Optimal models of prosodic prominence using the Bayesian information criterion

Tim Mahrt[1]*, Jui-Ting Huang*[2]*, Yoonsook Mo*[1]*,*
Margaret Fleck[3]*, Mark Hasegawa-Johnson*[2]*, and Jennifer Cole*[1]

[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, jhuang29@illinois.edu, ymo@illinois.edu,
mfleck@illinoi.edu, jhasegaw@illinois.edu, jscole@illinois.edu

Abstract

This study investigated the relation between various acoustic features and prominence. Past research has suggested that duration, pitch, and intensity all play a role in the perception of prominence. In our past work, we found a correlation between these acoustic features and speaker agreement over the placement of prominence. The current study was motivated by a need to enrich our understanding of this correlation. Using the Bayesian information criterion, we show that the best model for a feature that cues prosody is not necessarily a single Gaussian. Rather, the best model depends on the feature. This finding has consequences for our understanding of the role of these features in the perception of prosody and for prosody recognition systems.

Index Terms: prosody, prominence, Bayesian Information Criterion

1. Introduction

Prosody serves an important role in signaling the phrasing and information structure of an utterance, through the assignment of prosodic phrase boundaries and prominence. This study investigates acoustic encoding of prominence in American English spontaneous speech, and the perception of prominence by listeners. Past studies have found evidence that a number of acoustic features contribute to the perception of prominence, including duration, intensity, and F0. However, these studies do not always agree on which acoustic features contribute to the perception of prominence or to the degree that they contribute [1, 2, 3, 4, 5, 6].

Kochanski et al. ran a Bayesian classifier over acoustic features extracted from a corpus of read and spontaneous speech [1]. The task was to classify whether a word was prominent or not based on these acoustic features, and classification accuracy was measured against expert prosody transcription. They found that loudness provided the best classification accuracy. A phone-level duration measurement provided the second-best results while, comparatively, f0 was not useful for classifying prominence.

Wagner investigated the role of of the listener's expectation of prominence in its perception [4]. To do this she conducted an experimental study where subjects marked words in a text on a sliding scale of prominence while hearing the text at a normal rate and then at an accelerated rate and finally while reading the orthography. It was observed that the fast speech had nearly flat-F0 contours and little durational variation. From the high agreement seen across all three tests, Wagner concludes that speakers of English perceive prominence based on their expectations in the absence of acoustic cues. Wagner also found that predictions of prominence aligned with the duration measurements and the f0 measurements, suggesting that f0 is useful for the perception of prominence, counter to the findings of Kochanski et al.

The present study looks at some of these features that cue prosody and attempts to understand the role each feature plays in the perception of prominence as a phonological, contrastive feature.

1.1. Past Work

In our previous work, we collected prominence judgments for a 35,009 word subset of the Buckeye Corpus of spontaneous speech [7], using Rapid Prosody Transcription, a method we developed for prosodic labeling of spontaneous speech [8]. Excerpts were transcribed for prosodic prominence by teams of 15-20 naive speakers of English. Their task was to label each word as prominent or non-prominent in real time as they listened to short (15-60 s) excerpts. For each word, we added up the number of labelers who labeled the word as prominent and then divided this sum by the number of labelers for that word. We call this value the p-score. Figure 1 shows the distribution of p-scores over all words in our data set. Notice that the distribution is skewed, with most words having very low p-scores (the p-score value 0 represents 40% of the distribution).

In a later study, we looked for a relation between acoustic features and the ratings of prominence and found a positive correlation between them; as acoustic values become more extreme (e.g. duration becomes longer or intensity becomes higher) the number of listeners who rate the word as prominent increases as well [5]. The results of that study show that p-scores (i.e., measures of the likelihood that a word will be perceived as prominent) co-vary with word frequency and the acoustic cues to prominence.

1.2. Research Questions and Hypotheses

In linguistic models, prominence is encoded through the assignment of a pitch accent to a word based on its position in the hierarchical metrical structure, and/or based on its status related to information structure [2, 3]. This yields a categorical distinction between prominent words that bear a pitch accent and non-prominent words that are unaccented. Previous studies show that there are multiple correlates of prominence, so

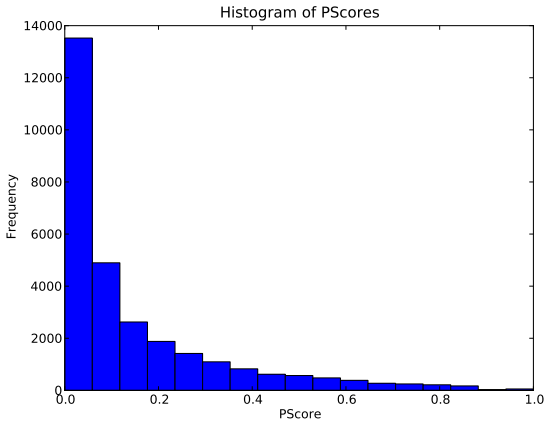

Figure 1: *The distribution of p-scores.*

we ask if each correlate provides a discrete distinction between prominent (accented) and non-prominent (unaccented) words. Our first hypothesis is that each correlate of prominence defines two distributions, one corresponding to low p-scores and one corresponding to high p-scores. If so, then the correlate provides a means of marking the contrast between prominent and non-prominent words, by providing a criterion for the p-score threshold between the two distributions. This finding alone would not confirm that prominence is a binary feature, but it would lend evidence that the prominence feature is discrete. Furthermore, if there are p-score thresholds for multiple correlates, but located at different values along the p-score continuum (0-1), that would suggest either that the prominence distinction is not discrete, or that there are more than two levels of prominence. Our second hypothesis then is that the distributions of values for all correlates of prominence will share the same p-score threshold.

2. Methodology and Results

2.1. Features

Previous studies show that the perception of prominence is multi-dimensional, with listeners taking into consideration various acoustic features, as well as their own expectation as to where prominence should occur, based on the perceived syntactic, semantic, and pragmatic properties of the utterance. For this study, we examine the relationship between perceived prominence on one hand and a subset of features found to cue prominence on the other hand. We test the hypothesis that values of each feature comprise two distributions, with one cluster for words with low p-score values (words that listeners are less likely to judge as prominent), and one cluster for words with high p-score values (words that listeners are more likely to judge as prominent).

For the durational measures, we extracted timestamps from the phoneme-level transcriptions provided by the Buckeye corpus. The word duration was calculated by taking the difference of the timestamps that marked the beginning and end of the word utterance.

For the stressed vowel duration, we first needed to find the stressed syllable, which is not labeled in the Buckeye Corpus. Using the International Speech Lexicon (ISLEX) dictionary, which contains phoneme-level dictionary pronunciation with stress markings, we estimated the location of the vowel carrying primary stress in the transcriptions in the Buckeye Corpus. With the location of this vowel, we were able to determine its duration from the phoneme-level timestamps provided by the Buckeye corpus. The word frequency was taken from Google's unigram dictionary which is a list of word frequencies extracted from a corpus of over one trillion words. We have included word frequency because in a previous study we showed that there was a negative correlation between word frequency and prominence [6].

Pre-word pause was included as a model of disfluencies, which have been found to be correlated with the introduction of new information (and thus prominence). Arnold et al. investigated the relation between disfluencies and word expectations in listeners [9]. Given a partial phrase such as "Click on the uhh red...", subjects were asked to choose the next word. They were given the choice between a new word or a word already in the discourse. They found that the presence of a disfluency causes the listener to expect the introduction of a new word. As prominence is used to introduce new items into the discourse, we include pre-word pause duration (considered as a kind of disfluency) as a feature in this study. In addition to indicating disfluency, pauses in speech mark the prosodic phrase boundaries. In English, the final word in a prosodic phrase often bears nuclear pitch accent and as such, prominence and the post-word pause phrase boundaries often coincide. For this reason, we consider the post-word pause in our set of features. Both the pre-word pause and the post-word pause are calculated from the phone-level timestamps between words in the Buckeye corpus.

After collecting the values for these features, the log transform was taken for each value. Values of 0 were discarded, which in the case of the pause durational measures constituted a significant proportion of the distribution. Due to the annotation scheme used in Buckeye where the last timestamp for one word often coincides with the start timestamp of the next word, there is no pause value between approximately 80% of the words. As a result, pauses are under-reported in the transcription.

2.2. Methodology

We compared different partitions for each of the following individual features: word duration, stressed vowel duration, word frequency, intensity, pre-word pause duration, and post-word pause duration. We partitioned the features into two halves at sixteen different thresholds based on their associated p-score. Thus, the left partition would contain feature values for words with fewer prominence judgments and the right partition would contain feature values for words with greater prominence judgments. We also considered the original unpartitioned feature sets, corresponding to a model where the given feature does not yield a discrete prominence distinction.

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) [10]. 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 likeli-

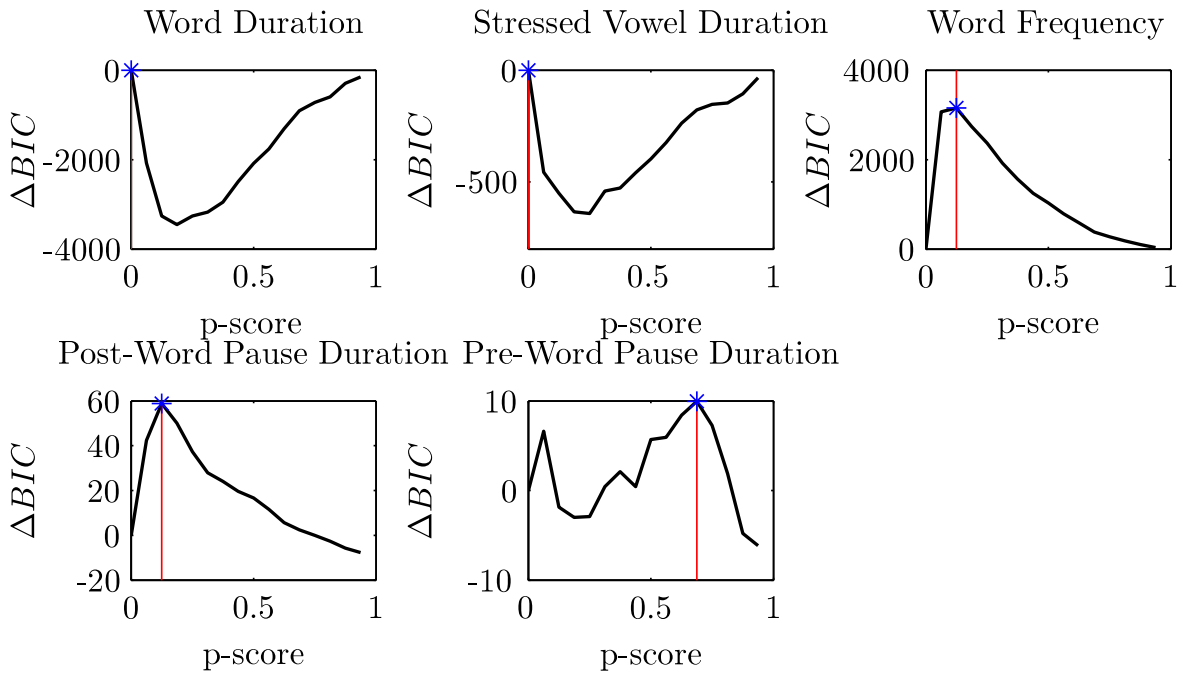

Figure 2: *These graphs plot the p-score threshold as a function of the ΔBIC. BIC scores are calculated from the Gaussian distributions created by the p-score threshold. A p-score of zero corresponds to the ΔBIC score for the single Gaussian model. The largest BIC score on each plot is marked with a '*' and a vertical line.*

hood function,

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

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

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

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.

For each model, the BIC outputs a score. To compare scores, for each feature we computed the difference between each BIC score with the BIC score of the single Gaussian model. The BIC score difference as a function of the p-score is rendered in figure 2. The model with the highest score for each feature is considered to be the optimal model for that feature. The optimal models are summarized in table 1.

Feature	Model	P-score threshold
Word duration	Single Gaussian	N/A
Stressed vowel duration	Single Gaussian	N/A
Word frequency	Two Gaussians	0.125
Post-word pause duration	Two Gaussians	0.125
Pre-word pause duration	Two Gaussians	0.6875

Table 1: Summary of optimal models for prosodic features.

3. Discussion

The first thing to note about the results (figure 2 and table 1) is the lack of uniformity in the distribution of BIC scores. While some features do share the same p-score threshold, the distribution of BIC values for each feature is different. Moreover, some features are better modeled by a single Gaussian while others are better modeled by two.

Given the distribution of p-scores (figure 1), with about 40% of the values equal to 0.0, we would hypothesize that the optimal partitions (if any) would act to separate very low-valued p-scores from the high-valued p-scores where the high-valued p-scores would be more strongly correlated with increased feature values. This expectation is realized in word frequency and post-word pause duration which not only peak at relatively low p-scores but the BIC scores decrease monotonically as the p-score increases. Furthermore, for these two features, the single Gaussian is one of the worst models of the distribution– partitioning the distribution almost anywhere will yield a better result than not partitioning it.

For the other features, however, we found very different results. For the word duration and stressed vowel duration, the shape of the curve is an inversion of word frequency and post-word pause duration, flipped over the x axis, with partitions at a low p-score providing the worst split. Instead, the optimal distribution contains only a single Gaussian. However, note that a partition at a very-high p-score would have almost the same BIC score as a single Gaussian. This suggests that it may be valid to model these features with two Gaussians (a near-optimal model).

The distribution of BIC scores for the pre-word pause du-

ration is the most varied of all the features. It is also the only distribution where the best partition point is at a high p-score.

It should also be noted for all five features that they have a local maximum BIC score at a low-valued p-score. In the case of the word duration, stressed vowel duration, and pre-word pause duration, that local maximum is not the optimal partition point. Regardless, this local maximum aligns with our predictions that there is a meaningful distinction between words with very small prominence ratings and words with larger prominence ratings.

Our first hypothesis stated that each feature will contain two distributions, divided by some p-score. This would reflect a binary prominence feature, as proposed in a linguistic model that equates prominence with the presence of pitch-accent. As we have seen, our hypothesis was validated in three of the five features. For the word duration and stressed vowel duration, we cannot say with certainty that one Gaussian is the correct model–more investigation is needed to uncover the optimal model for these distributions.

Our second hypothesis stated that all features will contain the same partition point. Considering only word frequency, post-word pause duration, and pre-word pause duration, the three features that were best modeled by two distributions, we can see that they do not share the same partition point, disproving our second hypothesis.

How can we explain the difference in the distribution of these features along the p-score continuum? One possible difference corresponds to different listeners' sensitivities. If we assume that the prominence feature is binary, the data could still be explained if the perception of prominence is idiosyncratic. For example, one listener may be more attuned to the duration of the stressed vowel, while another listener may be more attuned to the intensity of the word. If such a situation were to apply, the results of this study could be influenced by the presence of speakers with very different strategies.

On the other hand, if we assume that the perception of prominence is consistent across listeners, then the prominence feature cannot be binary but may instead be gradient. Otherwise, we would expect all features to have the same partition point. In other words, if features had the same partition point, then we would end up with the same two distributions of p-scores for every feature, regardless of the feature value. We would expect these two distributions to correspond to non-prominence and prominence. Thus, if we assume that the perception of prominence is consistent across listeners, then in order to account for the difference in partition point and number of distributions, we would conclude that the prominence feature is gradient. An alternative account would be that there are multiple discrete prominence features which together form multiple "levels" of prominence.

It is also possible, and in our view more likely, that the reality is some combination of speaker variation and a gradience of prominence. If the perception of prominence is purely idiosyncratic, then prominence would not be useful for the purposes of carrying linguistic information. On the other hand, there is evidence for some level of idiosyncracy. In our previous work, we showed that the acoustic correlates of prominence vary from speaker to speaker. In that study, we ran a linear regression for each speaker, predicting P-scores from a set of acoustic measures. We found that the contribution of each acoustic variable to the variability in P-scores varied from speaker to speaker[5].

Although our second hypothesis about a uniform prosody threshold across features was disproved, the results are still consistent with the phonological model of prosody. The participants in our study were making binary decisions as to whether or not a word was prominent, but as noted above, this distinction may draw on multiple distinct types of prominence, e.g., contrastive focus, new information focus, or purely 'rhythmic' prominence. More work is needed to investigate whether the perception of prominence is gradient or binary.

In future studies, we will apply the methodology of BIC analysis to investigate more correlates of prominence and will investigate the relationship between regression models and the distributions found in this study.

4. Conclusions

Using the BIC we were able to show that some features known to be correlates of prominence are better modeled by two Gaussians rather than one, that features have very different BIC score distributions, and that the optimal point to partition a feature varies depending on the feature. From our results we can conclude that listeners vary in their sensitivity to different cues to prominence or that the prominence feature is gradient or some combination of these two. Further work conducted on individual speakers and listeners may shed more light on this situation.

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

[5] 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.

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

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

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

[9] Arnold, J.E., Kam, C. L. H., Tanenhaus, M.K. and Arnold, J., "If you say thee uh-you're describing something hard: The on-line attribution of disfluency during reference comprehension", Journal of Experimental Psychology: Learning, Memory, and Cognition, vol 33, 5, 914-930, 2007.

[10] 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.