Accounting for Speaker Variation in the Production of Prominence using the Bayesian Information Criterion

Tim Mahrt[1], *Jennifer Cole*[1], *Margaret Fleck*[3], *Mark Hasegawa-Johnson*[2]

[1]Department of Linguistics, [2]Department of Electrical and Computer Engineering,
[3]Department of Computer Science, University of Illinois, Urbana-Champaign, Illinois

tmahrt2@illinois.edu, jscole@illinois.edu, mfleck@illinoi.edu, jhasegaw@illinois.edu

Abstract

This study investigated speaker variation in the production of various acoustic cues of prominence, including durational and intensity measures. This study stems from our prior work where we used the Bayesian Information Criterion to determine whether each cue were gradiently or discretely associated with prominence. In our prior work, we found that features vary as to whether they are gradient or discrete. In the present study, we found a similar result. For all speakers, some features are gradient and some features are discrete in the manner in which they cue prominence. Under a metrical stress notion of hierarchically layered prominence, our result would suggest that some speakers do not exploit the full range of prominence distinctions offered in English.

Index Terms: speech prosody, prominence, Bayesian Information Criterion, speaker variation, corpus linguistics

1. Introduction

Functionally, prominence is used in English to mark the prosodic structure of a phrase, to distinguish new information from old information, and to introduce contrast. Phonetically, studies have shown that prominence can be realized through an increase in the duration of the stressed vowel, an increase in intensity, and the presence of a pitch accent [1, 2, 3, 4, 5]. Furthermore, there is also evidence that listeners use their expectations of prominence placement to perceive prominence. [6]

There are several ways that one could represent the phonological notion of prominence. One such representation would be prominence as a binary feature, where each word is either prominent or non prominent. Another representation is offered by Metrical Stress Theory [7]. Under metrical stress theory, prosodic units are arranged in a strong-weak patterning and layered on top of each other in intonational phrases, prosodic words, syllables, feet, and mora. In this manner, prominence is also hierarchical and can be phonetically realized gradiently. The results from our previous work, described in detail in the next section, agree with this representation of prominence as hierarchical, however, much of the variance between features in the results remained unaccounted for. One hypothesis, and the point of departure for this study, is that speakers vary in which features they used to cue prominence.

In the present study our research questions are as follows. Across speakers, is there consistent use of correlates of prominence? What phonological model can be used to account for differences across speakers or features in cuing prominence?

1.1. Past Work

In our past work, we collected prominence judgments from a 35,000 subset of the Ohio State Buckeye Corpus [8] using the Rapid Prosody Transcription method–a method used for obtaining word-level prominence judgments from naive, native speakers of English [9]. Teams of 15-20 native speakers of English were used to transcribe short excerpts (15-60s) of the Buckeye Corpus with a binary prominent or non-prominent label, in real time. Then, for each word, we took the number of subjects that labeled the word as prominent and divided it by the number of labels. We term this value as the p-score. Utilizing this method, approximately five hours of data was annotated with p-scores.

In a later study, we investigated the relationship between p-scores and various acoustic measures that have been reported in the literature to be correlated with prominence and we found a positive correlation [4]. In other words, as the number of listeners, who thought a word was prominent increased, the acoustic measures also increased (durational measures became longer, intensity measures became louder, etc.).

1.2. Bayesian Information Criterion

In a subsequent study, we attempted to further refine our understanding of the relation between p-scores and prominence cues [10]. We investigated whether each cue was better modeled by a single Gaussian distribution or two Gaussian distributions. To obtain two distributions from a single cue we divided measures by their associated p-score, as shown in Figure 3. This process was done at every p-score interval. The best model between the single distribution and the many two-distribution models was obtained by calculating the Bayesian Information Criteria (described in detail in Section 2.2.2).

If a cue was best modeled by two distributions, with one distribution associated with low p-scores and non-prominence and the other associated with high p-scores and prominence this would suggest that this cue operated in a binary fashion. In contrast, a single distribution would suggest a gradient notion of prominence, which would be consistent with a metrical stress notion of hierarchically layered prominence. Note that if we expect that these cues all contribute the same information to the perception of prominence and that speakers act in a uniform manner, then we would expect that all features would be best modeled in the same way (i.e. they would all be best modeled by a single distribution or all best be modeled by two distributions). If best modeled by two distributions, would we expect that the p-score threshold dividing those two distributions would be the same for each cue.

In that study, we found that some features were best modeled by a single distribution and others were better modeled by

Feature	R^2	P-Value
Max Intensity of the Stressed Vowel	0.024	0
Min Intensity of the Stressed Vowel	0.024	0
Min Intensity of the Last Vowel	0.039	0
RMS Intensity of the Last Vowel	0.039	0
Max Intensity of the Last Vowel	0.039	0
Stressed Vowel Duration	0.042	0
Duration of the Last Vowel	0.049	0
Log Stressed Vowel Duration	0.057	0
Log Duration of the Last Vowel	0.058	0
Word Duration	0.227	0
Log Word Duration	0.248	0

Table 1: Table showing positive correlation between acoustic features and p-scores.

two. Furthermore, for those features best modeled by two distributions, there were very different p-score thresholds between features. This result suggests that either prominence is a gradient feature or speakers do not utilize the same set of cues, or possibly a combination of these two. It is this outstanding issue that the current study investigates by conducting a BIC analysis on each feature for individual speakers.

2. Methodology and Results

2.1. Features

We began this study by sampling various measures of cues found to be correlated with prominence from the stressed vowel, the last vowel, the whole word, and the following word. One cue that we measured was duration. For calculating the duration of the last vowel and the whole word, we used timestamps provided by the phoneme-level transcriptions in the Buckeye corpus.

As the Buckeye corpus does not contain stress information, additional work was needed to extract the duration of the stressed vowel. Using the International Speech Lexicon (ISLEX) dictionary, which contains phoneme-level dictionary pronunciations with stress markings, we were able to estimate the location of the vowel carrying primary stress and use that phoneme index within the Buckeye phoneme-level transcriptions to calculate the stressed vowel duration.

After calculating these raw duration measures, we created a second set of measures by taking the log values of all of the raw values.

We also calculated the minimum, maximum, and root mean square intensity. The raw intensity was extracted automatically using a praat script which sampled the sound files every millisecond.

Thus, in total we used four durational measures, four log durational measures, and twelve intensity measures for a total of twenty features. After these twenty features were extracted a regression analysis was conducted to confirm that they were indeed positively correlated with p-scores. Any feature that was not correlated was discarded, which left eleven features to be analyzed. Table 1 summarizes the features that were found to be positively correlated with p-scores.

2.2. Methodology

For each feature we compared models containing different partitions of the data produced by individual speakers. Partitions were made as shown in Figure 3. We first considered the original distribution as a distinct model. Then, for sixteen unique p-scores, a value ranging between 0 and 1, we used each p-score as a threshold and split the original distribution using it. All of the feature values associated with a p-score less than or equal to that threshold were placed in one distribution and all of the feature values associated with a p-score higher than that value were placed in another distribution.

This strategy was motivated by the idea that if prominence is binary, then we might expect to have two populations within our feature data, where one population is associated with low p-scores and one population is associated with high p-scores. Furthermore, assuming the feature is binary, we do not know where the threshold should be made, thus, we run our analysis over every unique p-score value.

2.2.1. Fitting Data to a Model

In this study, we investigate whether our data is best modeled by one Gaussian distribution or two. A Gaussian distribution is characterized by a mean and a standard deviation. For a given set of data we can calculate a mean and a standard deviation, thus "fitting" our data to the Gaussian defined by those parameters. It is possible that the data will not all fall within the the Gaussian distribution. Thus, if we split the data in two, we can model the data with two Gaussian distributions, providing a better fit for the data. This process can be continued until the set of Gaussian distributions perfectly represents the data. Note however, that increasing the number of Gaussian distributions increases the complexity of the model.

To find the best model, we can use a log likelihood estimate if the number of distributions in our two models is the same. In this study, however, we compared a model with a single distribution and several models with two distributions. Thus, we used the Bayesian Information Criterion (BIC) (eq. 1).

2.2.2. Bayesian Information Criterion

The problem of determining where to segment p-scores is qualitatively similar, in some ways, to the problem of segmenting meeting-room speech into segments corresponding to different talkers. In both cases, we wish to make as few assumptions as possible, e.g., we do not want to assume that we know how many segments there should be. The problem of speaker segmentation is often solved using a Bayesian Information Criterion (BIC) [11]. The BIC measures the mutual information between the parameters of any given model and the observed data, under the assumption that the parameters themselves are random variables generated by randomly resampling the training data. The BIC thus takes the form of a penalized log likelihood function,

$$BIC(X;\Lambda) = logF(X;\Lambda) - (k/2)ln(n) \qquad (1)$$

where Λ is a parameterized distribution model containing k parameters, and X is a dataset containing n observations. The likelihood $F(X;\Lambda)$ is guaranteed to increase when the dataset is segmented, and separate model parameters are trained using each half of the data. The entropy penalty $(k/2)ln(n)$ measures, in effect, the expected increase in the log likelihood. Thus we can compare two models by computing

$$\Delta BIC = BIC(X;\Lambda_1) - BIC(X;\Lambda_2) \qquad (2)$$

If ΔBIC is positive, it means that model Λ_1 fits X better than Λ_2 by a greater-than-expected amount; if ΔBIC is negative, the improvement in fit is less than expected. This is not a

significance test; $\Delta BIC > 0$ does not mean that Λ_2 is rejected with 95% confidence, it only means that Λ_1 is better.

Within a given feature, after calculating the BIC score for each model, we calculated equation 2, where Λ_2, the baseline, was the model with a single Gaussian distribution. From these ΔBIC scores, the highest score indicates that the associated p-score threshold is the optimal partition point. Note that any ΔBIC score with a value greater than zero suggests that this feature is better modeled by a two-Gaussian distribution. If none of the ΔBIC scores is higher than zero, this suggests that this feature is better modeled by a single Gaussian distribution.

After calculating the optimal BIC partitions, to make patterns more clear, we placed the optimal p-scores into five bins: 0, 0.25, 0.5, 0.75, 1.0 where p-scores were less than or equal to the bin they were placed in. We then observed the data by looking at how speakers produce prominence within each feature (Figure 1). We also inverted the observation and looked at the feature variation within each speaker (Figure 2).

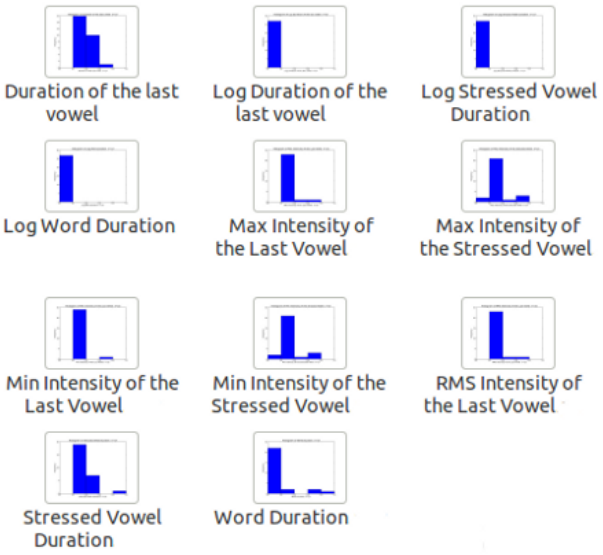

Figure 1: *Histograms of binned p-score thresholds for individual features (n=27, which corresponds to the number of speakers).*

3. Discussion

Considering the data in Figure 1, we see that all features are best modeled by two distributions with a low p-score threshold, with the exception of the log duration measures and the word duration, which are best modeled by a single distribution. Furthermore, we see some variation in p-score threshold in most features.

Considering the data in Figure 2, we see that, across all speakers, there are some features that are best modeled by two distributions with a low p-score threshold. Thus, all speakers make a binary prominent and non-prominent distinction with a low p-score threshold. About half of the speakers also make a binary distinction with a high p-score threshold. At the same time, everyone also has some features that are best modeled by a signle distribution. Thus, all speakers also use cues that are gradiently associated with prominence.

The results presented here do conform with the result found in our prior study. Some of the variance in our previous study

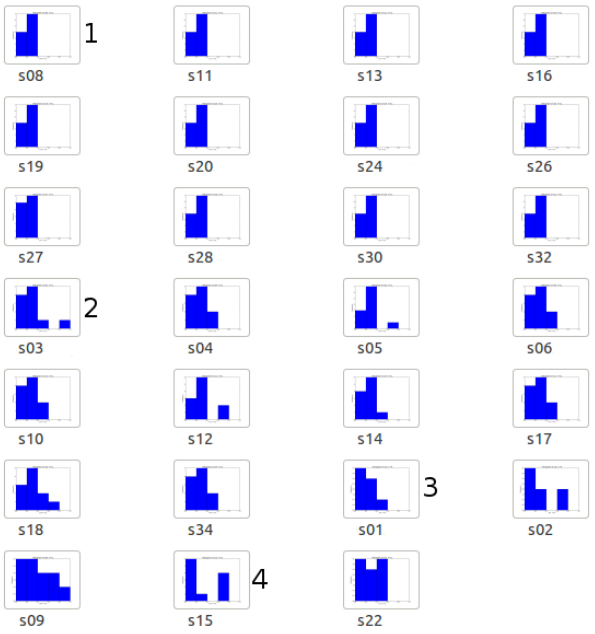

Figure 2: *Histograms of binned p-score thresholds for individual speakers (n=11, which corresponds to the number of features). Speakers are clustered by the group they fall into. 1) Indicates the speaker was best modeled by just two bins (0.0 and 0.25. 2) Indicates 0.25 was the largest bin. 3) Indicates 0.0 was the largest bin. Entries in 4) did not fall into any category.*

may be explained by speaker variation, as speakers do vary to a degree in how they cue prominence. However, by and large, within a feature, speakers are mostly consistent as shown in Figure 1. Thus, different acoustic cues are utilized consistently across speakers as either binary or gradient.

As with our previous work, we found some features were best modeled as binary across a low threshold, binary across a high threshold, and gradient. Our results for individual speakers confirms that no one uses a uniformly gradient or uniformly binary strategy in the production of prominence. If we consider a metrical stress notion of prominence, where prominence is hierarchically layered, one possible way to account for the plurality of strategies is to consider that some speakers "flatten" the hierarchy. In other words, they do not fully exploit the possible range of prosodic levels. If we consider the nuclear stress to be one extreme on a "prominence continuum" and an unstressed word that carries given information on the other end, our results would suggest that not all speakers are utilizing the levels between these two extremes.

More work is needed to investigate the reality of prominence in the mind of the speaker. The role of the listener as a source of variance also needs to be investigated.

4. Conclusion

In this study we have attempted to refine our understanding of variance across speakers in the production of prosodic prominence. Our findings show that while there is some variation in how speakers cue prominence, these variations are not fundamentally different–all speakers cue prominence using features that are gradiently associated with prominence and features that are binarily associated with prominence. Our result suggests

that prominence is gradient, but variation in our results suggests that not all speakers are exploiting all of the possible levels of the prosodic hierarchy. Further studies will continue to investigate the issue of variation in prominence production.

5. Acknowledgements

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.

6. References

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

[2] Ladd, D.R., "Intonational phonology", Cambridge University Press, 2008.

[3] Calhoun, S., "Information structure and the prosodic structure of English", University of Edinburgh, 2006.

[4] Mo, Y., Cole, J. and Hasegawa-Johnson, J., "How do ordinary listeners perceive prosodic prominence? Syntagmatic vs. paradigmatic comparison", Poster presented at the 157th Meeting of the Acoustical Society of America, Portland, Oregon, 2009. Columbus, OH: Department of Psychology, Ohio State University, 2007. Retrieved March 15, 2006, from www.buckeyecorpus.osu.edu.

[5] Cole, J., Mo, Y. and Hasegawa-Johnson, M., "Signal-based and expectation-based factors in the perception of prosodic prominence", Laboratory Phonology, 1, 425452, 2010.

[6] Wagner, P., "Great expectations–Introspective vs. perceptual prominence ratings and their acoustic correlates", In Ninth European Conference on Speech Communication and Technology, ISCA, 2005.

[7] Hayes, B., "Metrical stress theory: Principles and case studies", University of Chicago Press, 1995.

[8] Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and et al., "Buckeye corpus of conversational speech (2nd release)".

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

[10] Mahrt, T., Huang, J.T., Mo, Y., Fleck, M., Hasegawa-Johnson, M. and Cole, J., "Optimal models of prosodic prominence using the Bayesian information criterion", Eleventh Annual Conference of the International Speech Communication Association

[11] Moschou, V., Kotti, M., Benetos, E. and Kotropoulos, C., "Systematic comparison of BIC-based speaker segmentation systems", IEEE Workshop on Multimodal and Multimedia Signal Processing, 2007.

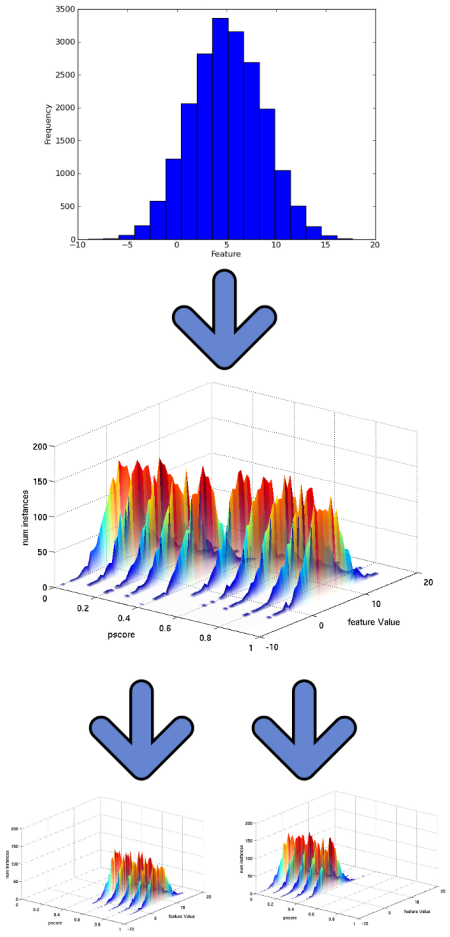

Figure 3: *Schematic of partitioning features using randomly generated data. Note that the example data contains data sampled from two distinct Gaussian populations. Further note that these two distributions cannot be seen in the 1D histogram. The first step to creating two distributions for use in the BIC analysis is to pair together feature values with their associated p-scores. This can be visualized in a 2D histogram. From here, we choose some p-score threshold (e.g. 0.4). All of the feature values associated with a p-score less than or equal to that threshold are isolated in a separate distribution from those feature values associated with a p-score greater than the threshold.*