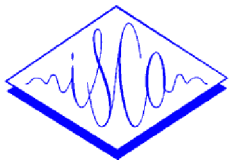

Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus

Tae-Jin Yoon

Department of Linguistics and Languages, McMaster University, Canada

Abstract

An analysis is presented on the rate of inter-speaker consistency in the way multiple speakers realize prosodic events when they read the same scripts. The analysis is made on the Boston University Radio Speech Corpus (BURSC). The BURSC consists of data from five speakers (3 female and 2 male), each reading the same scripts that comprise more than 110 different sentences. The design of the corpus, thus, proves to be a useful basis on which we can measure the degree of speaker variation or speaker consistency in prosodic realization. A pair-wise comparison of inter-speaker consistency is made regarding the rendition of prosodic prominence. The results indicate that the average rate of consistency on the presence or absence of pitch accent is 89.81%. An average consistency of 72.17% is achieved for the rate of consistency for the types of the pitch accent. The finding implies that there is a constraint that is imposed on an utterance by speakers regarding prosodic prominence placement, as well as certain degree of variation between speakers in rendering prosodic prominence.

Index Terms: The Boston University Radio Speech Corpus, ToBI, pitch accents, pair-wise comparison of prosodic prominence

1. Introduction

When different listeners listen to an utterance produced by the same speaker, how consistent are they in their perception of prosodic structure that is encoded in the utterances? In a similar vein, when different speakers are telling the same stories or reading the same scripts, how similar are they in their rendition of prosodic structure? Inter-transcriber reliability study will prove to be useful in studying the first type of question. And studies exist that report the inter-transcriber reliability on prosodic events annotated by trained or naïve labelers [1][2][3][4][5][6]. 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. Probably, there would not be a single instance in which two speakers realize exactly the same prosodic structure phonetically. But different phonetic realization of the intended prosodic structure may be perceived to be the same by listeners. In English as well as in other languages, speech utterances are chunked into smaller prosodic phrases, and within each prosodic phrase, a certain word or syllable may stand out due to the phrasal stress. Many factors contribute to making a word or a syllable stood out in the prosodic phrase, including focus and new information. In this paper, I use the term prosodic prominence to refer to the phrasal stress realized on a certain word or syllable in a prosodic phrase. Due to the lack of one-to-one correspondence between phonetic realization of prosodic prominence and its perception by listeners, researchers often adopt a prosodic annotation system to label the presence and type of prosodic prominence in utterances. In this paper, I report a pair-wise comparison of speaker consistency in rendering prosodic prominence, by analyzing the prosodic labels annotated using the ToBI (Tones and Break Indices) prosody annotation system [7] on a spoken corpus produced by professional radio announcers.

The work reported in this paper is based on the labnews portion of the Boston University Radio Speech Corpus (BURSC) [8]. The labnews portion of the BURSC consists of data from five speakers (3 female and 2 male), each reading the same scripts. The corpus is prosodically annotated using the ToBI system [8]. The design of the corpus, thus, proves to be a useful basis on which we can measure the degree of speaker consistency in prosodic realization. In addition, works exist on investigating the inter-transcriber reliability on portions of the corpus, which is useful in appreciating the degree of constancy in annotating prosodic prominence on the same set of data by different annotators. The paper is organized as follows: I will introduce the BURSC and the ToBI framework for prosody annotation. Then, after providing an informal account of perceptual prosodic labels and their phonetic F0 realization, I will review earlier inter-transcriber reliability studies conducted on the BURSC, and present my analysis of pair-wise comparison results on the speaker consistency on prosodic prominence.

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) [8]. 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. 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 four news scripts were collected in studio recordings, and were later recorded in the laboratory by multiple announcers. The stories represent independent data, covering different topics and a different time period. This BURSC is the richest data set that has prosody annotations, and is 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 [9][10][11][12]. In addition, 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, which is the topic of this paper.

2.2. The ToBI prosodic annotation system

The ToBI (Break and Indices) system is a standard prosodic annotation system [7], and is a variant of the prosodic model originally proposed by Pierrehumbert (1980) [13] and subsequently developed together with her colleagues (cf. [14]). In the ToBI system, two kinds of prosodic information are encoded: tonal information and information on the degree of juncture between words. The analysis in this paper is concerned with the tonal information; hence a brief introduction on the tonal information follows.

The ToBI system transcribes on a tonal tier labels for distinctive pitch events such as pitch accents, phrasal accents, and boundary tones. Pitch accents are marked using a star * at the stressed syllable in the lexical item, and types of pitch accents include H*, L*, L*+H, L+H*, and downstepped H*+!H*. A phrasal accent is assigned either a H-, !H- or L- marker at the phrasal right-edge corresponding to a final high, downstepped high, or low tone, respectively. A boundary tone is marked by either L% or H%. The relatively high boundary tone that sometimes observed at the beginning of an utterance is marked with %H. 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 diacritics are reserved for unspecified or uncertain tonal events, including symbols such as '?', and 'X'. For example, X*? means that a syllable is accented but it is not clear what type of accent must be assigned to the syllable.

2.3. Phonetic F0 contour vs. perceptual prosodic prominence

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. Below I compare the realization of prosodic prominence produced by multiple speakers in terms of the F0 contour with the perceptual prominence as indicated by the ToBI labels.

Probably, there would not be a single instance in which two speakers realize exactly the same prosodic structure phonetically. But the phonetic realization of the intended prosodic structure is not random either.

Figure 1 illustrates raw F0 contours of the phrase "Massachusetts may now…" produced by 5 different speakers (3 female and 2 male) in the radio speech corpus. The corresponding ToBI labels (transcribed by other researchers [8]) are in Table 1. Note that there must be multiple files that are prosodically labeled by each transcriber, given the reliability study previously conducted on this corpus. But the released corpus contains only consensus ToBI labels that transcribers agreed upon. I used the consensus labels for my experiments in this paper. Note also that prosodic labels are missing in some small portion of the corpus. For example, it appears that the speech files of a male radio announcer 1 (M1) are not prosodically transcribed. And some portion of the speech material produced by a female speaker 3 (F3) is missing. So it is not possible to illustrate examples by using utterances produced by the same number of speakers.

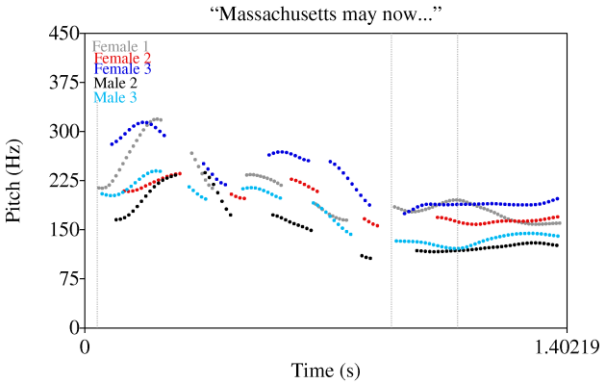

*Figure 1: Overlapped raw F0 contours of the phrase "*Massachussetts may now*…" rendered by 3 female and 2 male speakers. The vertical dotted lines indicate the word boundaries (i.e., "*Massachussetts | may | now")

*Table 1: ToBI labeling of the phrase "*Massachusetts may now *…." In the leftmost column in the table, F and M stands for the gender of the speaker; F for female and M for male.*

	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*

Figure 2 illustrates the raw F0 contours of the phrase "… of the Massachusetts Bar Association…" produced by 4 different speakers (2 female and 2 male) in the radio speech corpus, with the corresponding ToBI labels in Table 2.

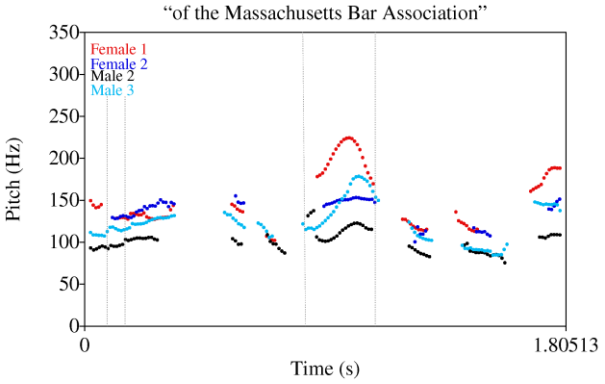

Figure 2: Overlapped raw F0 contours of the phrase "… of the Massachusetts Bar Association*…" rendered by 4 different speakers (2 female and 2 male). The vertical dotted lines indicate boundaries between two adjacent words.*

Table 2: ToBI labeling of the phrase "… of the Massachusetts Bar Association *…"*

	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%

Despite similarity in the F0 contours produced by multiple speakers and higher rate of consistency in transcribed prosodic labels, there is some discrepancy between tune and prosodic transcription. For example, similar shapes can lead to different transcriptions and different shapes may lead to the same transcription. Specifically, in Figure 2, the F0 contours of the word "bar" produced by Male 3 and produced by Female 1 look more similar to each other than the F0 contours produced by other speakers. But, one is transcribed with a rising "L+H*" accent, and the other is transcribed with a plain "H*" accent. The example may be a case where the F0 contour is not in a perfect mapping relationship with a perceptual prosodic event.

There are two sources of a mismatch between an F0 contour and the corresponding labeled tonal event. One is inconsistency in prosodic labeling, for which the studies on transcriber reliability are useful. A couple of previous studies exist that report the inter-transcriber reliability in the ToBI analysis on a small set of the BURSC. Thus I will summarize the previous studies on the inter-transcriber reliability below. The other source of the mismatch is that F0 contours are only one of the properties that determine perceptual prosodic events. It should be noted that while the ToBI label is influenced by the visual display of F0 contours, the system is not designed to be a phonetic transcription system, but a phonological model of intonation. This may be another reason why the phonetic F0 shape and its perceived prosodic event may or may not be in a perfect mapping relation. As to the question of how much the phonetic F0 shape contributes to the prosodic label, prosodic detection modeling using features obtained from those F0 shapes would prove to be useful. For example, Yoon (2007) [13] reports that the presence or absence of pitch accent can be predicted with 73.62% of accuracy when third order polynomial coefficients obtained from the F0 trajectory in the target word are used as the only feature in the classification system.

2.4. Inter-transcriber reliability

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. [1][2][4][6], among others). Two reliability studies have been conducted specifically for the BURSC; One by Ostendorf, Price, and Shattuck-Hufnagel (1995) [8] and the other by Dilley, Breen, Gibson, Bolivar, & Kraemer (2006) [5].

Ostendorf et al. [8] report that the transcriber agreement on the BURSC is relatively high. Transcribers agree on the presence or absence of a pitch accent on a particular word in the test sample 91% of the time, and on the type of pitch accent 60% of the time. Disagreement about pitch accent type is mostly concerned with the choice between H* and L+H* (and !H* and L+!H*). When these two accent types are combined into one category the level of transcriber agreement for accent type rises to 81%. Dilley et al. [5] also report on reliability conducted on a subset of the BURSC, which amounts to 20 minutes, or 5939 syllables. In [5], the transcribers are 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. About 87% is reported for the agreement rate on the presence of a pitch accent, and 80% for the rate on the type of pitch accent.

2.5. Pair-wise comparison of speaker consistency

Consistency is measured as follows: first, prosodic events are aligned for a pair of speakers along each word in an utterance using orthographic words as the time indices, as shown in *Table 3*. Second, the number of prosodic events which the two speakers share in common is counted, and then divided by the total number of words (i.e. 1129). For example, if the task is to compute consistency regarding the presence of pitch accent, then all types of pitch accent (e.g. L+H*, H*, L*+H, etc.) are treated as belonging to the same category "pitch accent.", and the numbers are used in calculating the pair-wise consistency rate on the presence and absence of pitch accent.

Table 3: An example of aligning word-prosody pair of a pair of speakers (e.g. Female 1 and Male 2)

	Female 1	Female 2
…		
of		
the		
Massachusetts		L+H*
Bar	H*	H*
Association	L-H%	L-H%
…		

3. Results

A pair-wise comparison of inter-speaker consistency regarding the rendition of prosodic prominence is reported below: In Table 4 and Table 5, the rates of consistency for all pairs of speakers are reported. In the first two columns, F and M stands for the gender of the speaker (F for female and M for male), and the number next to the F or M indicates speaker index.

Table 4 shows the rates of speaker consistency regarding the presence or absence of pitch accent. The presence or absence of pitch accent is calculated if two speakers have any type of pitch accent on the aligned words. On average, the rate of consistency on the presence or absence of pitch accent is 79.81%.

Table 4: Rate of consistency on the presence or absence of pitch accent for each pair of speakers. Average consistency rate is 79.81%

Speaker A	Speaker B	Ratio	Consistency
F1	F2	912/1129	80.77%
F1	F3	878/1129	77.76%
F1	M2	886/1129	78.47%
F1	M3	897/1129	79.45%
F2	F3	899/1129	79.62%
F2	M2	911/1129	80.69%
F2	M3	904/1129	80.07%
F2	M2	901/1129	79.80%
F3	M3	906/1129	80.24%
M2	M3	918/1129	81.31%

Table 5 presents the pair-wise consistency rate of types of pitch accents. Here, the types are broadly classified to be H*, !H*, L* and no pitch accent, on the basis of the tonal target (i.e. starred tone). Any pitch accents containing H* (i.e. H*, L+H*) and H+!H* are classified to be H*. Note that H*+!H* is treated as H*, not as !H*, because H*+!H* has high tone target preceded by (or a step down from) an even higher pitch [7]. Any pitch accents containing downstepped !H* except H*+!H* (i.e., !H*, L+!H*) are treated as !H*. Finally, both L* and L*+H are treated as members of the L* category. If two speakers share in the production of the broad types of pitch accents, then it is decided that they are consistent in rendering the type of prosodic prominence. Overall, an average consistency of 72.17% is achieved for the rate of consistency for the types of the pitch accent.

Table 5: Rate of consistency on the types of pitch accent for each pair of speakers. Average consistency rate is 72.17%.

Speaker A	Speaker B	Ratio	Consistency
F1	F2	815/1129	72.18%
F1	F3	788/1129	69.79%
F1	M2	788/1129	69.79%
F1	M3	810/1129	71.74%
F2	F3	813/1129	72.01%
F2	M2	813/1129	72.01%
F2	M3	812/1129	71.92%
F2	M2	820/1129	72.63%
F3	M3	842/1129	74.57%
M2	M3	848/1129	75.11%

4. Discussion and Conclusions

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), in addition to or instead of being influenced by the information status such as topic or focus of the word (cf. [15]).

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 pitch accent indicates that despite the observed inter-speaker variation, there must be constraints imposed on prosodic structure. It is also the case that effectiveness in encoding prosodic structure is different among different speakers. Informally, I observed that the male speaker M3 speaks some utterances rather in a slurring manner and that the ToBI transcription of those intervals contains uncertainty or ambiguous labels.

As indicated above, the finding reported in this paper implies that there is a constraint that is imposed on an utterance by speakers regarding prosodic prominence placement, as well as certain degree of variation between speakers in rendering prosodic prominence. It will be beneficial to develop a computational algorithm of predicting prosodic prominence by making use of the constraints. For example, Yuan, Brenier, & Jurafsky (2005) [10] attempt to develop a classifier of predicting the presence or absence of pitch accent by incorporating inter-speaker variation. The aim of their study is to test whether inter-speaker variability have an effect on the task of predicting the presence or absence of pitch accent using non-parametric classification and regression tree (CART) algorithm. The prosodic information of one speaker is trained and the trained model is applied to other speakers.

5. References

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

[2] Syrdal, A. and McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling, In *Proceedings of the International Conference on Spoken Language (ICSLP)*, Beijing, China, pp. 235-238.

[3] Gut, U. and Bayerl, S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Proceedings of the 2nd International Conference on Speech Prosody*, Nara, Japan, pp. 565-568.

[4] Yoon, T., Chavarría, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. *ICSA International Conference on Spoken Language Processing. Interspeech 2004*, Jeju, Korea. Pp. 2729-2732.

[5] 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)*, Pittsburgh, PA, pp. 1619-1622.

[6] Cole, J., Mo, Y., and Baek, S. (2008). The role of syntactic structure in guiding prosody perception with spontaneous speech. Presented at Experimental and Theoretical Approach to Prosody (ETAP) 2008, April 13, 2008, Ithaca, New York. Cornell University.

[7] Beckman, M. & Ayers, G. (1997). *Guidelines for ToBI labeling* (version 3.0). Manuscript and accompanying speech materials. The Ohio State University.

[8] Ostendorf, M., Price P., and Shattuck-Hufnagel, S. (1995). "The Boston University Radio News Corpus," Boston University Technical Report ECS-95-001.

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

[10] Brenier, M., Cer, D., and Jurafsky, D. (2005). The detection of emphatic words using acoustic and lexical features. In *Proceedigns of Eurospeech*, Lisbon, Portugal, pp. 3297-3300.

[11] Yoon, T. (2006). Predicting prosodic boundaries using linguistic features. In Speech Prosody 2006, Dresden, Germany.

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

[13] Pierrehumber, J. (1980). *The Phonology and Phonetics of English Intonation*. Ph.D. dissertation, MIT.

[14] Jun, S.-A. (ed.) (2005). *Prosodic Typology: The Phonology and Intonation and Phrasing*. Oxford: Oxford University Press.

[15] Selkirk, E. (1984). *Phonology and Syntax*. Cambridge, Mass.: The MIT Press