



Syllable-Level Prominence Detection with Acoustic Evidence

Je Hun Jeon, Yang Liu

Computer Science Department
 The University of Texas at Dallas, Richardson, TX, USA
 {jhjeon, yangl}@hlt.utdallas.edu

Abstract

Accurate prominence annotation benefits many spoken language understanding tasks as well as speech synthesis. In this work, we conduct a thorough study using acoustic prosodic cues for prominence detection in speech. This study is different from previous work in several aspects. In addition to the widely used prosodic features, such as pitch, energy, and duration, we introduce the use of cepstral features. Furthermore, we evaluate the effect of different features, speaker dependency and variation, different classifiers, and contextual information. Our experiments on the Boston University Radio News Corpus show that although the cepstral features alone do not perform well, when combined with prosodic features they yield some performance gain and, more importantly, can reduce much of the speaker variation in this task. We find that the previous context is more informative than the following context, and their combination achieves the best performance. The final result using selected features with context information is significantly better than that in previous work.

Index Terms: prosody, prominence, pitch accent, speaker variation

1. Introduction

In English, prominence¹ refers to an intonational prominence associated with a metrically strong syllable or word at the sentence level, which is often marked by pitch movements, increased energy, or prolonged duration. The task of automatic detection of prominence has received a considerable amount of research attention. It plays an important role not only in rich annotation of speech corpora but also in text-to-speech systems. Prominence also provides useful information in other tasks, such as automatic speech recognition, information extraction, keyword detection and summarization of spoken documents.

Prior research for automatic detection of prominence in speech has been conducted at both the syllable [1, 2, 3] and word level [4, 5, 6]. Most of previous efforts adopted supervised approaches using various machine learning techniques such as decision trees, neural networks, maximum entropy models, Gaussian mixture models, support vector machines, and boosting methods. Some also used unsupervised and semi-supervised methods [7, 8]. These approaches typically use prosodic evidence from the speech waveform such as pitch, energy, and duration, as well as lexical/syntactic information from text such as part-of-speech, syllable/word identity, and term frequency. Prosodic and lexical information sources are often used separately, and combined at the decision level for the final system output. Most studies reported increased accuracy (to different extents) by adding more features or selecting better classifiers.

¹This is also called pitch accent, or stress in different studies.

Some previous work has showed that the performance using lexical/syntactic information is better than that using acoustic cues. However, most of them used human transcripts to obtain textual features. We expect that when there are automatic speech recognition (ASR) errors, lexical and syntactic information is not reliable. In this study, we focus on acoustic prosodic cues since these features are potentially more robust in face of ASR errors. In addition, there are a lot of variations in acoustic prosodic cues. A better understanding and model of these variations will greatly help prominence detection and many speech applications.

We aim to answer the following questions with this work: (1) *Are acoustic cepstral features useful for prominence detection?* Previous work has mainly used prosodic cues such as pitch, energy and duration in this task, and ignored cepstral features. We evaluate the effectiveness of cepstral features for prominence detection, as well as evaluating their effectiveness when combined with prosodic information. (2) *What is the effect of speaker variation?* We compare the speaker independent and dependent setups in order to understand speaker dependency and variation. (3) *Is there any gain from feature selection?* We expect that feature selection may eliminate redundant or noisy features and yield performance gain. (4) *Is contextual information needed for this task?* We incorporate features from the previous and following syllables and evaluate system performance. (5) *Are the findings sensitive to different classifiers?* We compare two different classifiers, multilayer perceptron (MLP) and support vector machines (SVM), for different configurations. Our experimental results show that we can achieve much better performance using selected features with context information, compared to that using prosodic features only. Our results also outperform those in previous work.

In the next section, we provide details on the tasks including the classification method and features used. In Section 3, we present our experimental setup and results. The final section gives a brief summary along with future directions.

2. Prominence Detection Approach

Automatic detection of prominence requires appropriate representation schemes that can characterize prosody in a standardized manner. One of the most popular prosodic event labeling schemes is the ToBI framework [9]. The prominence tones (*) are marked at every accented syllable and have five types according to pitch contour: H*, L*, L*+H, L+H*, H+!H*. Since there is only very limited data annotated with such prosodic events, using the detailed representation of prominence tones creates a serious sparse data problem. This problem can be alleviated by grouping ToBI labels into coarse categories, such as presence or absence of prominence. This also significantly reduces ambiguity of the task. In this paper, we thus use coarse

10.21437/Interspeech.2010-507

representation (presence versus absence) for prominence, similar to the task definition used in most previous work.

Prominence detection is performed at the syllable level, since stress is defined to occur on syllables, and it is also easy to transform prominence information from syllables to words if needed. We derive time information for syllables from the speech waveform and phone-level forced alignment of the transcriptions, similar to [2, 3, 10]. The prominence detection problem is a binary classification task, that is, a classifier is used to determine whether the syllable is prominent or not given the acoustic evidence. We assume that each prominent syllable is independent of others and it is only dependent on the acoustic observations in the corresponding location. However, we try to capture some of the dependency information at the feature level. For the classifier, we use the implementation of MLP and SVMs in the Weka toolkit [10]. The MLP classifier showed better performance in previous studies [3], and SVM was widely adopted for emotion and many paralinguistic effect recognition [11].

Most of previous studies only used prosodic features for prominence detection, including pitch, energy and duration. In addition to prosodic features, in this study, we add cepstral features, represented by Mel Frequency Cepstral Coefficients (MFCC) 1~12. Pitch, energy, and cepstral values are computed using Praat [12]. In order to reduce the effect by both inter-speaker and intra-speaker variation, both pitch and energy values were normalized (z-value) with utterance specific means and variances. The duration values are extracted from forced alignment data. Table 1 lists all the features we use.

Primary cues	Functions	# features
Pitch	Stat, contour, slope	18
Energy	Stat, contour	13
Cepstral	Stat	84
Duration	-	6
Total		121

Table 1: Features used for prominence detection.

For pitch, energy, and cepstral features, we apply several categories of functions to generate derived features.

- **Statistic functions:** minimum, maximum, range, mean, standard deviation, skewness and kurtosis value. These are used widely in prosodic event detection and emotion detection.
- **Contour:** This is approximated by taking 6 leading terms in the Legendre polynomial expansion. The approximation of the contour using the Legendre polynomial expansion has been successfully applied in quantitative phonetics [13] and in engineering applications [14]. Each term models a particular aspect of the contour such as the mean of the segment, the slope, and information about the curvature.
- **Slope changes:** first and last pitch slope, maximum plus and minus pitch slope, the number of changes in the pitch slope patterns. These are used in previous work (e.g., [15]) and are shown to be useful for various tasks. For estimating values of slope changes, we used the Momel algorithm [16] to reconstruct the pitch values of the unvoiced segment and then estimated pitch slope features.

Note that not all the functions are applied to each feature group. See Table 1 for the functions applied. For duration fea-

tures, raw, normalized, and relative durations (ms) of the syllable and vowel are used. Normalization (z-value) is performed based on each syllable and vowel statistics. The relative value is the difference between the normalized current duration and the following one.

3. Experiments

3.1. Experimental Setup

Boston Radio News Corpus (BU) [17] is commonly used in prosodic classification research. It consists of news stories read by professional radio announcers, and is partially annotated with pitch accents, boundary tones, and boundary break indices, based on the ToBI prosodic labeling conventions for American English. As mentioned earlier, we adopt a binary setup for prominence detection, i.e., collapsing all prominence tone types to a single prominence class (prominent). Among the ToBI annotated data, we use 253 utterances from 2 female (*f1a* and *f2b*) and 2 male (*m1b* and *m2b*) speakers without any duplicated sentences. Statistics about the data are shown in Table 2.

Speaker	# uttr	# word	# syl	# prominent
f1a	54	2,499	4,158	1,589
f2b	124	9,090	14,996	5,161
m1b	48	2,949	4,804	1,573
m2b	27	1,510	2,557	891

Table 2: The statistics of the BU data used in our experiments (number of utterances, words, syllables, and prominent syllables for each speaker).

We use two different data splits in order to evaluate the effect of using different speakers: speaker independent (SI) and speaker dependent (SD) setup. For SI, we applied a leave-one-speaker-out strategy. For each experiment, we used the data from three speakers for training and the remaining speaker was used for testing. Speaker *f2b* was never left out because it contains the most data and there would not be enough training data if she is used as the test speaker. Therefore there are three test folds (for speakers *f1a*, *m1b*, and *m2b* respectively). For SD, we use all the four speakers in both training and testing. We created three folds in order to have a similar setup as in SI. Each fold in SD has similar number of instances as in the corresponding SI fold (numbers vary across different folds). We also made sure that for each fold the data proportion for the four speakers is similar. Note that this SD set up is not the same as speaker dependent modeling in speech processing. It only means that the test speakers also appear in the training set (different data though). We use this term to indicate the difference with the SI setup. The performance of each experiment is evaluated using F-measure value. This is chosen since the data is not balanced (between the prominent and not prominent class) and thus accuracy is a bit biased.

3.2. Experimental Results

3.2.1. Results I: No contextual information

First we use only the features for the current syllable without any contextual information to examine the effect of various features. Figure 1 shows the results for different feature setups: using individual feature groups, their combination, and feature selection (details below); different classifiers (MLP and SVM);

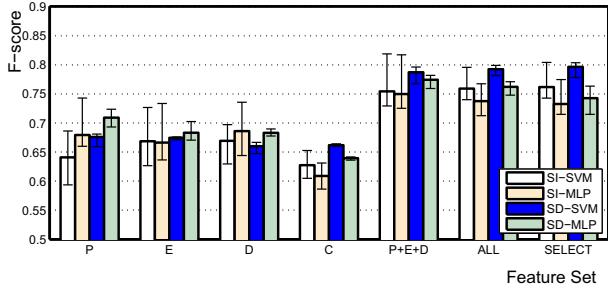


Figure 1: Results of prominence detection for different conditions: individual and combined features; SI vs. SD; MLP and SVM classifiers.

and SI and SD data splits. *P*, *E*, *D*, and *C* on the x-axis refer to pitch, energy, duration, and cepstral feature sets respectively. The average F-measure as well as the maximum and minimum scores (among different test folds) for each experiment are shown in the figure. We discuss some observations from the results in the following.

A. Using individual feature group

When only one feature category is used, prosodic features, pitch, energy, and duration, achieve better performance than cepstral features. Generally the performance on SD condition is better than SI, as expected. The performance gap between SD and SI setups is much bigger when pitch and cepstral features are used, even though pitch values are normalized to reduce inter-speaker variance. The energy and duration features show similar results for the two conditions. For performance variation ($max - min$ score among different folds) in each experiment, SI setup shows bigger variance than SD. This is because of the speaker difference in SI, but each fold of SD has similar proportion of each speaker, and the only difference is the number of instances. For SI condition, the variance using prosodic features is quite high, 8.3%, 9.7%, and 9.2% for pitch, energy, and duration respectively when using MLP. However, using cepstral features has a much lower variance (4.5%). Similar variance patterns are shown using SVM classifier. In SD condition, the variances are smaller than SI (3.1% for pitch, 3.2% for energy, 1.2% for duration, and 0.5% for cepstral). The most speaker dependent and variant information is pitch. The energy and duration features show less speaker dependent performance but bigger variance, whereas cepstral features have less variance from different folds although they are speaker dependent.

B. Using combined features

For the feature combination test, we use two combinations: *P+E+D* uses prosodic information only, while *ALL* uses prosodic and cepstral features. Similar to the results using feature groups separately, SD condition shows higher performance and lower variance. The performance gain using all features together over using only prosodic information is not significant, but the variance is much smaller, 5.6% for the former and 9.0% for the latter using SVM for the SI condition. These results indicate that introducing cepstral features reduces the speaker variance.

C. Feature selection

The last group in Figure 1 shows the results after using feature selection. We evaluated both forward selection and backward elimination, using 5-fold cross validation and SVM. Ta-

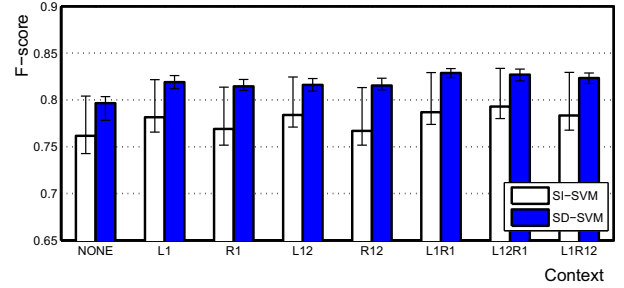


Figure 2: Results of prominence detection using different contextual information for both SI and SD conditions.

ble 3 shows the number of selected features using backward feature selection (the best result from forward selection is not as good as the backward method). The final feature set is less than half the size of the original one from Table 1. Most of the removed features are cepstral features (56 features), and the other 9 removed features are prosodic features. However, many cepstral features still remain in the selected features. This means that some cepstral features can contribute to prominence detection. Using the selected best feature set for prominence detection, there is no significant performance change; however, reducing the number of features has other benefits of less computational complexity and possibly increases robustness.

Primary cues	# features
Pitch	15
Energy	8
Cepstral	28
Duration	5
Total	56

Table 3: The number of selected features.

D. Classifier comparison

Regarding the two classifiers we used, MLP shows better performance in the lower feature dimension in the above experiments, but worse in the higher one. The performance of SVM is better when the number of features increases. One possible reason for MLP's lower performance in the high dimensional feature space is that there is not enough data to train weights for each node. We also tuned the options in these classifiers, such as the number of hidden nodes and the number of iterations in MLP, and kernel type and cost in SVM, but there is not significant performance improvement.

3.2.2. Results II: Effect of context

In the previous experiment, we assumed that a syllable's prominence tag is only dependent on its acoustic observations, and did not consider the effect of contextual information. To alleviate this restriction, we expand the features by incorporating information about the neighboring syllables. We evaluate using different context, left and right, and different distance. SVM classifier is used in this experiment using the best feature subset from the previous feature selection results. Figure 2 shows the results, where L and R refers to previous or following syllables respectively, and the number after L or R refers to the adjacency (1, or both 1 and 2).

We can see that adding contextual information improves

performance for both SI and SD setups, with some differences between them. In the SI configuration, adding previous context (L1 and L12) are more effective than following contexts (R1 and R12). The best performance comes from the combination of two previous and one following syllables (L12R1). For SD testing, using previous context shows better performance than using the following ones. This is similar to SI, but the effect for SD is rather small. Using one previous and one following context (L1R1) shows the best performance and adding more contexts does not yield further gain.

3.2.3. Results III: Final result using the best setup

The final result using the best setup is shown in Table 4. This uses the best selected feature subset, and one previous and one following context. To compare the performance of our models with previous work on BU with similar configuration, we also include some representative results from previous work. Our SD setup is similar to the 5-fold cross validation used in previous work, except that we use fewer test folds (3 folds). As shown in our experiments, the variation among the test folds in SD configuration is very small, so SD is comparable to 5-fold cross validation in prior work. We can see from the results that the performance using our expanded feature set including cepstral and contextual information is significantly better than previous work.

	Test set	accuracy	F-score
Ananthakrishnan et al. [2]	5-fold	80.1%	-
Jeon and Liu [3]	5-fold	83.5%	0.75
Our approach	SI	85.7%	0.79
	SD	88.3%	0.83

Table 4: Results of prominence detection using the best setup (selected subset of features and contextual information), in comparison with previous work.

4. Conclusions

In this paper, our goal was to answer several questions in prominence detection regarding cepstral features, speaker variation, contextual effects, and classifiers. We find that when used individually, cepstral features showed worse performance than other prosodic cues (pitch, energy, and duration). After combining cepstral features with prosodic features we did not achieve significant additional performance gain, but cepstral features effectively reduced the variation among speakers. This might be because that cepstral features can discriminate syllable identities, which provide clue of prominence [2], resulting in less speaker variation. This reduction was shown when cepstral features were used alone as well. Feature selection did not improve performance but significantly reduced the number of features. Most performance gain was achieved by adding contextual information. The previous context was more effective than the following, and the best performance was achieved when the previous and following contexts were combined. We found that MLP classifier showed better performance than SVM when using smaller number of features, but when the number of features increased, SVM performed better. The final results using selected features with contextual information showed much better performance than that in previous work.

For the future work, we plan to apply feature expansion to other types of prosodic event detection, such as intonational

phrasal boundaries and break indices. Since these tasks use similar features as for prominence detection, we expect that the performance gain will also hold for those tasks. In addition, we plan to combine acoustic and lexical cues for prosodic event detection and also investigate semi-supervised methods.

5. References

- [1] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transaction on Speech and Audio Processing*, vol. 2(4), pp. 469–481, 1994.
- [2] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical and syntactic evidence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16(1), pp. 216–228, 2008.
- [3] J. H. Jeon and Y. Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features," *Proc. of ICASSP*, pp. 4565–4568, 2009.
- [4] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ann-based syntactic-prosodic model and gmm-based acoustic prosodic model," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1509–1512, 2004.
- [5] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky, "To memorize or to predict - prominence labeling in conversational speech," *Proc. of NAACL-HLT*, pp. 9–16, 2007.
- [6] A. Margolis and M. Ostendorf, "Acoustic-based pitch-accent detection in speech: dependence on word identity and insensitivity to variations in word usage," *Proc. of ICASSP*, pp. 4513–4516, 2009.
- [7] G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," *Proc. of HLT-NAACL*, pp. 224–231, 2006.
- [8] J. H. Jeon and Y. Liu, "Semi-supervised learning for automatic prosodic event detection using co-training algorithm," *Proc. of ACL-IJCNLP*, pp. 540–548, 2009.
- [9] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," *Proce. of ICSLP*, pp. 867–870, 1992.
- [10] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with java implementation," *ICONIP/ANZIIS/ANNES International Workshop*, pp. 192–196, 1999.
- [11] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," *Proc. of Interspeech*, pp. 312–315, 2009.
- [12] P. Boersma, "Praat, a system for doing phonetics by computer," *Glott International*, vol. 5(9-10), pp. 341–345, 2001.
- [13] E. Grabe, G. Kochanski, and J. Coleman, "Quantitative modelling of intonational variation," *Proc. of SASRTLM*, pp. 45–57, 2003.
- [14] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15(7), pp. 2095–2103, 2007.
- [15] E. Shriberg, L. Ferrer, S. S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46(3-4), pp. 455–472, 2005.
- [16] D. Hirst and R. Espesser, "Automatic modeling of fundamental frequency using a quadratic spline function," *Travaux de l'Institut de Phonetique d'Aixen-Provence*, vol. 15, pp. 75–85, 1993.
- [17] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pp. 147–154, 2003.