Acoustic Differentiation of ip and IP Boundary Levels: Comparison of L- and L-L% in the Switchboard Corpus

Sandra Chavarría[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.
{chavarri; tyoon; jscole; jhasegaw}@uiuc.edu

Abstract

Prosodic phrase boundaries, regardless of level of disjuncture, can be signaled by variation in pitch, loudness, and final-syllable length. In an attempt to find acoustically distinctive characteristics correlated with ip (intermediate phrase) versus IP (intonation phrase) labels in a ToBI-labeled subset of the Switchboard corpus, we compared F0 drop, intensity drop, and nucleus duration in the phrase-final rime for L- and L-L% boundary labels. The results indicate no significant difference in F0 or intensity drop, but final-syllable lengthening as measured by nucleus duration differed significantly between the two boundary levels. Additionally, F0 aperiodicity associated with creaky voice was found to occur more frequently at L-L% than at L- boundaries. These results provide empirical corroboration of the statement that F0 does not reliably differentiate L- from L-L% [1], and support previous findings that degree of final-syllable lengthening [3] and presence of creaky phonation [10] are correlated with differences in perceived level of phrasal disjuncture.

1. Introduction

As noted in the Guidelines for ToBI labeling [1], it is often difficult to differentiate between a low intermediate (L-) tone and a low intermediate + low intonation phrase tone sequence (L-L%) based on F0 alone, because fundamental frequency at both of these boundary types is near the bottom of the speaker's range. This makes deciding between L- and L-L% "more subjective," such that "transcribers must rely on the percept of degree of disjuncture with less help from the F0 contour" (p. 33). Although other phrase boundary characteristics such as lowered intensity, final-syllable lengthening in the pre-boundary word, and post-boundary pausing [3] contribute to the impression of disjuncture, it is not clear whether these acoustic correlates differ in degree between intermediate and intonation level boundaries such that two distinct categories are in fact produced in spontaneous speech.

This paper reports on a study of the acoustic correlates of L- and L-L% boundaries, based on ToBI-labeled files from the WS97 subset of the Switchboard corpus[1]. Acoustic correlates of L- and L-L% labeling were examined by comparing normalized duration of the nucleus and normalized pitch and intensity drop over the rime in the last syllable of the phrase-final word. Assuming that pitch drop, intensity drop, and lengthening are present in both L- and L-L% boundaries, we expected that the qualitatively different categories of ip (intermediate phrase) and IP (intonation phrase) would be indicated by quantitative differences in these acoustic cues.

[1] $http://www.isip.msstate.edu/projects/switchboard/$

2. ToBI Labeling of Switchboard Files

Switchboard is a corpus of spontaneous informal telephone conversations. The files in the WS97 subset are segmented by conversational turn and have word- and phone-aligned transcriptions. Two linguistics graduate students independently labeled 181 WS97 files, containing utterances from 79 different speakers and a total of 1698 words, according to the ToBI (Tones and Break Indices) system of prosodic transcription. In cases of inter-transcriber disagreement, agreement was reached through discussion between the transcribers and, in some cases, consultation with a third labeler. The ToBI labeling had been initiated as part of a prosody-based ASR (Automatic Speech Recognition) project, and the present study was largely motivated by the high level of disagreement over phrase boundary level as compared to disagreement over phrase tone and pitch accent location and type. Table 1 summarizes the distribution of the agreed-upon L- and L-L% labels according to presence and type of pitch accent in the phrase-final word[2]. Pitch accent information was taken into account because of its influence on the overall pitch, intensity, and duration of the word.

Table 1: *Distribution of L- and L-L% labels*

Boundary	Pitch Accent	N. of tokens
L-	no PA	137
	H*	101
	L*	8
	total	246
L-L%	no PA	36
	H*	76
	L*	6
	total	118

3. Methods

Based on Wightman et al.'s [11] finding that perception of prosodic phrase boundary level is highly correlated with the lengthening of the final-syllable nucleus in the pre-boundary word, we compared normalized durations for the nuclei of pre-boundary syllables for L- and L-L% tokens.

For duration normalization, we determined the mean (μ^k) and standard deviation (σ^k) of each vowel phone (x^k) as measured from the phone aligned transcriptions over the whole WS97 corpus. The phone-based normalization ($\bar{d}_i^k$) formula is given in (1):

[2] Pitch accents were labeled only for the starred tone; these data include pitch movements with specified leading or trailing tones.

$$\bar{d}_i^k = \frac{x_i^k - \mu^k}{\sigma^k} \quad (1)$$

where x_i^k is the observed duration of token x_i belonging to vowel phone class *k*.

In order to compare pitch contours for the two boundary levels, we analyzed both the size and the slope of the F0 drop. The size of the drop was calculated based on the normalized differences between the beginning and ending F0 values of the pre-boundary rimes. The rime beginning was hand-labeled based on the spectrogram and waveform, and the end was marked as the last pitch point calculated by Praat[3]. Instances of pitch failure over the sonorant portion of a rime were labeled "no pitch" and the rime was excluded from the F0 analysis. Values indicative of pitch doubling or halving were manually checked and corrected. Normalization was based on the mean pitch value over all of a speaker's turns in the conversation from which the WS97 file was extracted[4]. This normalization was intended to account for variability in pitch range across same-speaker utterances. The pitch drop normalization ($\triangle\bar{F}_{0i}$) is shown in (2):

$$\triangle\bar{F}_{0i} = \frac{(F_{t0} - F_{t1}) - \mu_i}{\sigma_i} \quad (2)$$

where F_{t0} and F_{t1} are the F0 values obtained at the rime beginning time (t0) and rime end time (t1), and μ_i and σ_i are the mean and standard deviation of F0 over all of the speaker's turns.

The F0 slope ($S_{\triangle\bar{F}_0}$) was calculated by taking the normalized F0 difference ($\triangle\bar{F}_0$) over the rime duration, which was calculated as the interval ($\triangle t$) between the beginning and ending F0 values of the pre-boundary rime, as in (3):

$$S_{\triangle\bar{F}_0} = \frac{\triangle\bar{F}_0}{\triangle t} \quad (3)$$

Measurements for both F0 difference and slope were categorized according to the presence and location of pitch accent on the pre-boundary word, as determined from the agreed-upon ToBI transcriptions. This allowed for the possibilities that F0 difference measurements might be affected by the presence of pitch accent on the rime, and that F0 slope values might be affected by the location of pitch accent on the rime versus on a preceding syllable.

Intensity drop ($\triangle\bar{I}_i$) for L- and L-L% was calculated using the normalized differences over the rime intervals, which were measured - as for F0 drop - between the rime beginning and the final F0 point, as in (4):

$$\triangle\bar{I}_i = \frac{(I_{t1} - I_{t0}) - \mu_i}{\sigma_i} \quad (4)$$

Use of the same interval for both pitch and intensity measurements is based on the premise that segmental interference with either F0 or intensity is more likely to be caused by obstruents. The portion of the rime over which pitch is calculable is comprised of sonorants, so that intensity measured over this interval should reflect macro- rather than micro prosodic effects. Normalization was based on the mean intensity for a speaker's turn in the conversation.

[3] $http://www.praat.org$

[4] Files consisting of all of one speaker's turns in a conversation are part of the MS98 subset of Switchboard. The total number of words in each MS98 file ranges from 700 to 1300.

It is possible that relevant F0 information is available beyond the last calculable pitch point; for example, the pitch-tracking algorithm may fail due to aperiodicity from creaky phonation. Creaky voice has been identified as a potential boundary cue [1, 3, 5], and aperiodicity at low F0 levels can thus be included among the acoustic cues for which we would predict quantitative differences between L- and L-L%. We have not yet developed a method for quantifying the degree of creakiness in individual tokens. However, it may be informative to compare the rate of occurrence of creaky voice for the two boundary levels; Redi and Shattuck-Hufnagel [10] found that phrase-final creakiness was significantly more frequent at IP than at ip boundaries. To investigate the relationship between creaky phonation and boundary level, the presence or absence of creak at each instance of pitch failure over the rime was manually identified using the waveform and spectrogram. The percentage of pitch failure attributable to creak was compared for the two boundary levels.

4. Results

The results for F0, intensity, and duration comparisons are shown in Figures 1-7. No significant differences were found for F0 drop (Figures 2-5) or intensity drop (Figures 6-7), but lengthening as measured by pre-boundary nucleus duration was found to differ significantly between the two boundary levels (Figure 1). Finally, pitch track failure due to creakiness was found to occur more frequently at L-L% than at L- boundaries (Table 2).

4.1. Duration

The error bars in Figure 1 show the difference in normalized nucleus duration between L- and L-L%.

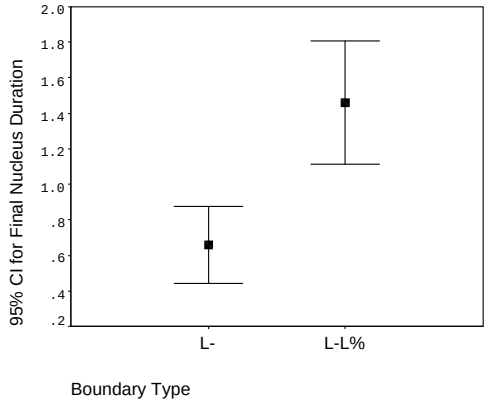

Figure 1: *95% CI for Normalized Final Nucleus Duration*

Nucleus durations for L- have a mean of 0.66 and standard deviation of 1.7, while the L-L% durations have a mean of 1.5 and standard deviation of 1.9. The positive mean values for both L- and L-L% indicate that final syllable lengthening is indeed characteristic of both boundary types, and the 95% Confidence Interval (CI) with $p < 0.001$ indicates that the two boundary types can be distinguished from one another by degree of final syllable lengthening.

4.2. Pitch

The scatter plots in Figure 2 show the magnitude of F0 drop in the non-normalized data for the two boundary levels. The distribution of the F0 values supports the ToBI guidelines statement [1] that L- and L-L% can not be reliably differentiated based on the magnitude of the F0 drop.

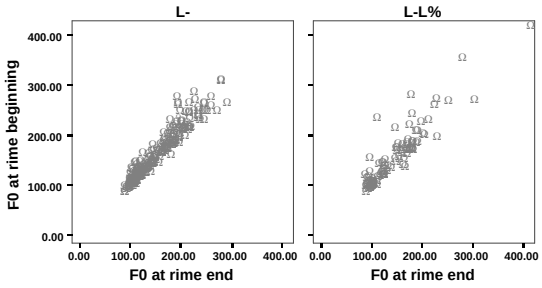

Figure 2: *F0's at the beginning and end of rime*

Comparison of the F0 drop measurements after normalization still does not indicate any significant difference between L- (mean = 0.33, SD = 0.39) and L-L% (mean = 0.36, SD = 0.60), as shown by the error bars in Figure 3 (p = 0.635 at α = 0.05):

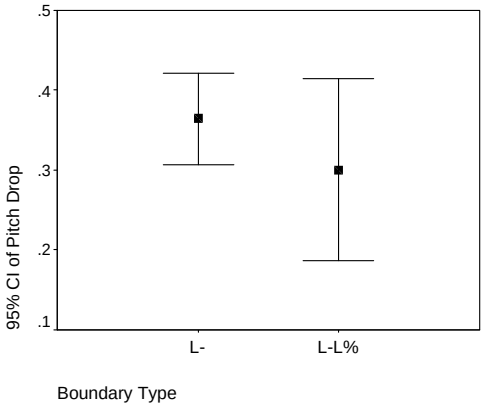

Figure 3: *95% CI for Normalized Pitch Drop*

F0 slope also failed to differ significantly between L- (mean = 1.85, SD = 2.26) and L-L% (mean = 1.84, SD = 2.26), as shown in Figure 4 (P = 0.98 at α = 0.05):

Although the presence and type of pitch accent on the pre-boundary word significantly affected F0 drop magnitude for both boundary types ($p<0.05$) as shown in Figure 5, neither the presence nor the location of pitch accent had any significant effect on the mean F0 values for either boundary type ($p > 0.900$ for each case).

4.3. Intensity

The scatter plots in Figure 6 show the magnitude of intensity drop in the raw data for the two boundary levels. The distribution of the intensity values indicates no difference between L- and L-L%.

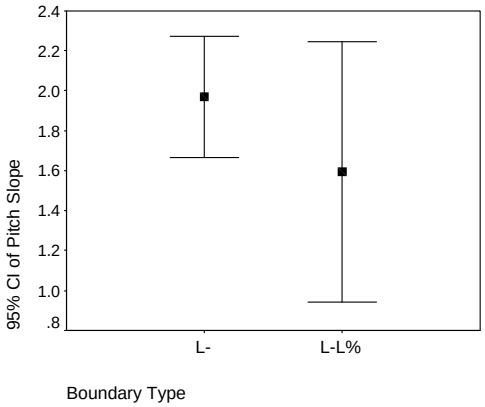

Figure 4: *95% CI for Normalized Pitch Slope*

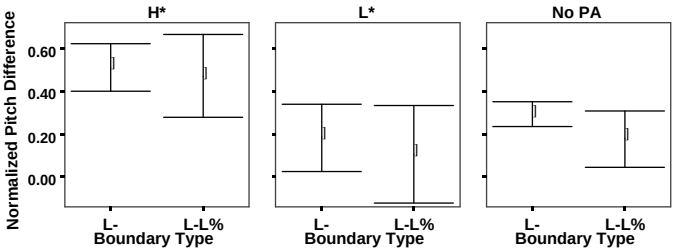

Figure 5: *95% CI for Normalized Pitch Drop Depending on Pitch Accent*

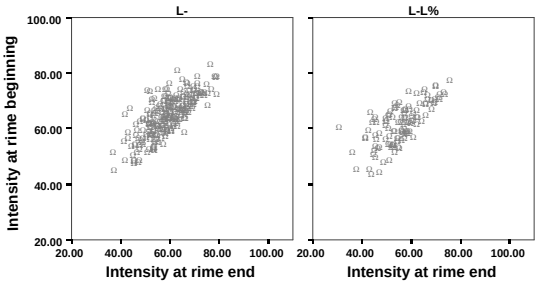

Figure 6: *Intensity at the beginning and end of rime*

The error bars in Figure 7 show the difference in intensity drop between the two boundary levels. The significant overlap of means and standard deviations strongly suggests that magnitude of intensity drop does not differentiate L- from L-L%.

4.4. Voice Quality

Table 2 shows the distribution of pitch failture due to creakiness for L- and L-L%. The percent of total boundary tokens for which pitch failure occurred, and the percent of total pitch

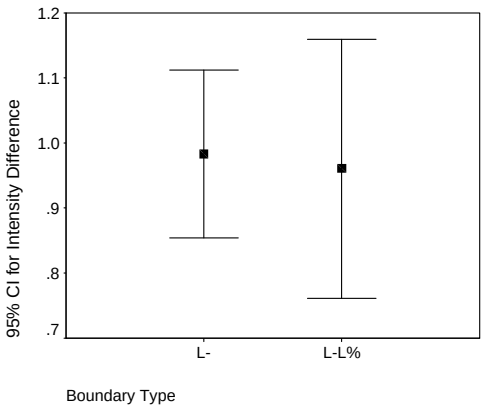

Figure 7: *95% Confidence Interval for Normalized Intensity Drop*

failures due to creakiness, are shown in parentheses.

Table 2: *Distribution of pitch failure due to creaky voice*

Boundary	N. of tokens	N. of creakiness
L-	9 (3.66 %)	6 (2.43 %)
L-L%	20 (16.95 %)	15 (12.71 %)

The percent of pitch failure occurrences due to creakiness is greater for L-L% than for L-, supporting Redi and Shattuck-Hufnagel's [10] finding that phrase-final creakiness is more likely to occur at IP than at ip boundaries.

5. Discussion

Although duration of the phrase-final syllable nucleus was the only cue for which we found a non-overlapping bimodal distribution, the possibility that F0 differs in degree between L- and L-L% can not be ruled out without further investigation. We are currently extracting measurements for comparing the minimum F0 value of pre-boundary words with the minimum F0 value of phrase medial words as an alternative basis for normalizing F0 drop. It is also necessary to investigate other prosodic cues, such as post-boundary pause, that might differ in degree or frequency of occurrence between L- and L-L% in particular, and between ip and IP in general. Further, comparisons of acoustic cues for boundary level across different phrase tones should be undertaken. Regarding creaky voice as a cue for differentiating between boundary levels, our findings were positive but limited in scope because we considered only those instances of creakiness that were correlated with pitch track failure. To further investigate creaky voice as a boundary level cue, we must identify all cases of preboundary creaky voice and compare the frequency of occurrence and possibly the degree of creakiness between ip and IP.

We find it encouraging that our results, obtained from spontaneous non-laboratory speech in the Switchboard corpus, are consistent with some of the findings for speech from more controlled or formal settings [5][10][11].

6. Acknowledgements

This work was funded through the University of Illinois Critical Research Initiative. Thanks to Chilin Shih and Ken Chen for helpful comments and advice.

7. References

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

[2] Beckman, M.E.; Hirschberg, J., 1994. *The ToBI annotation conventions*. ms, The Ohio State University and AT&T Bell Telephone Laboratories.

[3] Beckman, M.E.; Pierrehumbert, J.B., 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3, 255-309.

[4] Crystal, T.H.; House, A.S., 1990. Articulation rate and the duration of syllables and stress groups in connected speech. *JASA*, 88(1), 101-112.

[5] Epstein, M.A., 2002. *Voice Quality and Prosody in English*. Ph.D. dissertation, UCLA.

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

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

[8] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In *Intonations in Communication*, P. Cohen; J. Morgan; M. E. Pollack (eds.). Cambridge, Mass.; MIT Press, 271-311.

[9] Pitrelli, J.F.; Beckman, M.E.; Hirschberg, J., 1990. Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. *Proceedings of the International Conference on Spoken Language Processing.* Yokohama: Japan, 123-126.

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

[11] Wightman, C.W.; Shattuck-Hufnagel, S.; Ostendorf, M.; Price, P.J., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *JASA*, 91(3), 1707-1717.