Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception

Yoonsook Mo

Department of Linguistics
University of Illinois, Urbana-Champaign
ymo@uiuc.edu

Abstract

I investigate the acoustic correlates of prosodic prominence and boundary, as they are perceived by naïve listeners, in spontaneous speech from American English (Buckeye corpus). Prosodic prominence and phrasing serve different functions in speech communication: prosodic phrase boundaries demarcate speech chunks that typically cohere semantically, while prominences encode focus and possibly also rhythmic structure. The acoustic correlates of prominence and phrase boundary are examined through measures of vowel duration and overall intensity of stressed vowels, to see how those measures correlate, individually or in combination, with naïve listeners' perception of prominence and boundary. The results show that most stressed vowels are lengthened in pre-boundary words (i.e., those final in the prosodic phrase). Prosodic prominence is also cued by increased duration, but in combination with higher overall intensity for some vowels. These acoustic differences associated with perceived prominence and boundary suggest different mechanisms underlying their production. This claim finds support from consideration of the different functions that prominence and boundary play in encoding information structure, and in speech production planning.

1. Introduction

Prosodic structure encodes phonological phrasing and sentence-level prominence. In English, the words of an utterance are grouped into hierarchically layered phrasal constituents in which phrasal stress is located and marked with a pitch accent on top of a structurally defined lexical stress. Prosodic structures are phonetically encoded in patterns of pitch, duration, and loudness modulation, and in spectral features related to phone quality. In order to recover the syntactic, semantic and pragmatic content of an utterance, the listener must recognize the prosodic context from the acoustic correlates of prosody in the speech signal. Research concerned with how prosody encodes structure and meaning, and how listeners perceive the prosodic features in speech must first consider the nature of the acoustic encoding of prosody. This paper takes up that question, looking at the acoustic correlates of prosodic prominence and boundaries in stressed vowels of American English.

Pitch, the perceptual correlate of fundamental frequency, is traditionally regarded as a primary cue for prominence in many languages, including English, and previous studies have examined the relation between fundamental frequency (F0) and prominence perception. Gussenhoven et al. [3] examine the relation in Dutch between F0 maxima on accented syllables (i.e., intonationally marked syllables with phrasal stress) and F0 minima as the reference for the baseline of pitch range, and shows that the perception of prominence is related to the relative distance between pitch peaks as well as the degree of declination of the baseline.

Findings from other studies challenge the view that pitch is the most salient cue for prosodic prominence, although definitions of prominence vary in each study. Silpo and Greenberg [10] claim that F0 plays a minor role in the assignment of phrasal stress in American English, based on analysis of phrasal stress as marked by two linguistically trained labelers using three stress levels. Sluijter and van Heuven [11] investigate the hierarchical relations among various acoustic correlates of lexical (i.e., word-level) stress and accent in Dutch and American English and find that overall intensity as well as F0 are the primary correlates of focal accents (also termed phrasal stress), while duration and glottal parameters are correlates of both stress and focal accent. Kochanski et al. [7] also evaluate acoustic correlates of perceived prominence in varieties of British English, using a prominent/ non-prominent judgment classifier. The results show that prominence is coded by loudness and duration but various types of F0-related measurements play only a minor role. Heldner [5] examine how reliable overall intensity and spectral emphasis are to the perception of focal accents in Swedish. He finds that both overall intensity and spectral emphasis are reliable acoustic correlates, but spectral emphasis is more effective in separating focal from non-focal words.

There are also multiple studies that investigate the acoustic correlates of prosodic phrase boundary. Several studies report effects of prosodic boundaries on the duration of phrase-final segments and on pause duration in American English [1]. However, other studies report acoustic properties other than duration correlated with prosodic boundaries. Kim et al. [6] indicates that American English speakers differentiate two levels of phrase boundary (ip vs. IP) on the basis of final F0 at the rime end and glottalization as well as duration. Working with the same corpus of American English broadcast news speech, Choi et al. [2] find that all voice source measurements except pitch are correlated with perceived boundaries, and are relevant for automatic prosodic boundary detection.

I have examined acoustic correlates of prosodic prominence and boundary in the Midland Cities variety of American English [8] through measurements of F0, duration, pause, spectral tilt and overall intensity, and report the results of acoustic duration and overall intensity in the present study. This study differs from previous work in that the prosodic features that are associated with acoustic measures are collected from a large number of untrained listeners, "naïve" to the issues related to prosodic analysis, who have identified words in conversational speech as prominent or phrase-final in a real-time transcription task.

2. Methodology

2.1. Labeled transcripts

74 naïve listeners participated in two runs of the prosody transcription experiment (hereafter, Experiments 1 and 2), and were assigned to one of two groups within each experiment. Groups differed in the order of the transcription tasks (prominence and boundary marking). The materials are spontaenous speech excerpts selected from the Buckeye corpus of American English, with speakers from Columbus, Ohio [9]. Two excerpts (about 20 seconds long) from each of 18 speakers are extracted, totaling 36 excerpts. Transcribers mark the locations of prominence and boundary on words in a printed transcript, in real time as they listen. Transcribers are not guided by any visual display of the speech signal.

Each excerpt is transcribed by between 15 and 22 naïve listeners in separate tasks of prominence and boundary labeling. Transcriptions from all transcribers are pooled, and each word is assigned a probablistic prominence score and a probablistic boundary score that codes the number of transcribers who marked that word as prominent or final in a "chunk" (prosodic phrase). Fleiss' multi-raters' kappa coefficients and their corresponding z-scores are used to assess multi-transcriber agreement. All groups show significantly high agreement rates by z-score, showing consistent patterns of prosody perception (refer to Mo et al. in this volume of the proceedings of speech prosody).

2.2. Acoustic measurements

The waveforms for each excerpt were aligned with word and phone transcriptions, and measures of duration (ms) and overall rms intensity (dB) were extracted for all vowels. All the measures were z-normalized by speaker and by phone label. From each word I extracted the primary or secondary stressed vowels as identified in a reference dictionary [4], in order to analyze the effects of prominence (i.e., phrase-level stress) and phrase boundary while holding lexical stress constant. Table 1 shows the distribution of stressed vowels in the subset of the corpus used here.

Table 1: *The distribution of stressed vowels*[1]

vowel	aa	ae	ah	ao	aw	ay	eh
Freq.	81	129	211	58	28	140	187
vowel	er	ey	ih	iy	ow	uh	uw
Freq.	66	114	209	156	103	41	94

To analyze the acoustic correlates of perceived prominence on these vowels, I established three models of prosodic feature assignment that differ in the classification of words as prominent or phrase-final (the "boundary" condition) according to the prominence and boundary score the word receives from the pooled transcriptions. The critical scores are listed below (P for prominence, B for boundary), followed by the model descriptions in Table 2.

- Not-Prominent/Not-Boundary: Words which no transcribers marked as P/B.
- One-Prominent/One-Boundary: Words for which one or more transcribers marked a P/B.
- Half-Prominent/Half-Boundary: Words which at least one but fewer than half the transcribers marked as P/B.
- Prominent/Boundary: Words which half or more transcribers marked as P/B.

Table 2: *Three models for classification of words into groups by prosody scores*

model name	*prosody classification*
Model A	Not-P/Not-B *vs.* Half-P/Half-B *vs.* P/B
Model B	Half-P/Half-B *vs.* P/B
Model C	Not-P/Not-B *vs.* One-P/One-B

3. Results

One-way ANOVAs were performed to test for effects of prominence and boundary features (as perceived by naïve listeners) on acoustic measures. Separate ANOVAs were run for each of the three models from Table 3, to compare effects of prominence and boundary under each of the grouping criteria in the three models.

3.1.1 Effects of prominence

One-way ANOVAs showed that for most vowels, duration is significantly different according to the prominence classification of the word, as summarized in Table 3. *p*-values are reported in each cell and cell shading codes the level of statistical significance: duration differences that reach significance at α=0.05 appear in grey, those that are marginally significant are light grey, and those that are not significant remain white.

Among 14 vowels, the four vowels /aa, ao, er, uh/ show no effect of prominence on duration for any of the three models of prosody classification. For vowels that showed a main effect of prominence, post-hoc tests confirm that vowels classified as prominent are longer than vowels which are not prominent. Models A, B, and C produced only slightly different statistical results; while most vowels showed effects of prominence on the stressed vowel duration in all three models of prominence grouping, results for /ow/ and /uw/ differed depending on the classification model.

3.1.1 Effects of boundary

To examine the durational effects of boundary on stressed vowels, one-way ANOVAs were again conducted for each of the three models. Table 4 displays *p*-values for each vowel under each model. As seen, the durations of stressed vowels are significantly different depending on the phrase boundary condition. Duration effects on /ao, uh, uw/ were not significant under any classification model, while /aa/ showed significant duration effects under Model C but only marginal effects under Models A and B. Post hoc tests showed that all the vowels were lengthened at the preboundary location excluding the vowel /ao/ in Model B and /uh/ in Models A and B.

[1] The corresponding IPA symbols of vowels are: [ɑ, æ, ʌ, ɔ, aʊ, aɪ, ɛ, ɝ, eɪ, ɪ, i, oʊ, ʊ, u], in order.

Table 3: *p-values from ANOVA results for durational effects of prominence by three models of grouping (as in Table 2)*[2]

vowels	aa	ae	ah	ao	aw
Model A	.471	.003	.001	.363	.016
Model B	.537	.015	<.001	.154	.005
Model C	.423	.002	.001	.628	.178
vowels	ay	eh	er	ey	ih
Model A	<.001	<.001	.370	.001	.026
Model B	<.001	.028	.375	.009	.021
Model C	<.001	<.001	.180	.002	.043
vowels	iy	ow	uh	uw	
Model A	.003	.051	.749	.016	
Model B	.008	.023	.853	.539	
Model C	.005	.856	.447	.004	

Table 4: *p-values from ANOVA results for durational effects of boundary by three models of grouping (as in Table 2).*

vowels	aa	ae	ah	ao	aw
Model A	.059	<.001	<.001	.411	.001
Model B	.094	.034	<.001	.902	<.001
Model C	.019	<.001	<.001	.326	.034
vowels	ay	eh	er	ey	ih
Model A	.001	<.001	<.001	<.001	<.001
Model B	<.001	<.001	.003	<.001	<.001
Model C	.001	<.001	<.001	<.001	<.001
vowels	iy	ow	uh	uw	
Model A	.006	<.001	.633	.178	
Model B	.013	<.001	.426	.116	
Model C	.004	<.001	.912	.115	

3.1. Overall intensity meansurements

3.2.1 Effects of prominence

Effects of perceived prominence on the overall intensity of stressed vowels are shown in Table 5. Only the vowels /aa, eh/ showed an effect of prominence on overall intensity under all three models, while /ay, ih/ showed main effects in two of the three classification models and /er/ in one model. The vowels /ay, er/ also showed a marginally significant effect of prominence in Model B and in Model A, respectively. The three vowels /iy, ow, uw/ showed a marginally significant main effect of prominence only in Model C. Post hoc tests with vowels which reached statistical significance in overall intensity showed that vowels perceived as prominent were louder than vowels not perceived as prominent.

3.2.2 Effects of boundary

Results from one-way ANOVAs testing the effect of boundary on vowel intensity are shown in Table 6. Only the vowel /iy/ shows a main effect of boundary on intensity in all three models. The vowels /ao, ow/ show effects of boundary in one model with an additional marginal effect for /ao/ in Model A and for /ow/ in Model A and B. The vowel /ay/ also showed a marginal effect of boundary on overall intensity in Model A and C. However, counter to the effects of prominence on overall intensity, vowels in pre-boundary position are associated with a decrease in overall vowel intensity.

[2] Significant effects are shaded for $\alpha = 0.05$ in all the tables.

Table 5: *p-values from ANOVA results for intensity effects of prominence by three models of grouping (as in Table 2)*

Vowels	aa	ae	ah	ao	aw
Model A	.045	.236	.312	.720	.236
Model B	.039	.247	.187	.991	.692
Model C	.048	.484	.252	.461	.144
Vowels	ay	eh	er	ey	ih
Model A	.021	.021	.087	.213	.015
Model B	.060	.043	.033	.158	.004
Model C	.011	.014	.193	.531	.744
Vowels	iy	ow	uh	uw	
Model A	.187	.166	.185	.666	
Model B	.453	.173	.250	.753	
Model C	.068	.097	.080	.458	

Table 6: *p-values from ANOVA results for intensity effects of boundary by three models of grouping (as in Table 2)*

Vowels	aa	ae	ah	ao	aw
Model A	.979	.403	.340	.088	.470
Model B	.851	.184	.836	.017	.215
Model C	.860	.326	.201	.187	.493
Vowels	ay	eh	er	ey	ih
Model A	.085	.914	.576	.248	.295
Model B	.659	.916	.589	.101	.838
Model C	.088	.774	.294	.267	.142
Vowels	iy	ow	uh	uw	
Model A	.010	.054	.612	.616	
Model B	.014	.083	.375	.312	
Model C	.008	.016	.418	.612	

4. Discussion

This study finds robust effects of perceived prosody, including prominence and phrase boundary, on stressed vowel duration, with less consistent effects on stressed vowel intensity. Most stressed vowels that are categorized as prominent under one or more prominence models are significantly longer than those that are not. Similarly, most stressed vowels that are categorized as being in pre-boundary position in one or more models were significantly longer than the ones that are not. In other words, most stressed vowels do show statistically significant effects of both prosodic prominence and prosodic boundary on their duration values. Moreover, durational effects of prominence and boundary on stressed vowels are observed across all three models of prominence and of boundary classification. In relation to the effects of prominence, all stressed vowels except /aa, ao, er, uh/ show significant durational effects of prominence in at least two models. Concerning the effects of boundary, all stressed vowels except /ao, uh, uw/ demonstrate significant durational effects of boundary. Therefore, the results of durational measurements indicate that duration is a robust correlate of perceived prosodic prominence and phrase boundary. However, the fact that no single model is clearly superior indicates that individuals vary quite a bit in their sensitivity to acoustic indices of prominence and suggests that the acoustic evidence from duration alone is not a sufficient basis for modeling prominence judgments of the sort we analyze here.

Regarding effects of prominence and boundary on the overall intensity of stressed vowels, normalized rms overall

intensity measures were significantly correlated with prominence only for some stressed vowels. Vowels /aa, ay, eh/ showed significant or marginally significant effects of prominence on overall vowel intensity in all three models, while the vowels /er, ih/ show prominence effects in only two models. There were no significant prominence effects on overall intensity for vowels /iy, ow, uh/, though the intensity measures for these vowels were also marginally significantly increased under prominence in Model C. On the other hand, being in pre-boundary position has no significant effect on the overall intensity of stressed vowels in most cases. Only the vowel /iy/ exhibits a boundary effect on intensity in all models, while the vowels /ao, ay, ow/ show boundary effects in one or two models. More interestingly, post hoc tests indicate that overall intensity is elevated in words peceived as prominent, whereas intensity is diminished in words perceived to be in pre-boundary position. The asymmetry in these intensity results could be one reason why we do not observe significant effects of boundary on overall intensity. In our data, many nuclear prominences occur on the last word in the prosodic phrase (i.e., rightmost in phrase), which means that listeners perceived both prominence and boundary on those particular words. The enhancing effects of prominence might weaken the diminishing effects of boundary on overall intensity. Comparing the three models of prosody classification, there was no particular model that showed more consistent effects on intensity than other models of prominence or boundary classification, which is similar to what was observed for the effects on duration.

There are several implications that stem from these results. First, longer durations are correlated with both prominence and boundary perception, which means that duration on its own is an ambiguous cue to prosodic context. Second, unlike duration, overall intensity may distinguish between the two prosodic contexts, since intensity is higher in prominent words, but lower in phrase-final words. However, intensity is a less reliable cue for prosodic context, given the inconsistency of prosodic effects on intensity across vowels. Third, neither duration nor overall intensity serves as a sufficient cue for prosodic context over all 14 stressed vowels. For instance, prominence affects duration but not overall intensity of the stressed vowels /ae, ah, aw, ey, iy, ow, uw/. In contrast, the vowels /ay, eh, ih/ show significant effects of prominence and boundary on both duration and overall intensity while /ao, uh/ show no significant prosodic effects on either duration or overall intensity. The inconsistency of prosodic effects across vowels suggests that prosodic context may be signaled by multiple acoustic cues interacting with one another, rather than by a single cue that is invariant to vowel context.

5. Conclusion

This study shows that the duration and overall intensity of stressed vowels are correlated with naïve listeners' perception of prosodic prominence and boundary. Although duration effects from prosody are more consistent across vowels than are intensity effects, neither duration nor intensity alone provides reliable cues to prosodic context across vowels. More likely, listeners integrate information from multiple acoustic cues to prosody in judging the prosodic context from acoustic information. This study used graded prominence and boundary scores for each word, reflecting the perception of prosodic context from multiple transcribers. This method of coding prominence and boundary allows multiple ways of classifying words for prominence and location relative to a prosodic boundary, which differ in the threshold score for prominence and boundary classification, and also in the number of distinct prominence or boundary classes. This study compared three distinct models of prosody classification, but found no evident advantage in any single model. These findings call for future research to better understand the strength of acoustic factors in influencing prosody perception, and to identify additional factors that may help us understand the variation across listeners.

6. Acknowledgements

This research is supported by NSF grants IIS 07-03624 and IIS 04-14117 to Jennifer Cole and Mark Hasegawa-Johnson. I thank them and Chilin Shih, Margaret Fleck, Eun-Kyung Lee for advice and assistance.

7. References

[1] Chavarria, S; Yoon, T-J.; Cole, J.; Hasegawa-Johnson, M., 2004. Acoustic differentiation of ip and IP boundary level: Comparison of L- and L-L% in the Switchboard corpus. *Proceedings of ISCA* (Nara, Japan).

[2] Choi, J-Y.; Hasegawa-Johnson, M.; Cole, J., 2005. Finding intonational boundaies using acoustic cues related to the voice source. *Journal of the Acoustical Society of America.* 118 (4), 2579-2587.

[3] Gusssenhoven, C.; Rietveld, A. C. M, 1988. Fundamental frequency declination in Dutch: testing three hypotheses. *Journal of Phonetics*. 16, 355-369.

[4] Hasegawa-Johnson, M.; Fleck, M., ISLE Dictionary version 0.2.0," 2007, downloaded Oct. 19, 2007 from http://www.isle.uiuc.edu/dict/index.html

[5] Heldner, M., 2003. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*. 31, 39-62.

[6] Kim, H.; Yoon, T-J.; Cole, J.; Hasegawa-Johnson, M., 2006. Acoustic differentiation of L- and L-L% in Switchboard and Radio news speech. *Proceedings of ISCA* (Dresden, Germany).

[7] Kochanski, G.; Grabe, E.; Coleman, J.; Rosner, B., 2005. Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America.* 118 (2), 1038-1054.

[8] Labov, W.; Ash, S.; Boberg, C., 2006. The Atlas of North American English. Mouton de Gruyter (New York).

[9] Pitt, M.A.; Dilley, L.; Johnson, K.; Kiesling, S.; Raymond, W.; Hume, E.; Fosler-Lussier, E., 2007. *Buckeye Corpus of Conversational Speech* (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

[10] Silpo, R.; Greenberg, S., 2000. Prosodic stress revisited: Reassessing the role of fundamental frequency. *Proceedings of the NIST Speech Transcription Workshop* (College Park, MD).

[11] Sluijter, A. M. C.; Heuven, V. J. van, 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proceedings of ICSLP* (Philadelpia, PA). 630-633.