F0 and the Perception of Prominence

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@illinois.edu, jhasegaw@illinois.edu

Abstract

This study investigates the role F0 plays in the perception of prominence in American English. Raw, log and locally normalized measures of F0 were extracted from words in a 35K word corpus of spontaneous speech. Linear regression analyses were conducted to test the strength of these measures as cues to prominence, with prominence based on judgments made by ordinary listeners in real-time auditory perception. The Bayesian Information Criterion was used to further investigate whether these F0 measures cue prominence in a linear or piecewise linear function, corresponding to a linguistic model of prominence as a gradient or discrete feature. The results of this study show that F0 measures are similar to intensity measures in both their strength as cues to perceived prominence, and in signaling a discrete prominence distinction that distinguishes non- or weakly-prominent words from words with greater prominence. Our finding that F0 and intensity cue a discrete prominence distinction is compared with our prior finding that duration and word frequency signal gradient prominence distinctions. This apparent discrepancy is discussed in terms of the dual nature of prominence in English, as an expression of layered metrical (stress) structure in phonology, and as an expression of pragmatic focus.

Index Terms: speech prosody, prominence, Bayesian Information Criterion, F0

1. Introduction

This study investigates the role of F0 as an acoustic cue to the perception of prominence through the statistical analysis of a prosodically annotated corpus of spontaneous English speech. Prosodic prominence in English expresses metrical stress, and is also used to signal focused constituents, including narrow contrastive focus, and broad or new-information focus. A number of acoustic features have been reported as correlates of prominence including duration, intensity, and F0, however, the precise contribution of each these cues is disputed [1, 2, 3, 4]. For example, Turk and Sawusch found that while both duration and intensity cue prominence, intensity does not have an independent role from duration [5]. The role of F0 as a cue to prominence, expressing the presence of pitch accent, is widely discussed in the literature [6]. Kochanski and colleagues investigated prominence in a corpus of British English and found that while duration and intensity are important cues to prominence, F0, was not a reliable cue [7].

Another issue regarding prominence is the nature of the prominence distinctions that are encoded in phonological and phonetic form. One possibility is that prominence is a discrete binary feature, where words are either prominent or are not prominent. Another view of prominence is offered under metrical stress theory, where prosodic units are layered on top of each other in a hierarchical fashion [8]. This model yields a multi-layered, gradient notion of prominence, where the degree of prominence of an element depends on its depth of embedding in the metrical (stress) tree structure of the phonological phrase to which it belongs.

In our previous work, outlined in the next section, we investigated the role of a number of acoustic measures related to duration and intensity in the perception of prominence. We found that cues based on these measures do contribute to the perception of prominence, and furthermore, that there is significant variation among cues in the nature of the prominence distinction they signal, with several cues signaling gradient prominence distinctions In the present study, we extend our analysis to F0 cues to prominence. Our research questions are: to what extent does F0 contribute to the perception of prominence? Is F0 best modeled as a cue to a discrete or gradient prominence distinction? How does F0 compare to other prominence cues, in terms of its cue strength and in relation to models of prominence distinctions as discrete or gradient?

1.1. Past Work

The present study continues our work investigating the role of various acoustic features in the perception of prominence. To measure the degree to which a word is considered prominent we use p-scores, which represent prominence judgments obtained using Rapid Prosody Transcription, whereby naive subjects listen to a text and transcribe in real-time the words that they hear as prominent [9]. In this study we analyze a 35,000 word subset

of the Ohio State Buckeye Corpus, drawn from 27 speakers of American English [10]. Prominence annotation is based on transcriptions obtained from groups of 15-20 native speakers of English who transcribed short (15-60s) excerpts from the Buckeye corpus. Transcribers labeled each word they heard as being either prominent or not prominent, a binary distinction. For each word, the number of prominence judgments were summed and divided by the number of annotators; this is the p-score. Using this method, approximately 5 hours of speech data were annotated with p-scores. For the analyses presented here, each word in our prosodically annotated corpus has a vector of features associated with it, including the p-score, word frequency, and measures of duration, RMS intensity and F0 taken from the word and the stressed syllable.

Using the p-scores that were obtained using Rapid Prosody Transcription and a set of acoustic measures reported in the literature to correlate with prominence, a series of regression analyses were conducted to determine which word-level acoustic features from this corpus cue prominence as perceived by ordinary listeners [11]. The results show that despite substantial variation among speakers in their phonetic implementation of prominence, there is a strong positive correlation between p-scores and these acoustic features.

1.2. Bayesian Information Criterion

Given that the acoustic cues to prominence vary along continuous dimensions of duration, intensity and F0, we ask whether the prominence distinctions they signal are correspondingly gradient. In our recent and ongoing work, we pursue this question using the Bayesian Information Criterion to examine the role of acoustic features in the perception of prominence [12]. Specifically, for each cue to prominence we want to know if it acts as a cue to a discrete binary prominence distinction, modeled by two Gaussian distributions, or as a cue to a single gradient prominence distinction, modeled by a single Gaussian distribution. In order to obtain the two-Gaussian model, we pair the individual feature values with their associated p-score and divide the acoustic values into two groups based on some p-score threshold. Sixteen different p-score thresholds were used, corresponding roughly to the number of labelers that were used in gathering the p-scores. Using the Bayesian Information Criterion, the optimal model for each cue was selected from these sixteen two-Gaussian models and the one single-Gaussian model.

In the case that a two-Gaussian model is found to be optimal for a given cue, this suggests that the cue encodes a binary notion of prominence, with low values of that cue being associated with weak or non- prominence and high values of that cue being associated with prominence. On the other hand, a single-Gaussian model suggests that this cue works in a gradient fashion: as the cue value increases, so do p-scores, indicating a stronger degree of perceived prominence. In our prior study we conducted this analysis with measures of Word Duration, Stressed Vowel Duration, Word Frequency, Post-Word Pause Duration, and Pre-Word Pause Duration and found that different cues had different optimal models, with some cues optimally modeled as a single distribution co-varying with p-scores, and other cues optimally modeled in terms of two distributions, one co-varying with low p-scores, and the other co-varying with higher p-scores.

Taken as a whole, our prior results with duration and intensity measures suggest either (i) that speakers utilize different cues in the production of prominence, (ii) that prominence is gradient, with each cue contributing differently, or (iii) that there are different types of prominence that are cued differently. In our past BIC analyses we considered segment duration measurements, word frequency, pause duration measurements, and intensity measurements. In the present study we extend our BIC analysis onto F0 to investigate the role that F0 has in the perception of prominence.

2. Methodology and Results

2.1. Features

We extracted raw F0 values using the ESPS pitch tracker. F0 measures were taken from the whole word, the stressed vowel, the final vowel, and the following word. Portions of these domains where the pitch was 0, such as during the production of a voiceless consonant, were not considered for analysis. Phone-level time stamps within the Buckeye corpus were used to automatically extract values from the whole word and from the vowel of the final syllable. The stressed vowel is of interest in the analysis of prominence, as pitch accents are assigned to stressed vowels, which are also the locus of prominence at the word and phrase levels. Since stressed vowels are not labeled as such in the Buckeye corpus, the stressed vowel of each word (excluding unstressed function words) was located using the International Speech Lexicon (ISLEX) dictionary, which contains phoneme-level dictionary pronunciations as well as stress markings.

For each word we extracted the minimum, maximum, and average F0 values from the whole word, the final-syllable vowel, and the stressed vowel. These were analyzed as raw and log-transformed F0 measures for a vector of 24 unique features. For these 24 features, we applied z-normalization to each value in a window of length 3, consisting of a target value and the value before and after it, and a window of length 5, consisting of a target value, the two values preceding it and the two values following it. For example, the Max F0 measurement of the word was normalized in a window of 3 and 5 words,

Feature	Log	Norm.	r^2
Max F0 of the Stressed Vowel	No	0	.022
Mean F0 of the Last Vowel	Yes	5	.024
Mean F0 of the Last Vowel	No	5	.025
Mean F0 of the Last Vowel	Yes	3	.025
Mean F0 of the Last Vowel	No	3	.025
Max F0 of the Last Vowel	No	3	.028
Max F0 of the Last Vowel	Yes	3	.029
Max F0 of the Last Vowel	Yes	5	.029
Max F0 of the Last Vowel	No	5	.030
Mean F0 of the Stressed Vowel	No	3	.044
Max F0 of the Stressed Vowel	No	3	.045
Mean F0 of the Stressed Vowel	Yes	3	.046
Max F0 of the Stressed Vowel	Yes	3	.048
Mean F0 of the Stressed Vowel	Yes	5	.049
Mean F0 of the Stressed Vowel	No	5	.049
Max F0 of the Stressed Vowel	No	5	.052
Max F0 of the Stressed Vowel	Yes	5	.056

Table 1: *Table showing F0 features with a positive correlation between acoustic features and p-scores through Pearson's r where $r^2 > 0.02$. All correlations were found to be statistically significant with $p < 0.05$.*

centered on the target word, while the Max F0 measurement of the stressed vowel was normalized in a window of 3 and 5 stressed vowels, centered on the target stressed vowel. Our final data set contains 24 unnormalized features, 24 normalized in a window length of 3, and 24 normalized in a window length of 5, for total set of 72 features.

We then conducted a series of regression analyses on these features to determine which were significantly (positively) correlated with prominence. Features that had an r^2 of 0.02 or higher (with $p < 0.05$) were considered for further analysis using the Bayesian Information Criterion.

2.2. Results

The main questions we ask in this paper are 1) what is the utility of F0 in cueing prominence, which entails understanding the best way to process F0 and 2) is F0 a binary or gradient cue to prominence, which brings in comparison with features that we have investigated in our prior study.

2.3. Correlation between F0 and Prominence

To determine the utility of F0 as a cue to prominence we consider the regression analysis (Table 1) which contains the features that were found to be significantly (and positively) correlated to prominence. Here we see that only maximum and mean measurements are present, leaving out minimum measurements. Similarly whole-word measurements do not cue prominence well, although measurements extracted from the last vowel cue prominence to a small extent and measurements from the stressed vowel cue prominence about twice as well as those from the last vowel.

Feature	Threshold
Log word frequency	Low
Duration of the last vowel	Low
Max Intensity of the last vowel	Low
RMS Intensity of the last vowel	Low
Min Intensity of the last vowel	Low
Min Intensity of the stressed vowel	Low
Log Mean F0 of the stressed vowel (normalized)	Low
Log Max F0 of the stressed vowel (normalized)	Low
Stressed vowel duration	High
Max F0 of the stressed vowel (normalized)	High
Mean F0 of the stressed vowel (normalized)	High
Word Duration	None
Log word Duration	None

Table 2: *Table showing different models for some cues to prominence. A threshold of 'low' means an optimal p-score partition of .188 or lower. A 'high' threshold means an optimal p-score partition higher than .188. A threshold of 'none' means that the optimal model for that feature is a single Gaussian distribution.*

Perhaps surprisingly, (unnormalized) log F0 measures are not strong predictors of prominence. However, normalized and unnormalized raw F0 measures, as well as normalized log F0 measures, are well represented in the best F0 cues to prominence.

The most important salient finding is the importance of normalization. All but one of the best F0 cues to prominence were normalized across some window. This does not come as a surprise as the F0 values for a word will be dependent to a degree on the pitch of the phrase it is carried in. The difference between r^2 values for features normalized in a window of length 3 and a window of length 5 is small (although a window of length 5 appears to be slightly better).

2.4. The Bayesian Information Criterion and F0

Table 2 shows a sample of cues that correlate with prominence in our corpus and a relative p-score threshold for the associated optimal model. A cue with a low or high p-score threshold is best modeled by two Gaussian distributions, separated by a relatively low or high p-score value. No p-score threshold indicates that the cue is best modeled by a single-Gaussian model. Considering just the F0 cues, the normalized F0 measures had either a low or a high p-score threshold, with the log F0 measures having a low p-score threshold and the raw F0 values tending to have a high p-score threshold. Unnormalized measures, which had a much lower correlation, were

scattered across all three categories. In comparison, the word duration measures were best modeled by a single Gaussian distribution while the Intensity measures were mostly best modeled by a two Gaussian distributions with a low p-score.

3. Discussion

What do these results tell us about the nature of prominence? With different features best modeled in different ways, the results cannot be explained by a binary model of prominence (prominent vs non-prominent). Word duration, the strongest cue to prominence in our corpus, is best modeled by a single distribution. Intensity measures are best modeled by two-Gaussian distributions with a low p-score, as are normalized log F0 measures. These findings suggest that there are at least two different ways that prominence is cued. The evidence for a third distinction, two-Gaussian distributions with a high p-score threshold, is less certain. Although normalized raw F0 values did fall into this range, since F0 is perceived on a log scale, we may attribute greater importance to log F0 as a perceptual cue to prominence. With few other features in this third cue set (two-Gaussian with a high p-score threshold) contributing to the perception of prominence, there is little evidence justifying the existence of a unique high p-score threshold-based model.

Regardless, the distribution of cues to prominence and their associated optimal models does not support a strictly binary model of prominence. Instead, it suggests that prominence is gradient, or alternately, that prominence is not binary but makes a small number of distinctions (e.g. a three-way prominence distinction), or that different kinds of prominence, such as contrastive focus or new information, utilize different acoustic cues. This matter will be investigated in future studies.

4. Conclusions

Comparing several measures of F0 values against a measure of prominence, we found that F0 does contribute to the perception of prominence, particularly when the F0 measurement has been taken from the stressed vowel of a word, is on a log scale, and has been normalized against a local context. Using the Bayesian Information Criterion we found that these log F0 values are best modeled by two Gaussian distributions, with a low p-score threshold. Considering all of the cues together, prominence cannot be a binary feature, though more work is needed to better understand the nature of prominence in relation to rhythmic structure and pragmatic focus.

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

[2] Mo, Y., Cole, J. and Hasegawa-Johnson, M., "Prosodic effects on vowel production: Evidence from formant structure", Proceedings of Interspeech 2009, Brighton, UK.

[3] 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.

[4] Fant, G., Kruckenberg, A. and Liljencrants, J., "Acoustic-phonetic analysis of prominence in Swedish", Intonation. Analysis, Modelling and Technology. Kluwer Academic Publishers, 55-86, 2000.

[5] Turk, A.E. and Sawusch, J.R., "The processing of duration and intensity cues to prominence", Journal of the Acoustical Society of America, 99(6), 1996.

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

[7] 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.

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

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

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

[11] 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.

[12] 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", Proceedings of Interspeech 2011, Florence, IT.