PROSODIC BOUNDARY LEVELS CONDITIONED BY SYLLABLE-FINAL VOCALIC DURATION

Tae-Jin Yoon

McMaster University, Canada
tjyoon@mcmaster.ca

ABSTRACT

Prosodic structure encodes grouping of words into hierarchically layered prosodic constituents, including the prosodic word, intermediate phrase (ip) and intonational phrase (IP). This paper investigates the phonetic encoding of prosodic structure from a corpus of scripted broadcast news speech through analysis of the acoustic correlates of prosodic boundary at three levels of prosodic structure: Word, ip, and IP. Evidence for acoustic effects of prosodic boundary is shown in measures of vocalic duration local to the domain-final rhyme. These findings provide strong evidence for prosodic theory, showing acoustic correlates of a 3-way distinction in boundary level.

Keywords: prosodic word, intermediate phrase, intonational phrase, prosodic boundary, syllable rime duration

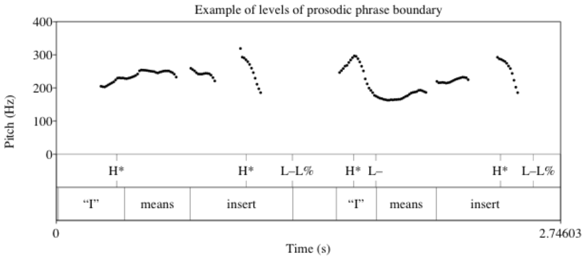
1. INTRODUCTION

This paper investigates the phonetic encoding of prosodic structure through analysis of the acoustic correlates of levels of prosodic boundary. A linear model of intonation structure, known as the autosegmental-metrical (AM) model of intonation, describes intonation in terms of a sequence of categorically distinct, non-interacting tonal events such as pitch accents and boundary tones (cf. [8]). In the linear model, the tonal events are assumed to be exclusively locally determined. Given the linear model of intonation, the grouping of words into hierarchically layered prosodic constituents is expected to be signaled by locally determined phonetic properties, such that increasingly longer duration may be observed from the prosodic word to intermediate phrase (ip), and to intonational phrase (IP) [1, 7, 8, 11].

However, little consensus has been reached regarding how the division between boundary tones in connected speech is phonetically signaled [4], and whether there exists another level of boundary that is higher than the prosodic word and lower then the boundary tone [11]. Some researchers assert that phonological criteria are sufficient enough to indicate where an intonational boundary should go in connected speech [5]. Other researchers express concern that they run into constant difficulty in identifying intonational groups in spontaneous speech [3].

Besides, an intermediate level of prosodic boundary such as Intermediate Phrase in American English and Accentual Phase in Korean and Japanese is claimed to exist above prosodic word [1, 2, 7, 11]. For example, Figure 1 illustrates an example of the two levels of prosodic boundary above prosodic word.

Figure 1: An illustration of two levels of prosodic boundary of intermediate and intonational phrase (p. 289 in [2]).



The two utterances in the figure are provided by [3] as a canonical minimal pair that necessitates a level of prosodic phrasing below intonational phrase (IP) and above a prosodic word, i.e., intermediate phrase (ip). In the figure, the F0 contours from the same strings "*I means insert*" are represented, which differ from each other regarding the prosodic realization of the subject "*I*". The utterance "*I means insert*" can be realized with one prosodic phrasing unit (L-L%), as in the first utterance on the left side, or it can be realized with two prosodic phrasing units (L- and L-L%), as in the second utterance on the right side. Whereas the subject *'I'* in the first utterance is not marked with any phrasal boundary, the subject '*I*' in the second utterance is marked with an intermediate phrase boundary (L-). It is interesting to note that even though silent pauses has been suggested to be a cue for prosodic boundary, the intermediate phrase boundary in Figure 1 appears to be signaled by other than a silent pause.

In this paper, using connected speech corpus, I investigate the phonetic encoding of prosodic structure at three levels of prosodic boundary: word, ip, and IP. Given the hierarchical organization, we expect to find non-elusive, audible acoustic correlates of prosodic boundaries at each of these levels, but especially at the phrase juncture of ip and IP, to guide the listener in chunking the speech signal. Guided by earlier evidence that boundary cues are local [11], evidence for acoustic effects of prosodic boundary is considered in measures of duration local to the domain-final rime. Given the inherent duration of vocalic segments [6], we expect that segmental effects will be observed in the vocalic duration in domain-final rime. Nevertheless, we expect that acoustic cues to prosodic boundaries will be observed in the length of segments in the preboundary syllable [11], especially through the lengthening of the preboundary rime. Furthermore, given the linear model of intonation [1, 7, 8], we expect to find greater effects on lengthening at successively higher levels of prosodic domains.

2. METHODS

2.1. Corpus

The corpus used for this work is drawn from a subset of recorded FM public radio news broadcasts spoken by five radio announcers, called the Boston University Radio Speech Corpus (BURSC)[10]. The BURSC is publicly available through the Linguistic Data Consortium (LDC). The BURSC is the richest data set that has prosodic annotation in the framework of ToBI (Tones and Break Indices)[1]. Radio speech appears to be a good style for prosody research, since the announcers strive to sound natural while reading with communicative intent. The corpus is also one of the most widely used corpora for studies of prosodic structure including computer algorithms designed to predict prosody prominence such as pitch accents and prosodic boundary such as intonational phrase boundary ([10]). Previously reported computer algorithms attempting to predict levels of prosodic boundary in this corpus is very low. It is less studied why these algorithms fail to predict levels of prosodic boundary with higher accuracy: it many be due to the lack of robust algorithm, or it may be due to the lack of phonetic evidence that signals levels of prosodic boundary in this corpus, albeit the perceptual responses to the levels of prosodic boundary by the labelers. Thus, this paper is designed to see whether there actually exist any quantifiable phonetic cues, with a focus on duration at the three levels of prosodic boundary.

The work reported in this paper is based on the labnews portion of the corpus, which consists of the recorded speech from 3 female and 2 male radio announcers. Each announcer read the same script of four news stories. Thus, each announcer read about 114 sentences whose average number of words is 16. The duration of the data subjected to analysis accounts to about 1 hour. The four news scripts were collected in studio recordings, and were later recorded in the laboratory by multiple announcers [9]. The stories represent independent data, covering different topics and a different time period.

2.2. Silence pause

Silent pause and pre-boundary lengthening are known to be acoustic correlates of prosodic boundary in English. While silent pause is neither a necessary nor sufficient boundary cue, the potential value of lengthening as a boundary cue is questionable as to whether different prosodic boundaries will be signaled by different values of lengthening.

There is a strong correlation between the presence of a pause and the perception of a prosodic boundary; however, the perception of a prosodic boundary does not depend on the occurrence of silent pause.

Table 1: Contingency table of the presence/absence of silent pause and the presence/absence of phrasal boundary (ether ip or IP) in the Boston University Radio Speech Corpus.

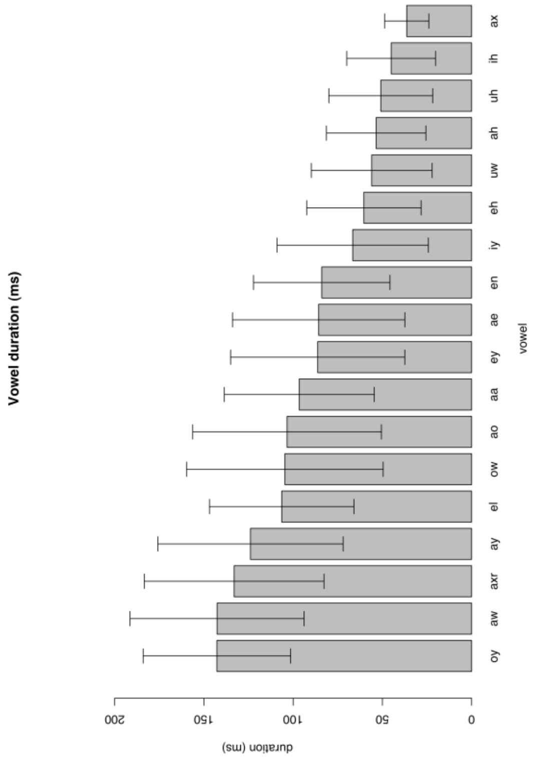
	Phrasal Boundary	No Boundary
Silence	984 (40.6%)	67 (0.8%)
No Silence	1439 (59.4%)	8056 (99.2%)

2.3. Duration measurement

Duration measures are taken for each segment following segmentation and phone labeling of the speech signal. Segmentation and labeling is automated by doing a forced alignment of the speech signal to a phone string. The phone string is taken from the dictionary encoding of each word, and forced-alignment is done using the HTK (Hidden Markov Model Toolki) [12]. The symbols in the y-axis on the right side in the figure are in

ARPABET format, in which two ASCII characters represent a phoneme (with an exception of ‘axr’). The x-axis indicates the range of raw duration measured in millisecond. The error bar at the center of each box plot indicates one standard deviation. The IPA (International Phonetic Alphabet) symbols that correspond to the ARPABET symbols in the figure are as follows: ɔɪ (oy), aʊ (aw), ɜɹ (axr), aɪ (ay), l̩ (el), o (ow), ɔ (ao), ɑ (aa), e (ey), æ (ae), n̩ (en), i (iy), ɛ (eh), u (uw), ʌ (ah), ʊ (uh), ɪ (ih), and ə (ax).

Figure 2: Mean and standard deviation of duration of each vowel in the Boston University Radio Speech Corpus, as obtained through HMM-based forced alignment.



3. RESULTS

The frequency table of vowels occurring at word-final syllable is shown in Table 2. In the table, the number of tokens of each vocalic type observed at word-final syllable is listed together with the total number of a vocalic type in the parenthesis. Note that syllabic nasals and liquids are included in the analysis.

Table 2: Frequency table of vowels and syllabic sonorants occurring at word-final syllable.

Phone	Total	Phone	Total
i	544 (1742)	o	201(377)
ɪ	606 (1671)	ʊ	27(84)
e	305 (892)	u	150(653)
ɛ	131 (536)	aʊ	64(132)
æ	197 (702)	aɪ	137(343)
ə	847 (1648)	oɪ	12(42)
ɑ	112 (353)	n̩	11(15)
ʌ	222 (756)	l̩	29(47)
ɔ	224 (543)	**Total**	**3819(10546)**

Due to the inherent duration differences among phone types, as shown in Figure 2, duration measure is normalized based on observed phone duration using the normalization method in [11], as in (1):

$$\bar{d}_i = \frac{x_i - \mu_i}{\sigma_i} \qquad (1)$$

where $x_i, \mu_i,$ and σ_i are, respectively, the observed duration, mean, and standard deviation of token *x*, belonging to vowel phone class *i*.

Normalized duration measures are taken from the nucleus segment(s) of syllables in word-final position in three prosodic contexts, as illustrated in Figure 3: (1) phrase-medial position, (2) intermediate-phrase final position, and (3) intonational-phrase final position. In the figure, measurements are taken from the syllable nucleus, as indicated by the circled x, in word-final position under the context of phrase-medial, intermediate phrase-final, and intonational phrase-final position.

Figure 3: Measurement domain for normalized duration.

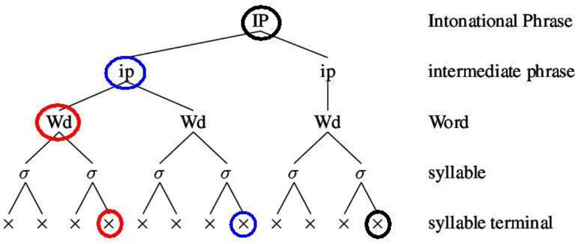
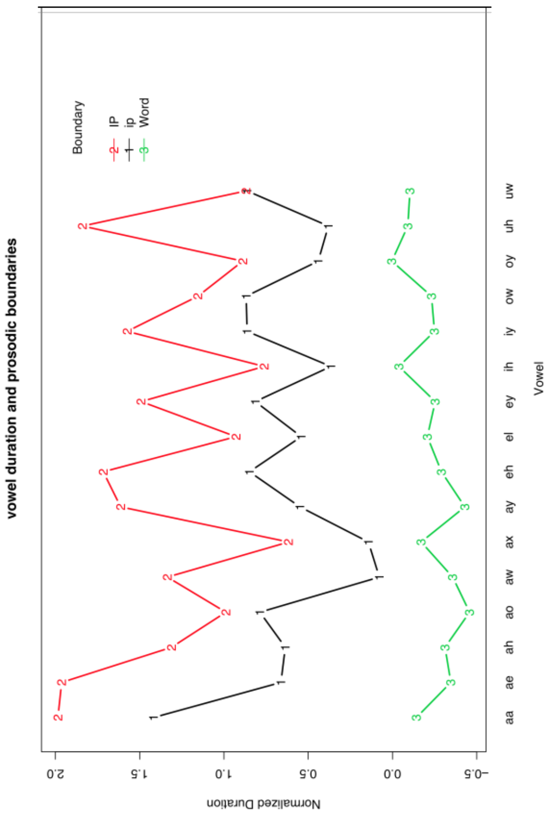


Figure 4 shows that there are consistent differences across phone types among the average normalized word-final rime durations of the three boundary levels. In this figure, we can see three levels of prosodic boundary on each vowel type, providing corroborative evidence that the boundary lengthening effects distinguish three levels of prosodic domains.

Figure 4: Average normalized rime duration of each phone type. Parameters are the level of prosodic boundary (i.e. ip, IP, and Word). As in Figure 2, the symbols in the y-axis on the right side in the figure are in ARPHABET format, and range of the normalize rime duration is shown on the x-axis.



4. CONCLUSION

The study provides evidence the levels of prosodic boundary signaled by vocalic duration. The boundary lengthening effects distinguish three levels of prosodic domains, and thus support a theory of prosodic structure that discriminates between levels of prosodic phrasing, such as ip and IP, in addition to the prosodic word. In addition, this study provides evidence for local effects of prosodic domains in the syllable at the right edge in support of the linear model of prosodic structure. Acoustic cues to prosodic boundaries will be observed in the length of segments in the preboundary syllable [2, 3, 4, 5, 11], especially in the lengthening of the preboundary rime, with greater effects on lengthening at successively higher levels of prosodic domains [5].

Further work is needed to find a way to use these durational differences in building a model of prosodic phrasing system in which intermediate phrasing is included in addition of intonational phrasing. In order to work on this, one needs to take into account the problem of data scarcity. The number of intermediate phrase is much smaller than those of words or intonational phrases. This scarcity of intermediate phrase in corpuses will pose a challenge to an attempt to build such prosodic phrasing determination models.

5. ACKNOWLEDGMENTS

This work is supported by NSERC (Natural Science and Engineering Council of Canada). Statements in this paper reflect the opinions and conclusions of the authors, and are not necessarily endorsed by the NSERC.

6. REFERENCES

[1] Beckman, M., Ayers, G.M. 1997. *Guidelines for ToBI Labeling* (ver. 3.0). The Ohio State University.

[2] Beckman, M.E., Pierrehumbert, J. 1986. Intonation structure in English and Japanese. *Phonology Yearbook 3*, 255-310.

[3] Brown, G., Currie, K.L., Kenworthy, J. 1980. *Questions of Intonation*. London: Croom Heim.

[4] Cruttenden, A. 1997. *Intonation* (2nd ed). Cambridge: Cambridge University Press.

[5] Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

[6] Klatt, D.H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of Acoustical Society of America* 59(5), 1208-1221.

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

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

[9] Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S. 1995. *The Boston University Radio News Corpus*. Technical Report ECS-95-0001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA.

[10] Ross, K.N., Ostendorf, M. 1996. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language* 10, 155-185.

[11] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P.J. 1992. Segmental duration in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91, 1707-1717.

[12] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. 2003. *The HTK Book.* Cambridge, UK: Cambridge University Engineering Department.