Feature Sets for the Automatic Detection of Prosodic Prominence

Tim Mahrt, Jui-Ting Huang, Yoonsook Mo, Jennifer Cole, Mark Hasegawa-Johnson, and Margaret Fleck
Beckman Institute, University of Illinois

Beckman Institute

Goal:

- Determine the utility of various acoustic features in the classification of words as prosodically prominent or nonprominent.

Prior Research:

- Several acoustic correlates are associated with prominence, including F0, duration, and intensity [3, 4, 1].
- The relative contribution of any one feature for prominence recognition is disputed [2, 5, 9].

Data:

- We used a 35,009 word subset of the Buckeye Speech Corpus [8], divided across fifty-four excerpts. In a previous study, the excerpts were transcribed for prosodic prominence by teams of 15-20 individuals using the method of Rapid Prosody Transcription developed in our prior work [6]. In the present study we mapped the quasi-continuous-values prosody labels from the transcribed protion of the corpus to a binary prominence label.
- If at least one rater deemed a word prominent, it was labeled 'prominent,' and otherwise it was labeled 'nonprominent.' 15,955 words were labeled prominent, yielding a baseline rate of prominence occurrences of 54.4%.
- Classification tests were run using SVM (libSVM) and HMM (HTK) models. 90% of the words were used in training the learning algorithms and the other 10% were used in testing.

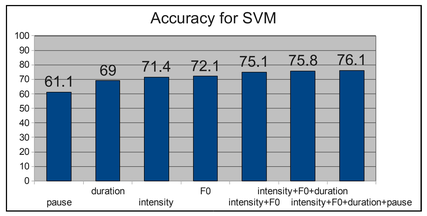
SVM Features:

- Pause features: post-word pause duration
- Raw duration features: duration of vowel in final syllable, stressed vowel duration
- Phone-normalized duration features: duration of vowel in final syllable, word duration, phone duration, duration of the longest phone
- F0 features: min f0, max f0, and mean f0 for the word, the next word, the final vowel, and the stressed vowel; the difference in F0 measures of the current word and the next word based on min F0 and max F0 in both.
- Intensity features: the min, max, and RMS energy for the current word, the next word, the final vowel, and the stressed vowel; the RMS energy difference between the current word and the next word

HMM Features:

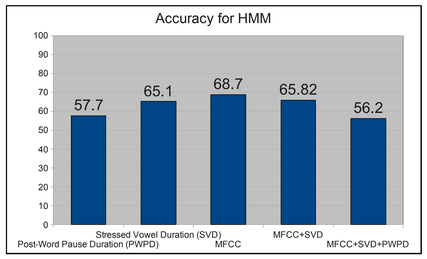
- MFCC vector, post-word pause duration, stressed vowel duration

Experiments



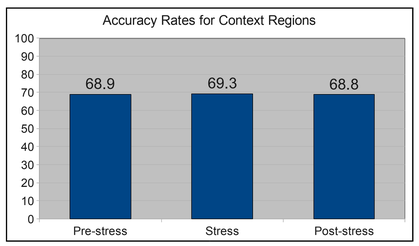
Experiment 1

We used SVMs to train over a set of 36 features divided across four feature vectors: pause, duration, intensity, and F0. An SVM model was selected because SVMs are well-suited to this vector-input, class-label-output task.



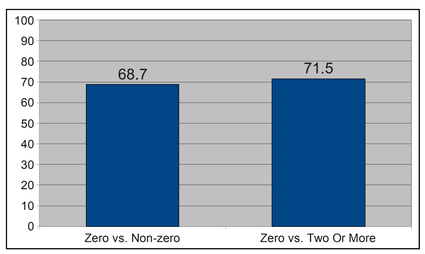
Experiment 2

We used an HMM to take advantage of sequential information. The MFCC feature vectors incorporate information about F0 and intensity. Extracted, explicit features (such as the post-word pause duration) were integrated into the MFCC to compare performance.



Experiment 3

In phonological models, prominence features are associated with the stressed vowel of a word. To test whether this observation holds any consequence for automatic prosody labeling, we trained an HMM model over MFCC feature vectors extracted from one of the three stress regions: pre-stress, stress, or post-stress.



Experiment 4

We anticipated that words which were only labeled as "prominent" by a few raters are more likely to have been mislabeled. To test this, using MFCC feature vectors in an HMM, we removed words where only a single labeler marked the word as "prominent."

Top Performing Features

The top ten features which provided the greatest boost in accuracy in experiment 1 were:

- the normalized minimum energy of the last vowel
- the RMS energy difference between the current word and the next word
- the pause duration
- the normalized word duration
- the normalized maximum energy of the last vowel
- the minimum F0 of the next word
- the normalized maximum energy of the stressed vowel
- the minimum energy of the next word
- the maximum energy of the current word

Results

- The ten best-performing features in experiment 1 included one or more features from each of the four categories as well as features that were phone-normalized.
- Comparing the results of experiment 1 to the results of experiment 2, the accuracy of the tests that used an SVM model are higher than the accuracy of the tests that used an HMM model. This could be due to differences in SVM and HMM or in the feature sets used.
- The results of experiment 3 suggest that cues of prominence exist throughout the word.
- Experiment 4 shows that our hypothesis was correct. By eliminating words which only one rater labeled as 'prominent' accuracy was improved.

Conclusions

- Normalized features improve accuracy beyond raw features, indicating that local changes in acoustic measures, and changes relative to mean values are important cues to prosody.
- Prominence classification based on stressed and unstressed regions results in comparable accuracy, contrary to predictions from the phonological model, where prominence features associate with the stressed syllable.
- Prominence is more accurately classified for words with higher inter-rater agreement. We speculate that data trimming eliminated words erroneously labeled prominent by a single transcriber, and also that higher agreement may occur for words with stronger acoustic cues to prominence.

References

[1] S. Calhoun. Information Structure and the Prosodic Structure of English. PhD thesis, University of Edinburgh, 2006.
[2] 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.
[3] 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.
[4] D. Ladd. Intonational phonology. Cambridge Univ Pr, 2008.
[5] 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.
[6] Y. Mo, J. Cole, and E. Lee. Naive listeners prominence and boundary perception. Proc. Speech Prosody, Campinas, Brazil, pages 735–738, 2008.
[7] J. Pierrehumbert. The phonology and phonetics of English intonation. MIT Cambridge, MA, 1980.
[8] 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.
[9] A. Rosenberg. Automatic Detection and Classification of Prosodic Events. PhD thesis, Columbia University, 2009.

This study is supported by NSF IIS-0703624 to Cole and Hasegawa-Johnson