Feature Sets for the Automatic Detection of Prosodic Prominence

Tim Mahrt, Jui-Ting Huang, Yoonsook Mo, Jennifer Cole, Mark Hasegawa-Johnson, and Margaret Fleck

This work presents a series of experiments which explore the utility of various acoustic features in the classification of words as prosodically prominent or nonprominent.

For this set of experiments, a 35,009 word subset of the Buckeye Speech Corpus was used [12]. This subset is divided across fifty-four segments of the Buckeye Speech Corpus. In a previous study, the words were transcribed for prosodic prominence by several teams, of sixteen naive native-speakers of English each, using the method Rapid Prosody Transcription developed in our prior work [10]. In the present study, we mapped the quasi-continuous valued prosody labels from the transcribed portion of the corpus to a binary prominence label. If at least one rater deemed a word prominent, it was labeled 'prominent' or otherwise it was labeled 'nonprominent.' 15,955 were labeled 'prominent,' yielding a baseline chance level of prominence assignment at 54.4%. 90% of the words were used in training the learning algorithms and the other 10% was used in testing.

Several acoustic correlates are associated with prominence, including F0, duration, and intensity [1, 2, 4, 17, 3, 6, 15, 16, 11, 7, 14, 18]. The relative contribution that these play in speech recognition and in recognition by humans is well discussed in the literature [5, 18, 9, 13].

In the first set of experiments, Support Vector Machines (SVM) were used. SVMs were chosen because the task is a vector-input, class-label-output task, and SVMs do well at such tasks. Here, a set of 36 features was used, including both features known to be correlated to prominence and features not known to be correlated, such as the length of the pause after a word. The ten best-performing features were, in order, the minimum energy of the final vowel normalized by phones, the ratio of the energy of the current word to the next word, the post-word pause duration, the word duration normalized by phones, the maximum energy of the last vowel normalized by phone-class, the minimum value of f0 in the following word, the maximum energy of the stressed vowel normalized by phone-class, the stressed vowel duration normalized by phone-class, the minimum energy of the next word, and the maximum energy of the word. The classification accuracy was tested with related features clustered together into four groups: pause, duration, intensity, and pitch. The results are reported in table 1

For the second set of experiments, Hidden Markov Models (HMM) with three hidden

SVM Features	SVM acc.	HMM Features	HMM acc.
pause feats.	61.1	Post-Word Pause Duration (PWPD)	57.7
duration feats.	69.0	Stressed Vowel Duration (SVD)	65.1
intensity feats.	71.4	-	-
pitch feats.	72.1	-	-
intensity + pitch	75.1	MFCC	68.7
intensity + pitch + duration	75.8	MFCC + SVD	65.82
intensity + pitch + duration + pause	76.1	MFCC + SDV + PWPD	56.2

Table 1: Classification accuracy percent using SVMs and HMMs

Context region	pre-stress	stressed syllable	post-stress
Classification accuracy	67.4	66.1	67.4

Table 2: Classification accuracy for context regions using HMMs

states were used. HMMs can take advantage of temporal information in the sequencing of units. Mel-frequency cepstral coefficients (MFCC) were generated using HTK and were used as the encoding of temporal features. These data were concatenated with per-word durational measures, taken from phoneme-occurrence timestamps in the Buckeye corpus. The post-word pause duration is the time between the end of the last phoneme in the current word and the beginning of the first phoneme in the next word. The results for these experiments are summarized in Table 1. Although the feature sets used between the HMM and the SVM are not the same, they correspond to each other. Note that the classification accuracy in the SVM is always higher. For this reason and that many of the top performing features used in the SVM experiment were normalized features, this suggests that temporal information is less useful than changes in the accoustic signal. These findings support evidence found in the human perception of prosody [8].

If temporal information is useful then some temporal regions may be more useful than others. In English, as prominence is primarily expressed on the stressed syllable [7], it may be expected that by extracting features only from the stressed syllable we would obtain provide better prominence classification results, with the other regions of the word contributing noise. However, prominence also has residual effects on the rest of the word. For example, F0 can peak in the post-stress syllable [11, 7].

To test prominence detection based on the stressed-unstressed distinction within the word, the words in Buckeye were split into three regions: pre-stress, stress, and post-stress. MFCC vectors were extracted from each of these regions and were tested independently of each other. The results for the three regions, reported in Table 2, are fairly similar to each other and to the results for the trials reported earlier using MFCCs extracted from the entire word. Thus, interesting information does exist throughout prominent words.

To see if making this contextual information more explicit could be used to improve accuracy, a new feature was created from the sum of the log-likelihoods of each frame being prominent given the model trained on MFCCs in the previous experiment. These values were trained on a new HMM with an accuracy of 56.2%, which suggests that the explicit contextual feature is not useful.

In our final experiment, we modified the classification task so that words were considered 'prominent' when two or more raters labeled a word as 'prominent' (rather than one or more). Words which were not labeled as 'prominent' by any raters were still considered 'nonprominent' but those which were labeled by only a single rater was thrown out. Agreement between labelers can provide greater confidence that the word is indeed prominent, whereas words with only a single 'prominent' judgment are more likely to be mistakes. The accuracy for this zero vs two or more classification task when only using MFCCs is 71.5% as compared to 68.7% for the zero vs one or more task, suggesting that words with only a single judgment of 'prominence' are indeed less reliable.

In this study we sought different strategies to improve learning performance. We found that normalized features are often more informative than raw features. The contribution of

temporal regions was observed and it was found that no one region was the most informative. And finally, by removing labels with low rater agreement, we were able to boost performance.

This study is supported by NSF IIS-0703624 to Cole and Hasegawa-Johnson. For their varied contributions, we would like to thank the members of the Illinois Prosody-ASR research group.

References

[1] M. Beckman. *Stress and non-stress accent*. Foris Pubns USA, 1986.

[2] M. Beckman and J. Edwards. Articulatory evidence for differentiating stress categories. *Phonological structure and phonetic form*, page 7, 1994.

[3] T. Cambier-Langeveld and A. Turk. A cross-linguistic study of accentual lengthening: Dutch vs. *English. Journal of Phonetics*, 27(3):255–280, 1999.

[4] J. Cole, H. Kim, H. Choi, and M. Hasegawa-Johnson. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, 35(2):180–209, 2007.

[5] A. Cutler, D. Dahan, and W. Van Donselaar. Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40(2):141, 1997.

[6] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118:1038, 2005.

[7] D. Ladd. *Intonational phonology*. Cambridge Univ Pr, 2008.

[8] Y. Mo. *Prosody production and perception with conversational speech*. PhD thesis, University of Illinois Urbana-Champaign, 2010.

[9] Y. Mo, J. Cole, and J. Hasegawa-Johnson. How do ordinary listeners perceive prosodic prominence? Syntagmatic vs. Paradigmatic comparison. In *Poster presented at the 157th Meeting of the Acoustical Society of America, Portland, Oregon.*, 2009.

[10] Y. Mo, J. Cole, and E. Lee. Naive listeners prominence and boundary perception. *Proc. Speech Prosody, Campinas, Brazil*, pages 735–738, 2008.

[11] J. Pierrehumbert. *The phonology and phonetics of English intonation*. MIT Cambridge, MA, 1980.

[12] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and et al. *Buckeye corpus of conversational speech (2nd release)*. Columbus, OH: Department of Psychology, Ohio State University, 2007. Retrieved March 15, 2006, from www.buckeyecorpus.osu.edu.

[13] A. Rosenberg. *Automatic Detection and Classification of Prosodic Events*. PhD thesis, Columbia University, 2009.

[14] C. S. *Information Structure and the Prosodic Structure of English*. PhD thesis, University of Edinburgh, 2006.

[15] A. Sluijter and V. Van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4):2471–2485, 1996.

[16] F. Tamburini and C. Caini. An automatic system for detecting prosodic prominence in American English continuous speech. *International Journal of Speech Technology*, 8(1):33–44, 2005.

[17] A. Turk and L. White. Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27(2):171–206, 1999.

[18] K. Yoon. Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody. *Journal of the Modern British & American Language & Literature*, 25(4):197–215, 2007.