Linguistic Research 31(1), 1-23

Speaker variation in English prosodic boundary*

Tae-Jin Yoon
(Cheongju University)

Yoon, Tae-Jin. 2014. Speaker variation in English prosodic boundary. *Linguistic Research* 31(1), 1-23. This paper analyses the rate of inter-speaker consistency in the way multiple speakers render prosodic events when they read the same scripts. Prosodically labeled data of five speakers from the Boston Radio Speech Corpus (BURSC) are used to measure the degree of speaker variation in rendering prosodic boundaries. The results indicate that the average rate of consistency on the presence or absence of boundary tones 89.71%. For the rate of consistency for the levels of boundary tones, an average consistency of 79.25% is achieved when three levels of category (i.e., ip, IP and no boundary) are compared in pairs of speakers. The rate is lowered to 76.74% when a comparison of prosodic strength is made after both speakers in a pair agreed that there are phrasal tones on aligned words. When a pair of speakers both have a prosodic boundary on a given word, the agreement rate on the type of phrasal tones is 50.95%. The rate of speakers' consistency in the presence of boundary tones is comparable to the rate of inter-transcriber reliability. The comparable rates with regard to locating prosodic boundaries in utterances by speakers and transcribers may be interpreted that the production and perception of prosodic phrasing are closely related to each other. The high rate of speakers' consistency is interpreted to be affected by syntactic structures, in spite of the lack of isomorphic relations between prosodic phrasing and syntactic phrasing. **(Cheongju University)**

Keywords prosodic boundary, ToBI, consistency, reliability, speaker variation

1. Introduction

Speech conveys information not only through segments such as vowels and consonants, but also through prosody. Two aspects of prosody are identified to be essential components for conveying information above segmental levels: prosodic phrasing and prosodic prominence. Prosodic phrasing is concerned with chunking words to perceptually coherent intonational contours. Prosodic prominence refers to perceptual salience of a word or syllable relative to other words or syllables in the

* I am grateful to two anonymous reviewers for their invaluable comments on the earlier version of the paper. The paper has significantly improved thanks to their constructive comments.

same prosodic phrase. Unlike the case of segments, one of the great difficulties about prosody research is what Xu (2011) calls the lack of reference problem. Reference is referred to by Xu (2011) as “a pivot that serves as both a starting point of inquest and a point that one can comfortably fall back on.” Aspects of prosody are phonetically expressed through F0, intensity, duration, voice quality, and the occurrence of silent pauses. Research to date has informed that the structure of prosody is based on complex interactions within and between different levels of linguistic and paralinguistic organization. Due to complex interaction of various linguistic and paralinguistic factors, even the same utterance is not produced identically by different speakers. It is also possible that the same utterance with the same prosodic structure may be perceived differently by different listeners.

Given the variation in the production and perception of prosody, we can use prosodic annotation system to ask questions concerning the consistency of prosodic structure of the same utterances perceived by different listeners and the consistency of prosodic structure of the same utterances rendered by different speakers. That is, we can address questions like the following: when different listeners listen to an utterance proposed by the same speaker, how consistent are they in their perception of prosody? When different speakers tell the same stories, how similar are they in their prosody realization?

The question of listener consistency has been studied and reported under the realm of inter-transcriber reliability study (Pitrelli, Beckman, and Hirschberg, 1994; Grice, et al., 1996; Syrdal and McGory, 2000; Jun et al., 2000; Gut and Bayerl, 2004; Yoon, Chavaria, Cole and Hasegawa-Johnson, 2004; Dilley, Breen, Bolivar, Kraemer, and Gibson, 2006; Breen, Dilley, Kraemer, & Gibson, 2012). For example, Pitrelli et al. (1994) studies inter-transcriber reliability for 26 labelers on 489 words taken from both read and spontaneous speech corpora. GA study by Syrdal and McGory (2000) employed six annotators who labeled 645 words. Yoon et al. (2004) investigated two labelers’ inter-transcriber reliability in ToBI for a larger corpus of spontaneous speech including 79 speakers and 1600 words. Dilley et al. (2006) presented inter-transcriber reliability studies by naive labelers and Breen et al. (2012) extended the reliability study by including expert labelers. The studies on inter-transcriber reliability have been done in other languages including German (Grice et al., 1996; Gut & Bayerl, 2004) and Korean (Jun et al., 2000). These previous studies of ToBI agreement have revealed some consistent findings. All of these prior studies of the agreement have demonstrated high agreement on the presence of a pitch accent and moderate agreement on pitch accent type. With regard to the presence of a pitch accent, more than 80% is reported to be the agreement rate by labelers. As for pitch accent type, labelers agreed more than 60%. These previous studies have demonstrated high agrement with regard to the presence or absence of intonational boundaries, with more than 89%.

On the other hand, less is known about the degree of consistency in the realization of prosodic structure when different speakers are telling the same stories in a natural setting. This paper seeks to answer the question of speaker consistency in the realization of prosody, especially prosodic phrasing. For this purpose, analyses are made on the prosodic labels annotated using the ToBI (Tones and Break Indices) prosody annotation system (Silverman et al., 1992; Beckman and Ayers, 1997) on a spoken corpus produced by professional radio announcers. Possible explanation on the rate of speaker consistency is discussed by reviewing previous works on computer-based prosodic phrasing prediction experiments.

The paper is organized as follows: Section 2 introduces the corpus that has prosody labels and that are used for the current paper. Reviews of inter-transcriber reliability studies on prosodic phrasing are presented in Section 3. Section 4 reviews the machine learning based studies of prosodic phrasing prediction experiments. Section 5 presents the study results of speaker consistency in rendering prosodic phrasing. Section 6 discusses the results presented in section 5 and concludes the paper.

2. The Boston university radio speech corpus (BURSC)

The corpus used for this work is drawn from a subset of recorded FM public radio news broadcasts spoken by five radio announcers. The corpus is called the Boston University Radio Speech Corpus (BURSC) (Ostendorf, Price, and Shattuck-Hufnagel, 1995). The BURSC is publicly available through the Linguistic Data Consortium (LCD). Radio speech appears to be a good style for prosody synthesis research, since the announcers strive to sound natural while reading with communicative intent. 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[1]. Announcers read the same script of four news stories. The four news scripts were collected in studio recordings, and were later recorded in the laboratory by multiple announcers. Examples of the script is in (4).

(4) Wanted: Chief Justice of the Massachusetts Supreme Court. In April, the S.J.C.'s current leader Edward Hennesy reaches a mandatory retirement age of seventy, and a successor is expected to be named in March. It may be the most important appointment Governor Michael Dukkakis makes during the remainder of his administration and one of the toughest. As WBUR's Margo Melnicov reports, Hennessy will be a hard act to follow. (taken from the file f1ajrl1.txt)

The stories represent independent data, covering different topics and a different time period. With these background on the corpus and the ToBI annotation system, an illustration of the ToBI annotated corpus is given in Figure 1.

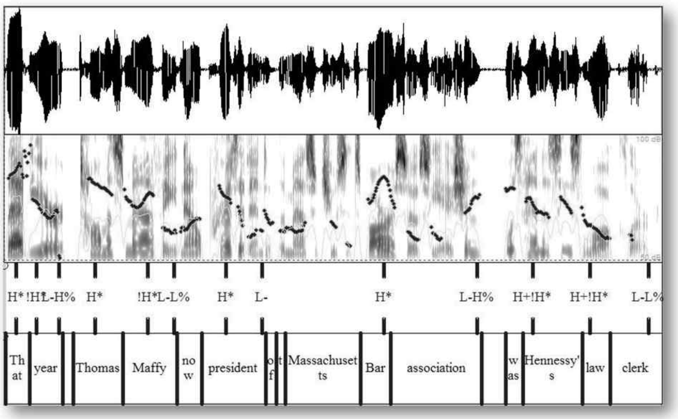

Figure 1. An illustration of parts of the corpus with ToBI annotation

In the figure, the two panels are waveform and spectrogram. The bottom two panels contain symbolic information. The first tier contained tonal information and the second tier lists time-delimited word information. The corpus contained data in

1 Five speakers are referred to as F1, F2, F3, M2, and M3, based on the convention in the released corpus. The speaker M1 is excluded from the analysis, because it does not have enough annotated prosodic labels.

which the ToBI prosodic marking system has been made by expert labelers for sentences that 4 announcers had read in a laboratory. Each announcer read about 114 sentences, each of which had 16 words in average. Raw statistics that are observed in the radio speech corpus for pitch accents are presented in Table 1, and boundary tones in Table 2.

Table 1. Distribution of pitch accents in the radio speech corpus[2]

Accents	Number of tokens	Percentage
H*	2589	46.89%
L+H*	1128	20.43%
!H*	712	12.89%
*?	291	5.27%
H+!H*	266	4.81%
L+!H*	245	4.43%
L*	228	4.12%
X*?	31	0.56%
L*+H	30	0.54%

Table 2. Distribution of phrasal tones (i.e., intermediate and intonational phrase) in the radio speech corpus[3]

Phrasal tones	Number of tokens	Percentage
L-L%	1026	35.60%
L-H%	709	24.60%
!H-	368	12.76%
L-	344	11.93%
H-	313	10.86%
H-L%	82	2.84%
!H-L%	19	0.65%
H-H%	12	0.41%
-?, %?, -X?	9	0.31%

The BURSC is the richest data set that has prosody annotations, and is one of

[2] It is sometimes difficult to decide whether categorical tones are present or not, and if so, what type of tones is present in the speech signal. Therefore, a few diacritic symbols are reserved for unspecified or uncertain tonal events, including symbols such as ‘?’, and ‘X.’ For example, *? means that it is not certain whether a syllable is accented or not. X*? means that a syllable is accented but it is not clear what type of accent must be assigned to the syllable.

[3] The three labels, -?, %?, and -X?, indicate an under-specification on the presence of phrasal accent (-?) or the boundary tone (%?), and an uncertainty on the type of phrasal accent (-X?).

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 (Sun, 2002; Brenier, Cer, and Jurfasky, 2005). The computer-based prediction experiments using speech corpora are characterized by their use of stochastic approaches rather than deterministic approaches.

There have been arguments for and against the use of frequency or probability in describing and explaining linguistic systems (Manning and Schütze, 1999). Some linguists hold the position that “[O]ne's ability to produce and recognize grammatical utterances is not based on notions of statistical approximations and the like” (Chomsky 1957:16), whereas others maintain that “[S]tatistical considerations are essential to an understanding of the operation and development of language” (Lyons 1968: 98). The recent advancement of methodologies for studying the role of frequency and probability in determining language patterns have fueled discussion on the nature of linguistic rules or constraints.[4] The analysis of frequency proves to be useful in evaluating the proposed theory of intonation, but more importantly, it can be employed in stochastic modeling of prosodic structure. In a comprehensive review of previous work on prosody, Cutler, Dahan, & van Doneselaar (1997) states that stochastic approaches are better suited to modeling of prosodic structure than algorithmic and deterministic approaches. Cooper and Paccia-Cooper (1980) presents an algorithmic approach in placing boundary tones and in determining strength of prosodic phrasing in a sentence. Even in his algorithmic approach, probabilities plays a role in determining phrasal tones and their strength. Jackendoff (2002) makes this point clear by stating that “the right approach to these correspondence [between phonology and syntax] sees Intonational Phrases as phonological units that on one hand constrain the domains of syllabification, stress, and intonation, and that one the other bear a loose relation to syntax (p. 119)” and then he postulates the following formulation rules in (5) for rules of intonational phrasing (where IntP stands for intonational phrase).

(5) a. An utterance consists of a series of one or more concatenated IntP's forming a flat structure. Each IntP is a sequence of Words.

---

4 See, for example, Bod, Hay & Jannedy (2003) for the role of probability in a range of subfield of linguisticss including phonology, morphology, syntax, and semantics.

b. Preferably, the IntPs are of equal length.
c. Preferably, the longest IntP is at the end.
d. (Possibly, some strong preferences on maximum duration of IntPs, e.g., try not to go over three seconds.) (Jackendoff 2002: 119)

If we agree with Jackendoff (2002) in using terms such as 'preferably' and 'possibly' in describing the mapping between prosodic structure and other grammatical structures, then we are led to the conclusion that prosody is better formalized through probabilistic or stochastic approaches than through deterministic or algorithmic approaches. Stochastic approaches are data-driven or, in other words, corpus-based. A corpus-based approach can be successfully implemented when there exists a rich set of labeled corpora.

The BURSC is a good resource of research into stochastic approaches to prosodic structure. Because multiple speakers produce the same scripts, it is possible to measure how similarly a number of different speakers render prosodic structure when reading the scripts. In fact, speaker consistency concerning the prosodic prominence is reported in Yoon (2013). Using the same dataset, pair-wise comparisons of inter-speaker consistency are calculated on prosodic prominence. The average rate of consistency on the presence of pitch accents is 79.81%. The finding indicated that constraints as well as degree of variation exist in rendering prosodic prominence on sentences. Prosodic phrasing is the other important aspect of prosodic structure. But it is quite rare to find research on the speakers' consistency in the rendition of prosodic phrasing. The paper aims at addressing speaker consistency with regard to prosodic phrasing.

3. Review of inter-transcriber reliability studies

As mentioned, the BURSC consists of spoken speech data recorded from five speakers (3 female and 2 male), each reading the same scripts that comprise more than 110 different sentences. Probably, there would not be a single instance in which two speakers realize exactly the same prosodic structure phonetically. Phonetic properties such as F0 contours are not in a perfect mapping relationship with a perceptual prosodic event. But the phonetic realization of the intended prosodic structure is not random either. Tables 3 and 4 illustrate the realization of prosodic events of the same chunks by multiple speakers. Focusing on prosodic phrasing, we can tell that there is a complete agreement in the location and the types of prosodic phrasing in Table 3. In Table 4 an agreement in the location is also observed but variability is manifest in the types of prosodic phrasing.

Table 3. ToBI labeling of the phrase "… *of the Massachusetts Bar Association* …" In the left most column in the table, F and M stands for the gender of the speaker; F for female and M for male

	of	the	Massachusetts	Bar	Association
F1				H*	L-H%
F2				H*	L* L-H%
M2			L+H*	H*	L-H%
M3			L+!H*	L+H*	L-H%

Table 4. ToBI labeling of the phrase "*Massachusetts may now* …."

	Massachusetts	may	now
F1	H* !H* L-	L+H*	
F2	H* !H* L-L%		L*+H
F3	H* L+!H* !H-		H*
M2	H* !H* L-		H*
M3	H* !H* !H-		H*

Despite higher rate of consistency in transcribed prosodic labels, there are discrepancies between tune and prosodic transcription. For example, similar F0 shapes can lead to different transcriptions and different shapes may lead to the same transcription. At least one source of mismatch can be identified between an F0 contour and the corresponding labels of tonal event: inconsistency in prosodic labeling. And the inconsistency in prosodic labeling can be examined by investigating transcriber reliability. Two previous studies exist that report the inter-transcriber reliability in the ToBI analysis on a small set of the BURSC. Thus I will summarize these two previous studies on the inter-transcriber reliability below.

The ToBI annotation system is, in essence, a perceptual labeling system. A trained transcriber decides prosodic labels perceptually and manually with the aids of audio-visual display of speech sounds. A number of concerns about the quality of labeling have been expressed for perceptual/manual labeling in general, and for ToBI labeling in particular. To assess the quality of the manual transcription of speech data, various methods have been proposed and used, including pair-wise comparisons between transcribers, and Cohen's or Fleiss' kappa coefficients (cf. Pitrelli, Beckman, and Hirschberg, 1994; Syrdal & McGory, 2000; Yoon, Chavaria, Cole & Hasegawa-Johnson, 2004; Cole, Mo, and Baek, 2008, among others). Two reliability studies have been conducted specifically for the BURSC; One by Ostendorf, Price, and Shattuck-Hufnagel (1995) and the other by Dilley, Breen, Gibson, Bolivar, & Kraemer (2006). The difference between the two studies is the degree of expertise of the labelers. Ostendorf et al. (1996) had expert transcribers label spoken utterances, whereas Dilly et al. (2006) reported inter-labeler agreement using labels produced by naive undergraduate students who have a brief period of training.

Ostendorf et al. (1995) report that the transcriber agreement on the BURSC is relatively high. Labeler consistency is conducted on a set of stories containing 1002 words. Two labelers marked 207 words with an intonational boundary. The boundary tone agrement rate was reported to be 93% for the words. When labelers agree that a phrasal tone was present, they agreed on phrasal tone 91%. These results are higher than that reported by Pitrelli et al., (1994), and this is in part due to the fact that the radio study has more clearly marked prosodic structure, and in part the due to the fact that the labelers are experienced labelers.

Dilley et al. (2006) also report on reliability conducted on a subset of the BURSC, which amounts to 20 minutes, or 5939 syllables. In Dilley et al. (2006), the transcribers were five naïve undergraduate students who have no previous prosodic annotation experience or phonetic training. The naïve transcribers are trained for ToBI labeling and then annotate about 20 minutes of read speech. The naïve transcribers spent two weeks in being trained in the ToBI labeling system, and then subsequent four weeks in labeling the speech data. As for the presence of a phrasal boundary, an agreement rate of 88% is achieved. An agreement rate for types of phrasal boundary is 76%. These two studies indicate that the consistency of listeners' perception of prosodic phrasing is quite high, despite different levels of expertise in prosodic labeling.

4. Earlier studies on stochastic prosodic phrasing prediction

One way of modeling in modeling prosodic phrasing using stochastic approaches is designing computational algorithm to attempt to develop a classifier of predicting prosodic phrasing. We may expect that because the reliability of inter-transcriber agreement is high, there will be cues from the texts or from the speech signals that can be used by stochastic machine-learning approaches that rely on frequent use of such cues or features. The simplest experimental design is classification of a prosodic boundary vs. a non-prosodic boundary, making only binary decisions at the juncture between words. This binary classification of prosodic phrasing is the most common approach found in the literature (e.g., Wang and Hirschberg, 1992; Black and Taylor, 1997). More sophisticated experiments can also be designed to distinguish multiple levels of prosodic boundary, for example, no prosodic boundary, and two levels of intermediate and intonational phrasal boundaries (e.g., Bachenko and Fitzpatrick 1990; Black and Taylor 1997; Ingulfsen 2004; Ross and Ostendorf 1996). Because the BURSC is the richest set of prosodically annotated database, this corpus has been used to develop machine-learning classifiers of prosodic phrasing. In this section, two such previous works, Cohen (2004) and Ingulfsen (2004), will be reviewed.

Cohen (2004) compares the performance of various machine learning algorithms as a tool to examine the effect of syntactic structure on prosodic phrasing. He uses the BURSC and utilize a full syntactic parser developed by Eugene Charniak (Charniak 1999) in extracting syntactic information including the part of speech of a word and the accumulated number of brackets at the end of the word as an indicator of the complexity or nesting of syntactic constituents. Table 5 is the result reported in Cohen (2004) for the task of predicting the presence or absence of an intonational phrase boundary on the basis of five machine learning algorithms including C4.5, SLIPPER, QUEST, Artificial Neural Network (ANN), and Naive Bayes classifier. In Table 5, Cohen (2004) reports the results in terms of training and testing errors, which I have converted into accuracy and put in parentheses.

Table 5. Results of Cohen (2004) on prosodic boundary prediction

Learner	Training Errors (Accuracy)	Testing Errors (Accuracy)
C4.5	7.6% (92.4%)	11.2% (88.8%)
SLIPPER	9.8% (90.2%)	10.2% (89.8%)
QUEST	9.7% (90.3%)	11.1% (88.9%)
ANN	10.1% (89.3%)	10.8% (89.2%)
Naive Bayes	11.3% (88.7%)	11.1% (88.9%)

The experimental results in Cohen (2004) indicate two points. First, syntactic constituency is an important feature for prosodic phrasing. Second, given fairly consistent performance across learning algorithms, it seems that linguistic information may be more important than the choice of a machine learning algorithm in improving the performance of the prosodic phrasing prediction.

Two levels of prosodic phrasing (i.e., intermediate phrase and intonational phrase) are usually assumed in the description of prosodic structure (Ladd, 1966). However, it turns out that correct classification of the two levels of prosodic phrasing is quite difficult to make. Ingulfsen (2004) conducted a series of experiments predicting levels of prosodic phrasing on the BURSC in addition to the above-mentioned prediction of binary prosodic boundary location. He reports that with the best setting obtained through the full-syntactic parsing (Charniak, 1999), the best performance achieves precision rate of 74.9% and recall rate of 77.9% in identifying break index 4 (or intonational phrase). As for the correct identification of break index 3 (or intermediate phrase), only recall rate of 0.56% and precision rate of 42.9% are achieved.

5. Pair-wise comparison of speaker consistency

As mentioned, studies exist on the inter-transcriber reliability and machine-learning based prosodic phrasing prediction that utilize the BURSC. But less is known about the degree of consistency in the realization of prosodic structure when different speakers are telling the same stories in a natural setting. This paper seeks to answer the question of speaker consistency in the realization of prosody, especially prosodic phrasing. The release of the BURSC is based on the agreed-upon labels by multiple labelers. Five speakers participated in the development of the corpus read the same scripts as naturally as possible, which resulted in variability among speakers in their rendition of prosodic structure. Due to the nature of the corpus design, we can ask how consistent the five speakers are in their rendition of prosodic structure. It is noted that the method of measuring the consistency rate of multiple speakers are the same as that of analysing the inter-transcriber reliability test

In order to measure the consistency, prosodic events were aligned for a pair of speakers along each word in an utterance using orthographic words as indices, as shown in Table 6.

Table 6. An example of alignment between prosody and word (one female and one male speakers)

Word	Female 1	Male 1
That	H*	H*
year	!H* L-H%	!H-
Thomas	H*	H*
Maffy,	H*+!H* L-L%	L+!H* L-L%
now		H*
president	H* L-	H* !H-
of		
the		
Massachusetts		L+H*
Bar	H*	H*
Association	L-H%	L-H%
was		
Hennessy's	H*	H+!H*
law	H*	H*
clerk.	L-L%	L-L%
...		

With the prepared dataframe as in Table 6, pair-wise comparisons of inter-speaker consistency were calculated in a couple of ways. One way of calculating the consistency was as follows: when the presence of boundary tone was of concern, types of boundary tones were collapsed and the rate of consistency was calculated. If levels of prosodic phrasing were concerned, prosodic boundaries were broadly classified into intermediate and intonational boundaries together with no boundaries.

With the post-processed dataframe with either two boundary types (i.e., presence of prosodic boundary and absence of prosodic boundary), or three types (i.e., intermediate prosodic boundary, intonational prosodic boundary, and absence of prosodic boundary), the number of prosodic events which the two speakers share in common was counted, and then divided by the total number of words, respectively. The total number of words in the present study was counted to be 1129. This way of calculating pair-wise consistency rate is reported in Syrdal & McGory (2000). It is also the way of calculating the accuracy of machine learning experiments of prosodic phrasing prediction in Ingulfsen (2004) and Cohen (2004).

The other way of calculating the consistency rate was finding out the words on which both speakers in each pair realize prosodic events, and then calculating either the strength or the type of phrasal tones. In this way of method, only words that were marked with any of phrasal accents by both speakers were used for calculation. When both speakers marked words as having no boundary tones, these words were excluded from the calculation. This is an approach adopted in Pitrelli et al. (1994). This alternative approach tends to lower the consistency rate compared to the first approach, because more words in the corpus tends not to bear prosodic events, and hence are excluded from the calculation.

In this paper, consistency rate calculated by both methods are reported. A pair-wise comparison of inter-speaker consistency regarding the presence or absence of boundary tones is reported first[5]. Here is an illustration of how the consistency rate was computed based on Table 6. In Table 6, the number of words is 15, and the two speakers have prosodic boundaries of any type in the same aligned words. Thus, the consistency rate computed happens to be 100%. As a matter of fact, mismatch can occur in which one speaker puts a boundary, whereas the other does not, making the consistency rate lower.

In Table 7, the rates of consistency for the presence or absence of boundary tones for all pairs of speakers are reported. The presence or absence of boundary tones was calculated if two speakers have any type of boundary tone (i.e., any of L-, H-, !H-, L-L%, L-H%, H-L%, !H-L%, and H-H% or even null boundary tone) on the aligned words. In the first two columns, F and M stands for the gender of the

---

5 Pair-wise consistency rate regarding pitch accents are presented in Yoon (2013). On average, the rate of consistency on the presence or absence of pitch accent is 79.81%, and an average consistency of 72.17% is achieved for the rate of consistency on the types of pitch accent.

speaker (F for female and M for male), and the number next to the F or M indicates speaker index.

Table 7. The rates of consistency for the presence or absence of boundary tone for pairs of speakers

Speaker A	Speaker B	Ratio	Consistency
F1	F2	1054/1129	93.35%
F1	F3	1016/1129	89.99%
F1	M2	995/1129	88.13%
F1	M3	1016/1129	89.99%
F2	F3	1035/1129	91.67%
F2	M2	1002/1129	88.75%
F2	M3	1023/1129	90.61%
F2	M2	984/1129	87.15%
F3	M3	1020/1129	90.34%
M2	M3	984/1129	87.15%
Average rate			89.71%

The result indicates that the consistency rate for the presence or absence of prosodic boundary tones ranges from 87% to 93%, with the overall rate of 89.71%. The seemingly high rate on the presence or absence of prosodic boundary tones may be in part due to the higher rate of absence of boundary tones on the aligned words than that of presence of boundary tones. This hypothesis is confirmed in Table 8. In Table 8, the number of words which bears no prosodic boundaries (indicated by '# of no bnd' in the second column) is presented along with the percentage of those words. Out of the 1129 aligned words, 772.2 words (or 68.4%) on average did not bear any prosodic boundary tones. When we factored out cases where both speakers did not have any type of prosodic boundary on the aligned words and considered only those cases in which at least one speaker had boundary tones, then the rate of consistency became lowered from 89.71% to 67.68%. This result implies that one can predict the presence of boundary with a success rate of at least 67.68% by using majority rule in a prosodic phrasing prediction test.

Table 8. The number and percentage of words on which both speakers in each pair do not have any prosodic boundaries

Speaker A	Speaker B	# of no bnd	Percentage
F1	F2	783	69.35%
F1	F3	769	68.11%
F1	M2	751	66.51%
F1	M3	794	70.32%
F2	F3	773	68.46%
F2	M2	749	66.34%
F2	M3	792	70.15%
F2	M2	745	65.98%
F3	M3	796	70.50%
M2	M3	770	68.20%
Average rate			68.40%

Table 9 presents the consistency rate for each pair of speakers, along with the number of boundary tones given any one of the speakers has a boundary tone in a given word.

Table 9. The pair-wise consistency rate on words with no inclusion of the absence of boundary tones on the aligned words

Speaker A	Speaker B	Ratio	Percentage
F1	F2	271/346	78.30%
F1	F3	247/360	68.59%
F1	M2	244/378	64.53%
F1	M3	223/335	66.54%
F2	F3	262/356	73.57%
F2	M2	253/380	66.56%
F2	M3	232/337	68.82%
F2	M2	239/384	62.22%
F3	M3	226/333	67.84%
M2	M3	215/359	59.87%
Average rate			67.68%

Now let's consider the consistency rate on the strength of phrasal tones. Again, there are at least two ways of computing the consistency rate for the boundary strength. In one way, the agreement rate was calculated without considering the presence or absence of prosodic boundaries. In the other way, the consistency rate was computed with a subset of data in which the aligned words had prosodic boundary tones marked by both speakers[6]. With Table 6, let me illustrate the calculation of the consistency rate of boundary strength in the two ways. As for the first way of computing, there are 15 aligned words, and one word that has different levels of prosodic boundary (i.e., *year*). Thus the rate of consistency for levels of prosodic phrasing is computed to be 93.3% (14/15*100=93.3). As for the second way, there are 5 words that are marked by prosodic boundary, and one word differs in terms of levels of prosodic boundary. Thus, the rate of consistency for boundary strength is 80.0% (4/5*100=80).

Results of speaker consistency for the two ways are presented below in Table 10 and Table 11. Table 10 is the result obtained by the first way. That is, speaker consistency with regard to prosodic strength is calculated by including cases of no boundary tones.

Table 10. Rate of consistency on the strength of prosodic phrasing. The categories are ip, IP and no boundary tone

Speaker A	Speaker B	Ratio	Consistency
F1	F2	1004/1129	88.92%
F1	F3	962/1129	85.20%
F1	M2	942/1129	83.43%
F1	M3	961/1129	85.11%
F2	F3	972/1129	86.09%
F2	M2	924/1129	81.83%
F2	M3	963/1129	85.29%
F2	M2	924/1129	81.83%
F3	M3	982/1129	86.97%
M2	M3	940/1129	83.25%
Average rate			84.79%

In Table 10, if two speakers produced the same level of prosodic phrasal boundary (i.e., ip boundary, IP boundary, or no phrasal boundary), then it was decided that they are consistent in rendering the same level of prosodic boundary. The average rate of 84.79% was calculated for the consistency rate on the strength

[6] A third way of calculating the consistency is excluding words when either of the paired speakers does not have any prosodic events. The rate of the consistency tends to be lower than the two methods described in the body of the paper.

of the prosodic phrasing.

Table 11 presents the rate of consistency on the strength of prosodic boundary after excluding words on which both speakers do not bear prosodic boundaries. In this case, the average rate of consistency lowers from 84.79% to 76.74%.

Table 11. Consistency rate of prosodic strength with no inclusion of the absence of boundary tones on the aligned words

Speaker A	Speaker B	Ratio	Consistency
F1	F2	221/271	81.51%
F1	F3	193/247	78.10%
F1	M2	191/244	78.24%
F1	M3	167/223	74.85%
F2	F3	199/262	75.92%
F2	M2	175/253	69.14%
F2	M3	171/232	73.67%
F2	M2	179/239	74.86%
F3	M3	186/226	82.26%
M2	M3	170/215	79.03%
Average rate			76.74%

With regard to the rate of consistency regarding the type of phrasal tones, there are also at least two ways of calculating the consistency rate. One way is including the no phrasal boundary tone as well as other boundary tonal types such as L-, H-, !H-, L-L%, L-H%, H-L%, !H-L%, and H-H%. The other way is by checking the type of phrasal boundary tones only when both speakers in a pair-wise comparison have any type of phrasal tones. In Table 6, the rate of consistency on the tonal type of phrasal boundary was computed to be 86.67% (13/15=86.67) when null phrasal tones were included in the calculation, and 60% (3/5=60) when only words with any tonal type of phrasal boundary were taken into consideration. In Table 12, the consistency rate regarding the type of phrasal tones is reported. All the 8 phrasal tones in Table 2 except for -?, %?, and −X? (i.e., L-, H-, !H-, L-L%, L-H%, H-L%, !H-L%, H-H%) were compared. The average rate of consistency is 79.25%.

Table 12. Rate of consistency on the type of phrasal tones for a pair of speakers

Speaker A	Speaker B	Ratio	Consistency
F1	F2	923/1129	81.75%
F1	F3	892/1129	79.00%
F1	M2	889/1129	78.74%
F1	M3	905/1129	80.15%
F2	F3	891/1129	78.91%
F2	M2	860/1129	76.17%
F2	M3	900/1129	79.71%
F2	M2	872/1129	77.23%
F3	M3	922/1129	81.66%
M2	M3	894/1129	79.18%
Average rate			79.25%

Despite the number of phrasal tones, the overall consistency rate of 79.25% seems to be higher than one might expect. One of the reasons for the higher consistency rate is due to the prevalence of words that do not have any prosodic events, as we have seen in Table 6.

The other consistency rate was calculated after excluding the words on which both speakers in each pair do not bear any prosodic boundary events. Table 13 is the result of the rate of consistency on the types of phrasal tones when both speakers bear phrasal tones on the aligned words. Because words without any phrasal tones were excluded, average consistency rate is dropped from 79.25% to 50.95%. The drop in the average consistency rate seems to be rather dramatic. Considering 8 possible phrasal tones, the consistency rate of 50.95% is still quite high. One factor that may contribute to the higher consistency rate is the uneven distribution of L-L% and L-H%. These two boundary types comprise about 60% of the phrasal tones in the corpus (see Table 2). To see how much these two boundary tone types could account for the average rate of consistency, a calculation was done only using these two boundary tones. The consistency rate on the agreement of only these two boundary tones could account for about 43.8%.

Table 13. Rate of consistency on the type of phrasal tone with no inclusion of the absence of boundary tones on the aligned words

Speaker A	Speaker B	Ratio	Consistency
F1	F2	140/271	51.64%
F1	F3	123/247	49.77%
F1	M2	138/244	56.53%
F1	M3	111/223	49.75%
F2	F3	118/262	45.02%
F2	M2	111/253	43.85%
F2	M3	108/232	46.53%
F3	M2	127/239	53.11%
F3	M3	126/226	55.72%
M2	M3	124/215	57.64%
Average rate			50.95%

6. Discussions and conclusion

It is acknowledged that the method of measuring the rate of speaker consistency for prosodic structure is rather coarse. The prosodic structure of prominence and phrasing may be influenced by each other, such that a pitch accent on a given word may be influenced by the presence of a prosodic boundary i.e., rhythmic factors (cf. Selkirk, 1984).

Nevertheless, the study of inter-speaker consistency as reported here provides us with some revealing insights: First, the high rate of consistency for the presence or absence of boundary tone indicates that despite the observed inter-speaker variation, there must be constraints imposed on the determination of prosodic phrasing. A speculation is that the syntactic phrasing, though not isomorphic, plays a significant role in determining the location of prosodic phrase boundaries. If prosodic phrasing were isomorphic with syntactic phrasing, the consistency would be 100%. The syntactic phrases may provide a placeholder for prosodic phrases to land and other various factors may decide the level and type of prosodic phrases.

The role of syntactic phrases as a placeholder may explain why the consistency of putting prosodic boundaries at certain locations in an utterances by multiple speakers is higher than the consistency in putting the same levels or the same types of prosodic boundaries. That is, the relatively high rate of 89.71% for choosing the presence or absence of prosodic boundary, compared to that of 79.25% for the consistency of choosing the level of prosodic boundary, may be an indirect indicator that there is more restriction for a given speaker in determining the presence or absence of a boundary than there is for choosing either levels or types of prosodic phrase boundary.

The role of syntactic phrases as a placeholder may also explain why both the agreement rate of inter-transcribers and the rate of speakers' consistency are high at least in locating the prosodic phrasing boundaries. If we regard the inter-transcribers' agreement as a measure of listeners' consistency in identifying prosodic phrases, and the speakers' consistency as a measure in rendering prosodic phrases, the high rate of both measurements can be interpreted as a close relationship between production and perception on the places of prosodic phrases. And the close relationship might be mitigated by the role of syntactic phrases as a placeholder.

If multiple speakers are highly consistent with each other on the way of putting prosodic boundary tones, it may indicate that less variation is observed and stochastic machine learning algorithms relying on frequent use of similar features will achieve high accuracy in predicting prosodic phrases. In fact, the machine learning approach to prosodic phrasing prediction, especially those conducted by Cohen (2004) and Ingulfsen (2004), used data taken from the BURSC. These experiments did not use speaker-dependent features, and obtained high rates of accuracy in predicting the presence or absence of prosodic boundary tones.

The discussion in this paper is motivated by the following points: Prosodic phrasing involved with various factors, which makes it harder to determine the phrasing in deterministic approach. Formalization of prosodic structure is better explained through probabilistic or stochastic approaches than deterministic, algorithmic approaches. The paper presented one database with prosodic labels, together with an analysis of the inter-transcriber reliability and the rate of speaker consistency. One possible source of high rate of consistency in determining the presence or absence is regarded as high correlation between syntactic phrasing and prosodic phrasing. Even if syntactic phrasing and prosodic phrasing is isomorphic, syntactic constituency is thought to provide places on which prosodic phrasing is realized. Given the high rate of intertranscriber reliability, the syntactic boundaries may also play a role in signaling listeners for possible prosodic phrasing boundaries.

It is acknowledged that the data used for the study is contributed by expert speakers who are professional announcers. They may produce the prepared scripts different from naive speakers who are not trained in the same way as radio speakers or voice actors or actresses. A follow-up study can be designed in which the prosodic structure rendered by naive speakers can be compared to that of professional speakers.

References

Bachenko, J. and Fitzpatrick, E. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics* 16: 155-170.

Beckman, M. and Ayers, G. 1997. *Guidelines for ToBI labeling* (version3.0). Manuscript and accompanying speech materials. The Ohio State University.

Black, A. and Taylor, P. 1997. Assigning phrase breaks from part-of-speech sequences. In *Proceedings of Eurospeech* '97, 995-998. Rhodes, Crete.

Bod, R., Hay, J. and Jannedy, S. 2003. *Probabilistic Linguistics*. Cambridge, Mass.: The MIT Press.

Breen, M., Dilley, L. C., Kraemer, J., and Gibson, E. 2012. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* 8: 277-312.

Brenier, M., Cer, D. and Jurafsky, D. 2005. The detection of emphatic words using acoustic and lexical features. In *Proceedings of Eurospeech*, 3297-3300. Lisbon, Portugal.

Charniak, E. 1999. A maximum-entropy-inspired parser. *Brown University Technical Report CS99-12*. Brown University.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Cohen, A. 2004. *A Survey of Machine Learning Methods for Predicting Prosody in Radio Speech*. M.Sc. Thesis, University of Illinois at Urbana-Champaign.

Cole, J., Mo, Y. and Baek, S. 2008. The role of syntactic structure in guiding prosody perception with spontaneous speech. Presented at *Experimental and The Theoretical Approach to Prosody (ETAP)* 2008, Ithaca, NewYork. Cornell University.

Cooper, W. and Paccia-Cooper, J. 1980. *Syntax and speech*. Cambridge, MA: Harvard University Press.

Cutler, A., Dahan, D., and van Doneselaar, W. 1997. Prosody in the comprehension of spoken language: a literature review. *Language and Speech* 40: 141-201.

Dainora, A. 2001. *An empirically based probabilistic model of intonation in American English.* Unpublished PhD dissertation, University of Chicago, Chicago.

Dilley, L., Breen, M., Bolivar, M., Kraemer, J. and Gibson, E. 2006. A comparison of inter-transcriber reliability for two systems of prosodic annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1619-1622, Pittsburgh, PA.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., and Batliner, A. 1996. Consistency in Transcription and Labelling of German Intonation with GToBI. In *Proceedings of International Conference on Spoken Language Processing (ICSLP),* 1716-1719.

Gut, U. and Bayerl, S. 2004. Measuring the reliability of manual annotations of speech corpora. In *Proceedings of the 2nd International Conferenceon Speech Prosody*, 565-568, Nara, Japan.

Ingulfsen, T. 2004. Influence of syntax on prosodic boundary prediction. *Technical Report 610*, University of Cambridge, Cambridge, UK.

Jackendoff, R. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jun, S.-A., Lee, S.-h., Kim, K., and Lee, Y.-J. 2000. Labeler Agreement in Transcribing Korean Intonation with K-ToBI. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)* (Vol. 3, pp. 211-214). Beijing, China.

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

Lyons, J. 1968. *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Manning, C., and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press.

Ostendorf, M., Price P. and Shattuck-Hufnagel, S. 1995. The Boston University Radio News Corpus. B*oston University Technical Report* ECS-95-001.

Pitrelli, J., Beckman, M.E. and Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. Proceedings of *the International Conference on Spoken Language Processing*, 123-126, Yokohama, Japan.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, C. 1991. The use of prosody in syntactic disambiguation. *Journal of the Acoustic Society of America* 90: 2956-2970.

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

Selkirk, E. 1984. *Phonology and Syntax*. Cambridge, Mass.: The MIT Press

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. 1992. TOBI: a standard for labeling English prosody, Proceedings of *the International Conference on Spoken Language Processing (ICSLP)*, 867-870, Banff, Alberta.

Sun, X. 2002. Pitch accent prediction using ensemble machine learning. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 953-956, Denver, Colorado.

Syrdal, A. and McGory, J. 2000. Inter-transcriber reliability of ToBI prosodic labeling, In

Proceedings of the International Conference on Spoken Language (ICSLP), 235-238, Beijing, China.

Wang, M., and Hirschberg, J. 1992. Automatic classification of intonational phrase Boundaries. *Computer Speech and Language* 6: 175-196.

Xu, Y. 2011. Speech prosody: A methodological review." *Journal of Speech Sciences* 1: 85-115.

Yoon, T. 2013. Speaker variation in rendering English prosodic prominence. *The Jungang Journal of English Language and Literature* 55: 465-481.

Yoon, T. 2007. *A Predictive Modeling of Prosody through Grammatical Interface: A Computational Approach*. Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Yoon, T, Chavarria, S., Cole, J. and Hasegawa-Johnson, M. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. *ICSA International Conference on Spoken Language Processing, Interspeech 2004*, 2729-2732, Jeju, Korea.

Yuan, J., Brenier, J. and Jurafsky, D. 2005. Pitch accent prediction: Effects of genre and speaker. In *Proceedings of Eurospeech,* 1409-1412, Lisbon, Portugal.

Tae-Jin Yoon
Department of English Language and Literature
Cheongju University
298 Daeseong-Ro, Sangdang-Gu, Cheongju-Si, Chungbuk, 360-764
E-mail: tyoon@cju.ac.kr

Received: 2014. 02. 22
Revised: 2014. 03. 31
Accepted: 2014. 04. 15