created by means of a forced alignment using a Kaldi-based ASR system with a lexicon containing on average $5.57-6.18$ pronunciation variants per word type [16]. Phonetically trained transcribers created prosodic annotations for a total of 5234 word tokens from 34 speakers of GRASS. The prominence annotations distinguished the prominence levels 0 (no prominence, *PL0*), 1 (weak prominence, *PL1*), 2 (strong prominence) and 3 (emphatic prominence) [17]. In this study, we combined prominence levels 2 and 3, as emphatic prominence occurred rarely (*PL2*). Annotations were created in three stages: One annotator created a first version, which later was corrected by her/him and subsequently corrected by one of the other annotators. Based on a small subset annotated by two different annotators in those stages, the inter-annotator agreement was calculated: The overall Cohen's kappa was 0.72 (598 tokens), 0.72 for level 0 vs. 1 (371 tokens), 0.92 for level 0 vs. $2/3$ (446 tokens) and 0.57 for level 1 vs. $2/3$ (275 tokens). Other studies obtained similar agreements of 0.53 [2] or 0.84 [7].

2.2. Basic Prosodic Features

84 F0 and RMS features: All features were calculated at the word level. We calculated F0 with the library AMFM decompy [18] which includes an implementation of the pitch detection algorithm YAAPT [19]. Intensity features were generated directly from the waveform by calculating the root mean square. For F0/RMS, and their respective first and second derivatives, we extracted 10 measurements: maximum, minimum, range, relative position of maximum and minimum in the word, mean, median, first and third quartile and standard deviation (60 features). Additionally, we extracted left and right slope of the maximum and minimum, absolute and relative onset and offset within the word, as well as the maximum, minimum, range and mean relative to the utterance (24 features).

12 Durational features (DUR): We extracted word duration, phrase-level speech rate (i.e., number of segments per phrase), local speech rate (i.e., the number of segments per word duration), and relative speech rates (i.e., the ratio of the local speech rate and the minimum, maximum or median of local speech rates within a phrase). Additionally, we calculated the minimum, maximum, range, mean, median and standard deviation of local speech rates within a phrase.

2.3. Entropy-based Features

Entropy measures the spread of probability distributions and provides a measure of uncertainty of a random variable X [20]. If the random variable X assumes values $x_i \in \mathscr{X}$ where $\mathscr{X}$ is a finite set, the definition of entropy can be stated as

$$(1) \quad H(X) = -\sum_i p_i \cdot \log p_i,$$

where $p_i = Pr\{X = x_i\}$ describes the probability of X taking the value x_i, assuming that $p_i \cdot \log p_i = 0$ for $p_i = 0$.

If we observe a sequence of N (non-negative) feature values $\langle f[1], f[2], \ldots, f[N] \rangle$ within a given word, we can measure the spread of these values also by a formal entropy where the (pseudo-)probability distribution is defined by normalizing the feature values

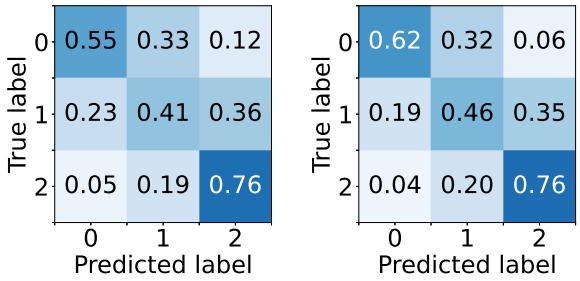
$$(2) \quad p_i = \frac{f[i]}{\sum_{i=1}^{N} f[i]}$$

such that the condition for the total probability is fulfilled: $\sum_{i=1}^{N} p_i = 1$.

With this definition, the entropy (1) achieves its maximum $H_{max} = \log N$ if the feature sequence is constant $f[1] = f[2] = \cdots = f[N] = \text{const.}$, and its minimum $H_{min} = 0$ if all probabilities according to equation (2) turn out to be close to either 1 or 0, e.g., for a very non-uniform feature sequence within the given word. Note that this entropy measures the (relative) feature variability within the word, but without accounting for the time order of the feature contour. Finally, we also experimented with a normalized entropy $\tilde{H}$ obtained from division by the sequence length N:

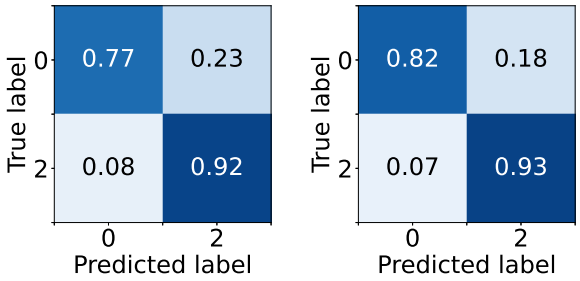
$$(3) \quad \tilde{H} = \frac{H}{N}.$$

For our experiments, we applied equation (2) to the extracted F0/RMS contours and calculated four



(a) F0+RMS **(b)** F0+RMS+DUR

Figure 1: Confusion matrices (3 classes) with F0 and RMS features **(a)** and all features **(b)**.



(a) F0+RMS **(b)** F0+RMS+DUR

Figure 2: Confusion matrices (2 classes) with F0 and RMS features **(a)** and all features **(b)**

additional entropy-based features (ENT) with equations (1) and (3) leading to two (pseudo-)entropies H and two normalized (pseudo-)entropies $\tilde{H}$ of F0/RMS.

Simulations with uniform and non-uniform distributions indicated that these entropy-based features depend primarily on the number of possible outcomes N, which in our case corresponds to word duration. Nevertheless, for words of similar lengths these measurements also encode contour variations by capturing deviations from uniform distributions.

2.4. Random Forest

We trained Random Forest classifiers (RFCs) with the *scikit learn* toolkit (version 0.21.3) [21] RFCs were built with 100 estimators, default maximum depth, a minimum samples split of 2 and the Gini impurity for measuring the quality of a split. For each of the different feature sets, we present results from two conditions, one for 2 classes (*PL0* vs. *PL2*), and one for 3 classes (*PL0* vs. *PL1* vs. *PL2*). Each classification experiment involved two steps: First, RFCs were trained with the entire feature set in order to learn about the feature's relative importance. Second, a (final) RFC was trained with the 15 most important features as given by the first step. The training and test sets were based on a random 80/20 split and we present associated F1-scores. Additionally, we provide means and standard deviations of accuracies resulting from 10-fold cross-validation experiments in order to estimate the model's generalization ability.

3. THE ROLE OF DURATIONAL FEATURES

In order to learn about the role of durational features for prominence detection, we conducted two classification experiments. While the first RFC uses all 96 F0, RMS and durational features (F0+RMS+DUR) described in Sec. 2.2, the second RFC uses the 84 F0/RMS related features only.

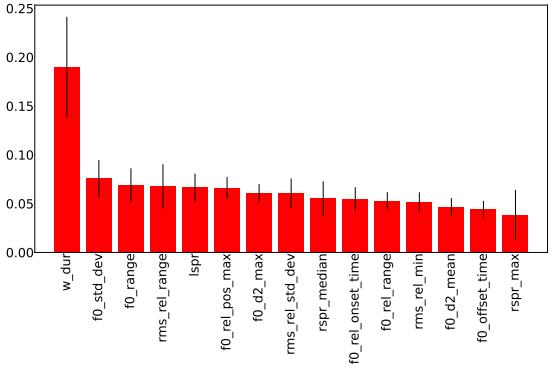


Figure 3: RFC feature importances for 3 class problem with F0, RMS and DUR features.

Fig. 1 and Fig. 2 show the confusion matrices of RFCs which were trained on the entire basic feature set (F0+RMS+DUR) and on a subset without durational features (F0+RMS). We observe that classification performance between non-prominent and highly prominent words is high in both cases, but that non-prominent words were better classified when the RFC was trained with the entire feature set than without DUR (recall 82% vs. 77%). For *PL0* the F1 increased form 80.3% to 84.4% by adding DUR, and for *PL2* from 89.8% to 91.6%. Corresponding cross-validation accuracies were 88% $\pm$ 4% (F0+RMS+DUR) and 85% $\pm$ 4% (F0+RMS).

RFCs with 3 classes (Fig. 1) showed a similar behaviour since the recall for *PL0* of 61.6% (F0+RMS+DUR) was higher than the recall of 55.2% (F0+RMS). However, in case of *PL1*, recalls of only 45.7% (F0+RMS+DUR) and 40.6% (F0+RMS) were achieved, while in both cases approx. 35% of tokens from *PL1* were predicted as *PL2*. Respective F1s of *PL0*/*PL1* were 63.1%/47.2% (F0+RMS+DUR) and 55.9%/43.2% (F0+RMS). Interestingly, recalls of highly prominent words were similar in both cases (approx. 76%). In this case, cross-validation accuracies were 63% $\pm$ 5% (F0+RMS+DUR) and 60% $\pm$ 6% (F0+RMS). Fig. 3 shows the feature ranking corre-

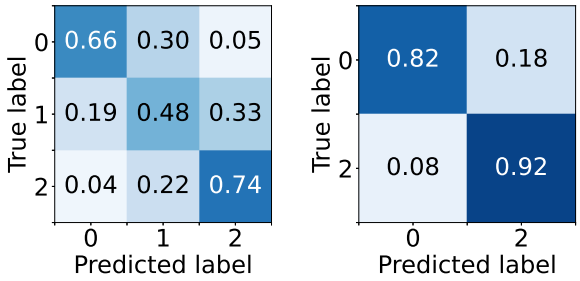


Figure 4: Confusion matrices from experiments with 15 best features (F0+RMS+DUR+ENT).

sponding to the averaged impurity decrease of the RFC for 3 classes trained with the 15 best features (F0+RMS+DUR). Word duration (w_dur) has by far the highest importance among all features, capturing almost 20% of the overall importance. Other durational features (see Sec. 2.2) like the local speech rate (lspr) or relative speech rates (rspr_median and rspr_max) were also present in the feature ranking and had similar importances as the relative relationships of the F0/RMS contours. This finding is in line with the study by [8], showing that word-duration is the most important feature for prominence detection in read speech.

Overall, durational features improve the RFC accuracy for detecting prominent words of conversational speech. Results from previous investigations pointed towards different trends. Whereas [4] found vowel duration to be an important cue to prominence in spontaneous speech, [6] concluded that F0 and [7] that RMS play a more important role. These studies, however, did not consider word duration, which in our experiments resulted to be the most important feature among all durational, F0/RMS features to classify prominence in conversational Austrian German.

4. THE ROLE OF ENTROPY-BASED FEATURES

To learn about the role of entropy-based features (ENT) for prominence detection, and whether they can complement phone-based durational features, we conducted two classification experiments. While the first RFC uses all 100 features (F0+RMS+DUR+ENT), the second RFC does not use any durational features (F0+RMS+ENT).

The RFC for 3 classes with F0+RMS+DUR+ENT features resulted in a large number of confusions of *PL1* (recall/F1: 48%/48.6%) with *PL0* or *PL2*, where approx. 19% of *PL1* was classified as *PL0* and 33% as *PL2*. In contrast, recalls/F1s of 65.6%/66.1% (*PL0*) and 74.1%/73.1% (*PL2*) indicated less confusions with others classes (i.e., only $4-5\%$ of non-prominent or highly-prominent words were classified as highly-prominent or non-prominent words). This result is to be expected, as also the inter-rater agreement showed to be lowest/highest for these classes. Overall cross-validation accuracies reached $62\% \pm 7\%$. Compared to the RFC without ENT, the classification of non-prominent words improves by adding the ENT features (recall $66\% > 62\%$). Furthermore, the comparison with the RFC trained without any durational features (F0+RMS+ENT) indicated that developed entropy-based features compensate for durational information (similar F1s for classes *PL0*/*PL2* of approx. 84%/91%). With respect to the feature importances for the RFC with F0+RMS+DUR+ENT, we observed that the 5 best features comprised word duration as well as the entropy and normalized entropy features, which all had average importances of $> 8\%$ (capturing approx. 45% of the overall importance), while all other features had importances $< 6.7\%$.

For both the 2 and the 3 class problem, prominence detection was best when adding entropy-based F0/RMS-features to the feature set. To the best of our knowledge, there exist no earlier studies on prominence detection using similar entropy-based F0 and RMS features.

5. CONCLUSION

This paper investigated different word-level features to detect prosodic prominence, to avoid the necessity of creating manual phonetic segmentations for conversational speech. Overall, the classification performances achieved with our different sets of features were in the range of the human inter-rater agreements for the respective classes. We found that durational features (incl. speech rate variations) have a higher importance than F0/RMS features, and that among them, word duration is by far the most important feature. Experiments with entropy-based F0/RMS features showed that they encode necessary durational information along with information about the features' distribution, making them useful for classifying prominence levels in conversational speech. In future, we will explore whether entropy-based F0/RMS features are also useful to capture other prosodic characteristics in speech, both with respect to speech analysis as well as in ASR.

6. ACKNOWLEDGEMENTS

This work was partly funded by grant P-32700-NB from FWF (Austrian Science Fund). We would like to thank the transcribers Nina Richter and Nikolaus Tlapak for their efforts.

7. REFERENCES

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

[2] Tamburini, F., Wagner, P. 2007. On automatic prominence detection for German. *Proc. of Interspeech*, 1809–1812.

[3] Mishra, T., Sridhar, V. R., Conkie, A. 2012. Word prominence detection using robust yet simple prosodic features. *Proc. Interspeech 2012*, 1864–1867.

[4] Cole, J., Mo, Y., Hasegawa-Johnson, M. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1, 425–452.

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

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

[7] Baumann, S., Winter, B. 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgements. *J. Phon* 70, 20–38.

[8] Linke, J., Kelterer, A., Dabrowski, M. A., Zarka, D. E., Schuppler, B. 2020. Towards automatic annotation of prosodic prominence levels in Austrian German. *Proc. Speech Prosody 2020*, 1000–1004.

[9] Johnson, K. 2004. Massive reduction in conversational American English. *Proc. in Spontaneous speech: Data and analysis. Proc. of the 1st session of the 10th international symposium*, 29–54.

[10] Ludusan, B., Schuppler, B. 2022. An analysis of prosodic boundaries across speaking styles in two varieties of German. *Speech Communication* 141, 93–106.

[11] Klabbers, E., Veldhuis, R. 2001. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing* 9, 39–51.

[12] Misra, H., Ikbal, S., Bourlard, H., Hermansky, H. 2004. Spectral entropy based feature for robust ASR. *ICASSP*. IEEE, 193–196.

[13] Rogério Scalassara., P., Eugenia Dajer., M., Dias Maciel., C., Carlos Pereira., J. 2008. Voice signals characterization through entropy measures. *Proceedings of the First International Conference on Bio-Inspired Systems and Signal Processing - Volume 2: BIOSIGNALS, (BIOSTEC 2008)*. INSTICC SciTePress, 163–170.

[14] Cohen, A. S., Hong, S. L., Guevara, A. 2010. Understanding emotional expression using prosodic analysis of natural speech: Refining the methodology. *Journal of behavior therapy and experimental psychiatry* 41 2, 150–7.

[15] Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., Pessentheiner, H. 2014. GRASS: The Graz corpus of Read And Spontaneous Speech. *Proc. of LREC*, 1465–1470.

[16] Wasserfall, S. 2020. Automatic Speech Segmentation using Kaldi. Master's thesis Technical University Graz.

[17] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94C, 62–74.

[18] Schmitt, B. J. B. 2018. AMFM decompy documentation 1.0.8. https://bjbschmitt.github.io/AMFM_decompy/.

[19] Zahorian, S., Hu, H. 2008. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America* 123, 4559–71.

[20] Cover, T. M., Thomas, J. A. 2006. Elements of Information Theory 2nd Edition. *Wiley Series in Telecommunications and Signal Processing*.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.