Local Acoustic Cues Distinguish Two Levels of Prosodic Phrasing: Speech Corpus Evidence

Tae-Jin Yoon, Heejin Kim, and Sandra Chavarría
Department of Linguistics
University of Illinois at Urbana-Champaign

INTRODUCTION

- Beckman and Pierrehumbert (1986) propose that "the phrase-accent plus boundary tone configuration [used as a complex tonological mark of phrase juncture] in Pierrehumbert (1980) should be reanalyzed as involving correlates of two levels of phrasing." They introduce the phrase-accent (or intermediate phrase) as a level below the intonation phrase in the prosodic hierarchy.
- This proposal of an independent intermediate phrase is based on the auditorially perceived degree of disjuncture and on observed F0 contours from a small set of data, including controlled speech data produced under laboratory conditions and synthesized speech data.
- Much of the recent research on intonation in speech relies on the intonation transcription standard of the ToBI (Tones and Break Indices) system (Beckman & Ayers 1997), which adopts the Pierrehumbert-Beckman model, including the distinction between the intermediate and intonational phrase levels.

- However, in actual labeling practice with non-laboratory speech, it can be very difficult to judge the level of phrase juncture in cases where the pitch contour alone is not definitive.
- For example, a phrase ending in a low or falling pitch contour could be analyzed with a low intermediate tone (L-) or with a sequence of a low intermediate tone followed by a low boundary tone (L-L%), as noted in the ToBI labeling guidelines.

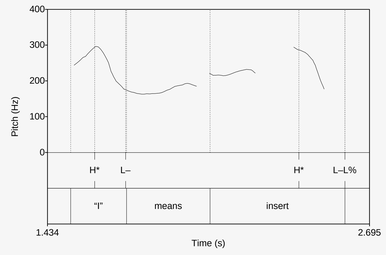

Figure 1: Pitch contour of an utterance "'I' means insert" illustrating two boundary levels – intermediate and intonational phrases (Beckman and Pierrehumbert 1986). Sound file along with ToBI transcription is taken from ToBI Guidelines (Beckman and Ayers 1997).

- In such cases the human labeler must rely on cues other than the gross pitch contour to identify the level of phrasal juncture.
- This labeling challenge can be especially acute in conversational speech style for speakers who exhibit less overall pitch variation and more frequent interruption of pitch contour due to disfluencies than is often found in laboratory speech.

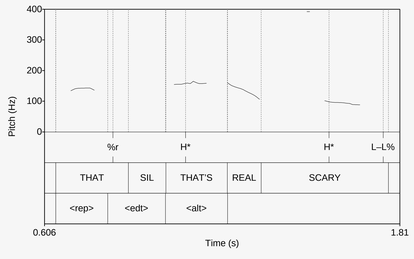

Figure 2: Pitch contour of an utterance "Tha[t] that's real scary" from a WS97 file of the Switchboard corpus illustrating prosodic disruption and less overall pitch variation.

RESEARCH QUESTIONS

Our current research addresses two questions:

(1) Do human labelers reliably distinguish two levels of phrase juncture in non-laboratory speech?
(2) If so, what are the acoustic factors that condition the perceived distinction?

CORPORA

- We analyzed intonation in two large speech corpora:

  (1) The Switchboard corpus of telephone conversation speech
  (2) The Boston University Radio News corpus of read speech

- For the Switchboard corpus, we produced our own ToBI labels of the WS97 subset (Yoon et al., to appear), and analyzed files with agreed-upon ToBI labeling (around 180 files, 80 speakers, 1700 words).
- For the Radio News corpus, we analyzed the lab news portion of two speakers (F1A and F2B), for which ToBI labeling is available (Ostendorf et al. 1995).
- The pitch accent inventory was collapsed into H* and L* for both corpora.

Table 1: Distribution of L- and L-L% tokens in a subset of 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 in 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

- 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. Creaky tokens were excluded from our analysis of pitch due to frequent pitch track failure.
- Published reliability studies, including Yoon et al. (to appear), show that human labelers do reliably distinguish the intermediate from intonational levels of phrase juncture. Agreement rates range from 80% to a little over 90 % depending on corpus and/or labeling inventories.

MEASUREMENTS

For the comparion of acoustic cues at the two preboundary levels, we applied the following normalization:

- **Duration**: We normalized vowel durations using the means and standard deviations of each phone obtained across all speakers for the Switchboard corpus and within speakers for the Radio News corpus.
- **Pitch and intensity**: For Switchboard, the domain for F0 and intensity normalization was defined over the individual utterance as delimited by the beginning and ending of the WS97 file, which contains approximately 3-4 intonational boundaries. For Radio News, pitch and intensity were not normalized, being analyzed within, rather than across, speakers.

The following acoustic measures of F0, intensity, and duration were taken from the phrase-final syllable rime for each boundary type from both corpora:

- **Beginning F0**: For preboundary syllables with an 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**: measured at the point of peak intensity in the rime
- **End F0 and end intensity**: taken at the end of the sonorant portion of the rime
- **F0 drop**: equal to end F0 minus beginning F0
- **Intensity drop**: equal to end intensity minus beginning intensity
- **F0 slope**: The F0 drop divided by the duration of the interval from beginning F0 to end F0

RESULTS

In both corpora there are significant acoustic correlates of phrase level expressed in the phrase-final syllable rime.

Duration

- Nucleus duration is longer in L-L% than in L- in both corpora (Switchboard: $F(1, 313) = 15.748$, $P < 0.001$; Radio News: F1A: $F(1, 245) = 20.069$, $P < 0.001$, F2B: $F(1, 362) = 7.967$, $p < 0.01$).

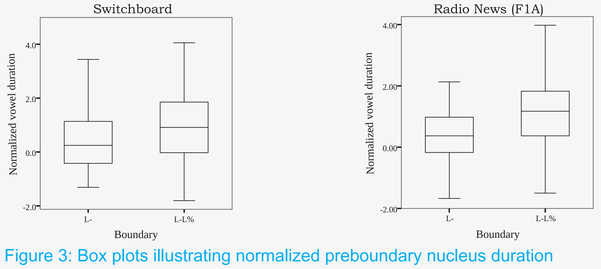

Figure 3: Box plots illustrating normalized preboundary nucleus duration

Pitch

- In both corpora, the F0 value at rime end is lower at L-L%, compared to L- (Switchboard: $F(1, 276) = 7.597$, $p < 0.01$; Radio News: F1A: $F(1, 90) = 20.371$, $p < 0.001$, F2B: $F(1, 94) = 19.316$, $p<0.001$).

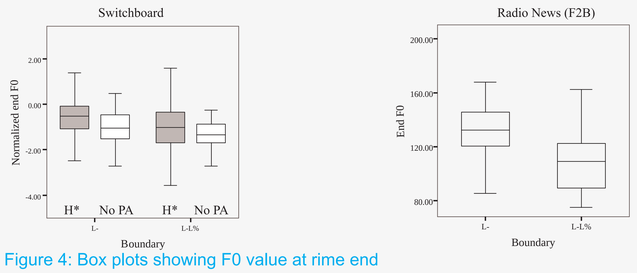

Figure 4: Box plots showing F0 value at rime end

- Further, the read speech style of the Radio News corpus manifests significant differences in F0 drop and F0 slope over the final rime (F0 drop: F1A: $F(1, 90) = 10.824$, $p < 0.05$; F2B: $F(1, 94) = 8.124$, $p < 0.01$, F0 slope: F1A: $F(1, 90) = 4.929$, $p < 0.01$; F2B: $F(1, 94) = 7.789$, $p < 0.01$). These differences were not found in the spontaneous speech of the Switchboard corpus.
- Beginning F0 is not different between two boundary levels for either corpus.

Intensity

- Peak intensity is significantly lower for L-L% than L- for Switchboard and for Radio News speaker F2B, but not F1A (Switchboard: $F(1, 276) = 12.769$, $p < 0.001$; Radio News: F2B: $F(1, 94) = 13.899$, $p < 0.001$).
- Speaker F2B shows additional significant difference in end intensity, with lower intensity value for L-L% ($F(1, 94) = 10.344$, $p < 0.01$).

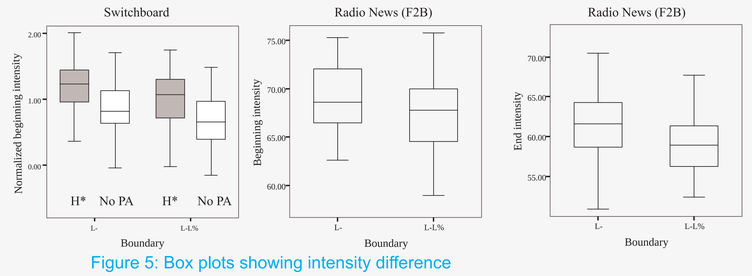

Figure 5: Box plots showing intensity difference

CONCLUSION

- Our findings provide important empirical support from non-laboratory speech for the Pierrehumbert-Beckman model in its distinction of two levels of phrase juncture.

- Our finding of acoustic correlates of phrase level in the phrase-final rime, and most often at the rime-end, offers critical support for the claim that prosodic features are locally rather than globally associated in phonological structure (Beckman and Ayers 1997).

ACKNOWLEMENT

Thanks to Jennifer Cole, Mark Hasegawa-Johnson, Chilin Shih and members of the prosody-based Automatic Speech Recognition (Prosody-ASR) Group. This work was funded through the University of Illinois Critical Research Initiative.

REFERENCES

Beckman, M. and G. Ayers. 1997. *Guidelines for ToBI labeling* (version 3.0). ms. The Ohio State University.
Beckman, M. and J. Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3, 255-309.
Chavarria, S., T. Yoon, J. Cole, and M. Hasegawa-Johnson. 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. [http://prosody.beckman.uiuc.edu]
Ostendorf, M., P.J. Price, and S. Shattuck-Hufnagel. 1995. *The Boston University Radio News Corpus*. [http://www.ldc.upenn.edu]
Patterson, D. 2000. *A linguistic approach to pitch range modeling*. PhD dissertation, University of Edinburgh.
Pierrehumbert, J. 1980. *The phonetics and phonology of English intonation*. PhD dissertation, MIT.
Wightman, C., S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. 1992. Segmental durations in the vincinity of prosodic phrase boundaries. *JASA* 91(3), 1707-1717.
Yoon, T., S. Chavarría, J. Cole, and M. Hasegawa-Johnson. (to appear). *Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI*. ICSA International Conference on Spoken Language Processing. [http://prosody.beckman.uiuc.edu]