Acoustic Differentiation of L- and L-L% in Switchboard and Radio News Speech

Heejin Kim[1], *Tae-Jin Yoon*[1], *Jennifer Cole*[1] *& Mark Hasegawa-Johnson*[2]

Department of Linguistics[1]; Department of Electrical and Computer Engineering[2]
University of Illinois at Urbana-Champaign, U.S.A.
{hkim17; tyoon; jscole; jhasegaw}@uiuc.edu

Abstract

Acoustic evidence for a distinction between low-toned intermediate (ip) and intonational phrase (IP) boundaries is presented from two speech corpora representing spontaneous, conversational speech and scripted broadcast speech. Robust effects of the two boundary levels are found in the phrase-final syllable rime in both corpora. Nucleus duration is longer and the F0 value at rime end is lower at IP boundaries compared to ip boundaries. Glottalization is also more frequent before an IP boundary. Other effects of boundary level on the F0 and intensity contours over the phrase-final rime are evident but variable across the two corpora. These findings support the Beckman-Pierrehumbert theory of intonation [1] in its recognition of two levels of prosodic phrasing.

1. Introduction

At least two levels of prosodic phrasing have been widely assumed in the description of prosodic structure offered by linguists and speech scientists. More recently, Beckman & Pierrehumbert [1] propose that there exists a level of phrasing between the prosodic word and the intonational phrase (IP), identified as the intermediate phrase (ip) in English and the accentual phrase (AP) in Japanese. As pointed out by Ladd [2][3], however, there is scant empirical evidence for the intermediate phrase in English, compared with the Accentual Phrase for Japanese [1][4]. Subsequent research provides evidence for differentiating ip from IP based on articulatory measures [5], acoustic duration [6], and the perceptual judgment of voice quality [7]. There remains little evidence for a distinction in phrase level in terms of F0 or intensity or their perceptual correlates in pitch and loudness, even though prosodic structure is generally defined in terms of categories that are instantiated by pitch and intensity as well as duration.

Our earlier study of prosodic phrasing in the Switchboard corpus of conversational telephone speech provided corroborating evidence of final lengthening and occurrence of creakiness in spontaneous speech, with greater lengthening and more frequent occurrence of glottalization at low-tone intonational phrase boundaries (L-L%) than at low-tone intermediate phrase boundaries (L-) [8]. However, we failed to find evidence that changes in pitch and intensity serve to differentiate between L- and L-L%. The lack of evidence for a phrase level distinction from pitch and intensity contours raises both theoretical and practical questions: (1) Are the acoustic correlates of phrase level in spontaneous speech different from those found in the kind of read speech that has formed the empirical basis for much prior work? (2) Is the distinction between ip and IP phrase boundaries made through acoustic cues other than pitch and intensity? If ip and IP phrase boundaries are not effectively distinguished by local pitch features, that would diminish the argument for Pierrhumbert's tone-sequence model over the superpositional model of intonation (cf. [9]).

In this paper, we present a further analysis of the acoustic differentiation of ip and IP boundaries comparing data from two speech corpora representing spontaneous, conversational speech and scripted broadcast speech. First, we present evidence for the acoustic differentiation of the two levels of intonational boundaries in the Switchboard corpus, based on locally normalized measures of F0 and intensity in the phrase-final rime, and on acoustically and perceptually identified glottalization. We find significant differences in F0 at the rime end, in peak rime intensity, and in the frequency of creak occurrence in the preboundary rime for the two boundary levels, L- and L-L%. Second, our parallel study of Radio News speech shows the same significant differences between L- and L-L% in nucleus duration, F0 at rime end, and frequency of creaky voice, and shows an additional effect of boundary level on F0 drop and F0 slope. The effect of boundary on intensity varied depending on speaker and voice quality. Taken together, our Switchboard and Radio News results indicate that preboundary lengthening, phrase-final F0, and the frequency of creak occurrence consistently differentiate boundary levels across speakers, in both scripted and non-scripted speech. On the other hand, the patterns of F0 and intensity variation over the phrase-final syllable rime signal boundary strength only for some speakers, and do not achieve the status of general characteristics of boundary strength.

2. Speech Corpora

2.1. Switchboard WS97 Corpus

Our analysis is based on the files in the WS97 subset of Switchboard [10], annotated with ToBI labels marking pitch accents, phrase accents, and boundary tones. This is the same set of ToBI-labeled WS97 files used in our earlier study comparing acoustic correlates for L- and L-L% [8].

2.2. Radio News Corpus

The speech used in this analysis is the lab news portion of the Boston University Radio News corpus [12]. This corpus includes ToBI labeling (cf. [11]) and word-level transcriptions for all files. As in the Switchboard labeling, the pitch accent inventory was collapsed into H* and L* for this study. We analyzed all files for two of the six speakers (F1A and F2B).

3. Methods

For the comparison of preboundary syllable nuclei between L- and L-L% tokens, we normalized vowel durations using the corresponding phone-aligned transcriptions for each corpus. Because the WS97 subset has at most two or three short files for each of the 79 speakers included in our analysis, duration was normalized across all speakers for the Switchboard corpus. The large data set available for each Radio News speaker made it possible to normalize duration within speaker for that corpus.

For all tokens, the rime beginning was hand-labeled based on the spectrogram and waveform view in Praat [13], and the rime end was marked at the end of the sonorant portion of the syllable. Tokens were divided into 'plain' and 'creaky'. Creaky tokens were identified when the pitch track failed or was in error over the sonorant portion of the rime, and when auditory impression along with visual inspection of the waveform and spectrogram indicated that pitch track error could have been caused by creak. Tokens with pitch track errors unrelated to creak were excluded from the F0 analysis. Unlike in [8], creaky and plain tokens were analyzed separately to allow for the lower intensity that characterizes creak. Because our preliminary examinations showed that F0 and intensity are affected by the presence and type of pitch accent (PA) on the final rime, tokens were categorized according to presence and type of pitch accent on the preboundary word. Distribution of tokens according to boundary types, pitch accent types and voice quality is shown for Switchboard in Table 1 and for Radio News in Table 2[1]. Although it is possible that some of our IP boundaries coincide with the end of higher-level domains such as the utterance, our analyses do not take into account domain levels higher than the phrase.

Table 1: *Distribution of L- and L-L% tokens (Switchboard)*

Boundary	Pitch Accent	Plain	Creak
L-	H*	106	3
	L*	7	2
	No PA	92	12
	Total	205	17
L-L%	H*	60	15
	L*	5	4
	No PA	22	11
	Total	87	30

Table 2: *Distribution of L- and L-L% tokens (Radio News)*

		Speaker F1A		Speaker F2B	
Bnd	PA	Plain	Creak	Plain	Creak
L-	H*	54	7	46	38
	L*	2	0	1	1
	No PA	19	5	10	3
	Total	75	12	57	42
L-L%	H*	43	85	55	136
	L*	1	2	2	12
	No PA	0	19	10	37
	Total	44	106	67	185

For F0 (Hz) and intensity (dB) comparisons, the following values were extracted:

[1] A few tokens that had been present in [8] were corrected while we checked again the validity of the labels of the Switchboard data.

Beginning F0 For preboundary syllables with a H* pitch accent, beginning F0 was measured at the accent peak. For non-pitch-accented syllables, beginning F0 was measured at the rime beginning.

Beginning intensity Because maximum intensity in the rime is usually not reached until some point after start of the rime beginning, beginning intensity was measured at the point of peak intensity in the rime.

End F0 and end intensity These measurements were taken at the end of the sonorant portion of the rime.

F0 drop and intensity drop F0 drop is equal to end F0 minus beginning F0, and intensity drop is equal to end intensity minus beginning intensity. Bigger negative values indicate greater magnitude of drop.

F0 slope This is the F0 drop divided by the duration of the interval from beginning F0 to end F0.

Again, because Switchboard is a multi-speaker corpus and the WS97 subset includes only a small data set for each speaker, normalization of F0 and intensity was necessary. One possible reason we couldn't get reliable results in [8] from the F0 and intesnity analyses may be that we failed to control the variation that were present in speaker's pitch range. Based on Patterson's [14] discussion of pitch range modeling, the domain for F0 normalization was defined over the individual utterance as delimited by the beginning and ending of the WS97 file. Whereas our previous F0 and intensity comparisons [8] were based on normalization over all of a speaker's turns in the conversation, the more locally based normalization took into account paralinguistic factors such as attitude or emotion that can affect the level and span of a speaker's pitch range [14]. Outliers due to pitch tracking errors such as doubling or halving were manually identified and corrected to avoid artificially compressed or expanded pitch range values.

In our statistical analysis, we compared L- and L-L% according to F0 drop magnitude, F0 slope, and intensity drop magnitude. We also compared beginning and end F0 and intensity values to determine whether there are more localized differences that might be obscured in looking only at the change over the whole rime. Since unaccented preboundary syllables were rare in the Radio News corpus and L* preboundary syllables were rare in both corpora, analyses of those items (unaccented in Radio News and L* in both) are not reported here.

4. Results

4.1. Duration

4.1.1. Switchboard Corpus

For both boundary types, the normalized nucleus duration is in general longer than the mean values, as reported in [8]. The increased preboundary nucleus duration is a general phrase boundary cue, while the significantly greater duration at L-L% compared to L- ($F(1, 313) = 15.748$, $p<.001$) indicates that degree of preboundary lengthening differentiates levels of phrasing across speakers.

4.1.2. Radio News Corpus

For both speakers, the preboundary nucleus durations of L-L% are greater than those of L-, and the difference is significant (F1A: $F(1, 245) = 20.969$, $p<.001$; F2B: $F(1, 362) = 7.967$, $p<.01$). The box plots in Figure 1 show the difference in nucleus duration between L- and L-L% for speaker F1A.

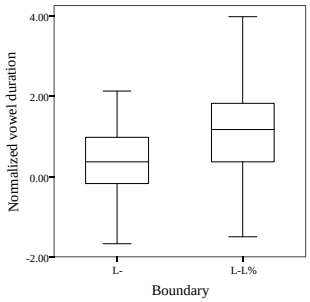

Figure 1: *Box plots for normalized preboundary nucleus duration (Radio News, speaker F1A)*

4.2. F0

4.2.1. Switchboard Corpus

Although L- tends to have a higher rime beginning F0 than L-L%, the difference is not significant. Rime end F0, however, is significantly lower for L-L% than for L- ($F(1, 276) = 7.597$, $p<.01$), indicating that end F0 is used across speakers to differentiate these two boundary types. The box plots in Figure 2 show the rime end F0 means for L- and L-L%.

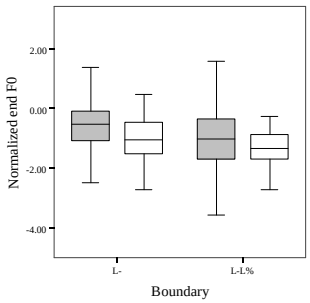

Figure 2: *Box plots for end F0, with gray bars for H* and white bars for no accent tokens (Switchboard)*

4.2.2. Radio News Corpus

Both speakers show significant differences between the two boundary levels in several F0 measures. F0 at the rime end is lower and F0 drop and slope are greater for L-L% than for L- (for end F0, F1A: $F(1, 90) = 20.371$, $p<.001$; F2B: $F(1, 94) = 19.316$, $p<.001$, for F0 drop, F1A: $F(1, 90) = 10.824$, $p<.05$; F2B: $F(1, 94) = 8.124$, $p<.01$, and for F0 slope, F1A: $F(1, 90) = 4.929$, $p<.01$; F2B: $F(1, 94) = 7.789$, $p<.01$). Beginning F0 is not different between the two boundary levels for either speaker. The box plot of end F0 for speaker F2B is shown in Figure 3.

4.3. Intensity

4.3.1. Switchboard Corpus

For plain tokens, intensity at the rime end is not significantly different between the two boundary types, but beginning intensity is significantly lower for L-L% than for L- ($F(1, 276) = 12.769$, $p<.001$). This indicates that beginning intensity is another of the acoustic features used to differentiate L- from L-L%. The box plots in Figure 4 show the means for beginning intensity for both L- and L-L%. No intensity differences are found for creaky tokens.

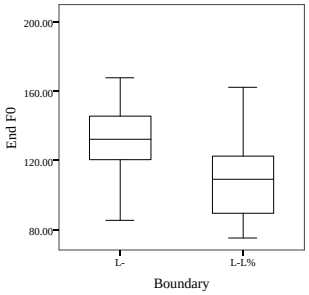

Figure 3: *Box plots for end F0 (Radio News, speaker F2B)*

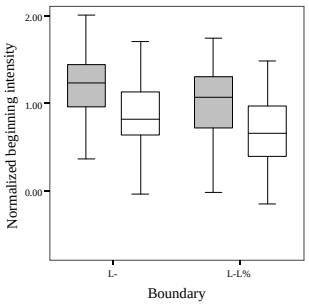

Figure 4: *Box plots for beginning intensity, with gray bars for H* and white bars for no accent tokens (Switchboard)*

4.3.2. Radio News Corpus

For plain tokens, speaker F1A shows no significant difference between the two boundary types for any measurement of intensity. On the other hand, speaker F2B shows a significant difference in beginning intensity (Fig. 5) and end intensity between L- and L-L%, with lower intensity values for L-L%, (beginning intensity; $F(1, 94) = 13.899$, $p<.001$; end intensity; $F(1, 94) = 10.344$, $p<.01$). Although L-L% has a greater magnitude of intensity drop than L-, the difference is not significant.

For creaky tokens, end intensity is significantly lower for L-L% for both speakers (F1A: $F(1, 85) = 6.605$, $p<.05$; F2B: $F(1, 166) = 14.455$, $p<.001$) and speaker F2B shows an additional difference in intensity drop ($F(1, 166) = 6.773$, $p<.05$), with a greater drop in L-L%.

4.4. Voice quality: creak

4.4.1. Switchboard Corpus

As in [8], frequency of creak occurrence is greater for L-L% than for L-. This supports and supplements our previous finding for creak distribution [8], which included only those cases of creaky voice that were identified through complete pitch tracking failure; the current analysis includes other creaky tokens

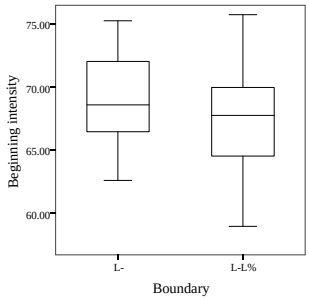

Figure 5: *Box plots for beginning intensity (Radio News, speaker F2B)*

that result in pitch tracking errors of doubling and halving.

4.4.2. Radio News Corpus

Creak is observed to be markedly more frequent in L-L% tokens than in L- tokens, as in Table 3.

Table 3: *Frequency of creak occurrence (Radio News)*

	Percentage of creak	
Boundary	Speaker F1A	Speaker F2B
L-	13.79% (12/87)	42.42% (42/99)
L-L%	70.67% (106/150)	73.41% (185/252)

5. Discussion

This study finds acoustic evidence for differentiation of low-toned intermediate and intonational boundaries in both spontaneous and read speech in American English. Duration measures of the nuclei of preboundary syllables show that preboundary lengthening differentiates boundary levels across speaking styles, corroborating previous findings [1][2]. The analysis of F0 and intensity contours in Switchboard reveals significant differences between L- and L-L% for only two measurements: F0 at rime end and peak rime intensity. On the other hand, the Radio News data show significant differences in F0 drop, F0 slope, and F0 at rime end. Beginning and end intensity measurements of plain tokens differentiate boundary types for speaker F2B, with no intensity differences for speaker F1A. For creaky tokens from Radio News, we find a significant difference in intensity at rime end in both speakers, but intensity drop differs only for speaker F2B. In addition, glottalization is more frequent at intonational boundaries in both corpora. Observing that it is not beginning F0, but F0 at rime end that differs between L- and L-L%, we note that the effect of boundary tone seems to be localized at the phrase edge, as suggested by [9][11][15]. The local expression of the phrase accent (L-) and boundary tone (L%) provides clear evidence in support of the tone-sequence model, and more generally, of the claim that prosodic phrases are encoded through demarcative features positioned at the end of those constituents.

Our results indicate that preboundary lengthening and F0 values, and the frequency of glottalization differentiate boundary levels across speakers, although no single feature is likely to serve as an effective classifier. Our speaker-dependent analysis of Radio News speech shows that speakers may vary in the prosodic features they use to mark boundary level distinctions, and we expect that a similar in-depth look at individual speaker data in Switchboard would result in additional findings of boundary level differentiation for conversational speech.

6. Acknowledgements

This work was funded through the University of Illinois Critical Research Initiative and through NSF award number IIS-0414117. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF or University of Illinois.

7. References

[1] Beckman, M.; Pierrehumbert, J., 1986. Intonational structure in Japanese and English, *Phonology Yearbook* 3, 255-30.

[2] Ladd, D.R., 1986. Intonational phrasing: the case for recursive prosodic structure, *Phonology Yearbook 3*, 311-340.

[3] Ladd, D.R., 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[4] Pierrehumbert, J.; Beckman, M., 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

[5] Fougeron, C.; Keating, P., 1997. Articulatory strengthening at edges of prosodic domains, *Journal of the Acoustical Society of America* 101(6), 3728-3740.

[6] Wightman, C.; Shattuck-Hufnagel, S.; Ostendorf, M.; Price, P.J., 1992. "Segmental durations in the vicinity of prosodic phrase boundaries", *Journal of the Acoustical Society of America*, 91(3), 1707-1717.

[7] Redi, L.; Shattuck-Hufnagel, S., 2001. Variation in the rate of glottalization in normal speakers. *Journal of Phonetics* 29, 407-427.

[8] Chavarría, S.; Yoon, T.; Cole, J.; Hasegawa-Johnson, M., 2004. "Acoustic differentiation of ip and IP boundary levels: Comparison of L- and L-L% in the Switchboard corpus", *Proceedings of the International Conference on Speech Prosody*, Nara: Japan, 333-336, 2004.

[9] Ladd, D.R., 2000. Bruce, Pierrehumbert, and the Elements of Intonational Phonology, In *Prosody: Theory and Experiment*, M. Horne (ed.). Dordrecht: Kluwer, 37-50.

[10] Godfrey, J.; Holliman, E.; McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development, *Proceedings of the International Conference on Audio, Speech and Signal Processing*, 517-520.

[11] Beckman, M.; Ayers, G., 1997. *Guidelines for ToBI Labelling* (version 3.0). ms., The Ohio State University.

[12] Ostendorf, M.; Price, P.; Shattuck-Hufnagel, S., 1995. *The Boston University Radio News Corpus*, from <http://www.ldc.upenn.edu>.

[13] Boersma, P.; Weenink, D., *Praat: A system for doing phonetics by computer*, Computer Program available at <http://www.praat.org>, 1992-2005.

[14] Patterson, D., 2000. *A linguistic approach to pitch range modelling*, PhD Thesis, University of Edinburgh.

[15] Pierrehumbert, J., 2000. Tonal elements and their alignment, In *Prosody: Theory and Experiment*, M. Horne (ed.). Dordrecht: Kluwer, 11-36.