A Comparison of Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices)

Laura Dilley[1]*, Mara Breen*[2]*, Edward Gibson*[2]*, Marti Bolivar*[2]*, John Kraemer*[2]

[1]Department of Psychology, Ohio State University, Columbus, OH, USA
[2]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
dilley.28@osu.edu; {mbreen, egibson, mbolivar, kraemer}@mit.edu

Abstract

Agreement was investigated among five labelers for the use of two prosodic annotation systems: the ToBI (Tones and Break Indices) system [1,2] and the RaP (Rhythm and Pitch) system [3]. Each system permits the labeling of pitch accents and two levels of phrasal boundaries; RaP also permits labeling of speech rhythm and distinguishes multiple levels of prominence on syllables. After training with computerized materials and getting expert feedback, coders applied each system to a corpus of read and spontaneous speech (36 minutes for ToBI and 19 for RaP). Inter-coder reliability was computed according to two metrics: transcriber-syllable-pairs and the kappa statistic. High agreement was obtained for both systems for pitch accent presence, pitch accent type, boundary presence, boundary type, and, for RaP, presence and strength of metrical prominences. Agreement levels for ToBI were similar to those of previous studies [4,5], indicating that participants were proficient coders. Moreover, the high level of agreement demonstrated for the RaP system indicates that RaP is a viable alternative to ToBI for prosodic labeling of large speech corpora.

1. Introduction

Researchers are increasingly recognizing the importance of prosody for both basic research into human speech communication and for the development of automatic spoken language systems. A practical means of assessing prosodic characteristics in speech is through the use of prosodic annotation by human listeners. The ToBI (Tones and Break Indices) system was introduced in the 1990's and has been adopted by a number of research labs. However, since that time questions have been raised about some of the distinctions which are assumed in ToBI [e.g., 6,7,8]. The present paper describes a new prosodic transcription system, the RaP (Rhythm and Pitch) system, which is based on recent empirical and theoretical work in phonetics, psychology, and linguistics. It also presents a test of inter-coder agreement for ToBI and RaP.

The RaP system was developed to fill several outstanding needs in the speech research and linguistics communities. First, recent phonetic and psycholinguistic evidence has suggested that some aspects of the mapping from phonetic attributes to categories of intonational contrast do not correspond precisely to those posited in the original work of Pierrehumbert [9], which forms the basis of ToBI categories. Second, ToBI does not permit the labeling of speech rhythm. However, a large body of research now indicates that speech rhythm is important for language processing by adults and infants [e.g., 10,11], highlighting the need for capturing rhythm in prosodic annotation. In the following we describe the basic components of the ToBI and RaP systems in more detail.

1.1. ToBI

A standard ToBI transcription consists of four tiers of symbolic labels which are time-aligned with the speech signal: an *orthographic tier* for the text transcription, a *tonal tier* for labeling pitch events, a *break index* tier for labeling perceived disjuncture between words, and a *miscellaneous tier* for additional information. The version of ToBI used in the current study also included a fifth tier, termed an *alternative* (or *alt*) tier; alternative choices for tonal and break index labels may optionally be indicated on this tier. Determination of prosodic labels is based both on a coder's perceptual impression of prosodic events, as well as on the visual characteristics of the fundamental frequency (F0) contour. In the following we describe in more detail the tonal and break index tiers, which form the core of a ToBI transcription.

1.1.1. Tonal tier

The tonal tier enables the labeling of two kinds of information: pitch accents and phrasal tones. There are five basic pitch accent types, which can be simple (H*, L*), or complex/bitonal (L+H*, L*+H, and H+!H*). Additionally, there are three "downstepped" accent variants (!H*, L+!H* and L*+!H). In lieu of using the ToBI X*? and *? labels to indicate uncertainty, coders used the *alt* tier to indicate alternative labels. Several labels are also available for indicating hierarchical phrasal information. Three labels (H-, !H-, and L-) indicate pitch movement at a "small" or intermediate intonational phrase boundary, while five complex labels (H-H%, L-L%, H-L%, !H-L% and L-H%) indicate pitch movement at a "large" or full intonational phrase boundary. All labels indicate unidirectional pitch movement, except for L-H%, which generally indicates bidirectional (falling-rising) movement.

Several observations can be made about the tonal inventory in ToBI. First, recent work in phonetics and psycholinguistics has called into question some ToBI categories. For example, H* and L+H* are often confused by trained ToBI labelers [12] and speakers do not distinguish these two categories in production tasks [7,8]. It has also been observed that multiple perceptual and acoustic factors distinguish ToBI tonal labels, making it difficult to define the phonetic properties which correspond to these labels [13]. Finally, there is inconsistency in phonetic exponents of pitch accents, which may be labeled when a pitch excursion is either present or absent. For example, in a stretch of monotone, low-pitched speech for which some syllables are perceived as accented, ToBI prescribes L* pitch accents [2].

1.1.2. Break index tier

A break index is a number from 0-4 which is assigned to the end of each word, building on the work of Price *et al.* [14]. In general, this number indicates the perceived degree of disjuncture between words. A 1 is used to indicate the "normal" degree of disjuncture. A 0 indicates a tight connection between words during fast speech. Moreover, labels of 3 and 4 generally indicate relatively large and maximal disjuncture, respectively.

There are two exceptions to the characterization of break indices as indicating degree of perceived disjuncture. The first stems from the stipulation that a 3 or 4, respectively, must be labeled whenever an intermediate or full intonational phrase tone is labeled on the tonal tier, regardless of the perceived degree of disjuncture. Second, the break index 2 is used to indicate a mismatch between tonal movement and perceived disjuncture. As a result, this label can either indicate a small degree of disjuncture comparable to a 1 or a large degree of disjuncture comparable to a 4 [4,6].

1.2. RaP

The RaP (Rhythm and Pitch) system [3] was developed to meet the needs of the speech research community by building on experimental work and theoretical advances that have taken place since the development of the ToBI system. It is based on the theoretical framework proposed by Dilley [7] as well as work in phonetics and theoretical linguistics [e.g., 8,15-17]. A RaP transcription is based on coders' *perceptual* impressions of prosodic events. Unlike ToBI, a visual display of the F0 contour is considered an aid to labeling, rather than a requirement. A transcription consists of four tiers of symbolic labels which are time-aligned with the speech signal: a *words tier* for indicating the text transcription, a *rhythm tier* for labeling metrical prominences and phrasal boundaries, a *tonal index* tier for labeling tonal information, and a *miscellaneous tier*. In the following discussion we focus on the rhythm and tonal tiers, which form the core of a RaP transcription.

1.2.1. Rhythm tier

The rhythm tier permits the labeling of metrical prominence. Several levels of metrical strength are distinguished. The label X indicates that a syllable is a very strong metrical beat, while x indicates that a syllable is a weaker metrical beat. Uncertainty about the strength and presence of a beat are indicated by X? and x?, respectively. Moreover, phrasal boundaries are labeled on word-final syllables; '))' and ')' indicate major and minor phrase boundaries, respectively. Phrasal labels are based strictly on perceived disjuncture. Finally, uncertainty about the type or presence of a phrasal boundary is indicated by the labels '))?' and ')?', respectively.

1.2.2. Tonal tier

The tonal tier permits labeling of accent-related and phrase-related tonal events. A pitch accent in RaP corresponds to a syllable which carries a beat as well as a pitch excursion; such syllables are labeled with H*, L*, and E* labels (i.e., "starred tones"). By distinguishing syllables which are pitch accented from those which are prominent for strictly rhythmic reasons, RaP provides another means of distinguishing degrees of prominence, in addition to rhythm tier labels. Moreover, tonal movements occurring at metrically weak positions are labeled with H, L, or E labels (i.e., "unstarred tones"). The use of separate labels for starred and unstarred tones is consistent with a growing body of production data on F0 timing [e.g., 15-17]. A '+' is used to indicate association with a preceding or following starred tone. These unstarred tones are also used to indicate phrase-related tonal movement. Finally, '!' indicates a small pitch excursion (i.e., a compressed pitch range), while '?' indicates uncertainty about tonal type or presence.

2. Method

2.1. Corpus

To assess inter-coder agreement for diverse styles of speech, materials were drawn from two speech corpora: a read speech corpus (the Boston Radio News Corpus of professional news broadcast speech, or BRNC [18]), and a spontaneous nonprofessional speech corpus (the CallHome corpus [19]). The amount of speech from each corpus which was labeled in each system is shown in Table 1.

Table 1. *Amount of speech (in minutes and syllables) from each corpus labeled in each system.*

System	Corpus	Minutes	Syllables	Coders/File
ToBI	CallHome	15.2	3680	3.5
	BRNC	20.9	5939	3.4
RaP	CallHome	9.6	2638	4.5
	BRNC	9.6	2889	4.7
	Total	55.2	15146	4.0

2.2. Procedure

Five naïve undergraduate students were hired to participate in the project; none had any previous prosodic annotation experience or phonetic training.

2.2.1. Training and initial testing of ToBI

Training for ToBI involved reading the associated manual and completing the computerized exercises [2], as well as receiving one-on-one feedback from an expert coder (the second author). In addition, all naïve coders participated in bi-weekly meetings with a group of four expert ToBI labelers throughout the course of ToBI training and testing. After two weeks of initial training, the coders annotated one minute of read speech. Feedback from two expert coders (the first two authors) was provided. Subsequently, the coders annotated one minute of spontaneous speech. Again, feedback from the two expert coders was provided.

After these two feedback rounds, the coders labeled 90 seconds of speech (60 seconds read, 30 seconds spontaneous). The annotations were evaluated by three expert coders using the following system. One or two points were deducted for each label with which the expert mildly or moderately disagreed, respectively. Three points were deducted when a label was strongly disagreed with and/or presented incorrect ToBI syntax. Experts also employed a subjective grading system ranging from excellent (5) to poor (1), indicating their overall impression of the labels. Three coders received average grades of 4 or higher from all three expert evaluators on both test files and began annotating the corpus. The other two coders received average grades of 3 from the experts, and were instructed to go back through the Guidelines, paying attention to the labels they had misused in the test labels. After another week of training, they also began corpus annotation.

Coders spent the next four weeks annotating 26.7 minutes of the corpus with the ToBI system (11 spontaneous, 15.7 read). The order of files in the corpus was pseudo-randomly determined so that coders would label approximately equal amounts of read and spontaneous speech. Following training and testing of the RaP system (as described below), coders participated in a second ToBI test phase in which they annotated another 9.4 minutes of the corpus using ToBI.

2.2.2. Training and testing of RaP

After this initial period of learning and applying ToBI, the coders spent two weeks learning the RaP system. Coders were introduced to RaP using the guidelines and computerized exercises in Dilley and Brown [3]. Coders initially participated in a week of intensive group training with the manual, and then continued to meet bi-weekly with an expert RaP labeler (the second author) throughout the course of RaP training and testing. After the first week of training, coders annotated a one-minute passage of read speech, and received feedback on their annotations from an expert RaP coder (the first author). Coders then labeled a one-minute passage of spontaneous speech and again received feedback from the expert coder.

After these two feedback rounds, the coders annotated 60 seconds of read and spontaneous speech. The expert RaP coder provided quantitative and subjective scores for their annotations, as described above. All coders received scores of “4” or higher, and were cleared to begin annotating the corpus using RaP.

Coders spent the next four weeks annotating 19.2 minutes of the corpus using the RaP system (9.6 spontaneous, 9.6 read). The files annotated with RaP were a subset of the 26.7 minutes of the corpus labeled in the first four weeks with ToBI.

2.3. Data analysis

2.3.1 Agreement metrics

Two measures of coder agreement were computed for the current study. First, a metric based on *transcriber-syllable-pairs* was computed by determining the total number of pairwise agreements between coders for each syllable, divided by the total number of possible pairwise agreements on all syllables (cf. [5]). Second, the current study also employed the Kappa statistic to correct for chance agreement, which is given by the following:

$$\kappa = (A_O - A_E)/(1 - A_E) \quad (1)$$

where A_O is the observed agreement and A_E is the expected agreement by chance, given the statistical distribution of labels in the population. A kappa statistic of .7 or higher indicates very good agreement. The distribution of labels across the entire corpus for each labeling system served as the basis for A_E, which was used to generate a kappa statistic for each pair of raters. An overall kappa was then determined by averaging the individual kappa scores.

2.3.2 Agreement comparisons

The first analysis concerned the presence of a pitch accent. For ToBI, two coders were said to agree if they each indicated that a syllable had a pitch accent (H*, L*, L+H*, L*+H, H+!H*, L*+!H, L+!H*, L+!H*), or had no pitch accent. For RaP, two coders were said to agree if they each indicated that a syllable had a pitch accent (H*, L*, E*) or had no pitch accent.

The next analysis concerned the type of pitch accent. For ToBI, two coders were said to agree if they each indicated that a syllable had (a) some variety of high pitch accent (H*, L+H*, !H*, L+!H*, H+!H*), (b) some variety of low accent (L*, L*+H, H+L*), or (c) no pitch accent. For RaP, two coders were said to agree if they each indicated that a syllable had (a) a high pitch accent (H*), (b) a low pitch accent (L*), (c) an equal pitch accent (E*) or (d) no pitch accent.

Next, we examined word-final syllables for agreement regarding the presence and type of a phrasal boundary. For ToBI, two coders were said to agree on the presence of a phrasal boundary if both coders indicated (a) an intermediate or full intonational phrase boundary (3, 3-, 3p, 4, 4-, 4p), or (b) no phrase boundary (0, 1, 1-, 1p, 2, 2-, 2p). For RaP, two coders were said to agree on the presence of a phrasal boundary if both coders indicated (a) a phrasal boundary (‘))’, ‘))?’, or ‘)’), or (b) no boundary (‘)?’ or no label).

Agreement on the type of phrasal boundary was also examined. For ToBI, two coders were said to agree on the type of a phrasal boundary if both coders indicated (a) a full intonational phrase boundary (4, 4-, 4p), (b) an intermediate intonational phrase boundary (3, 3-, 3p), or (c) no phrase boundary (0, 1, 1-, 1p, 2, 2-, 2p). In RaP, two coders were said to agree on the type of a phrasal boundary if both coders indicated (a) a large boundary (‘))’, ‘))?’), (b) a small boundary (‘)’ or ‘)?’), or (c) no boundary.

A final agreement analysis which applied only to the RaP system concerned the presence and type of beat (metrical prominence) on a syllable. Two coders were said to agree on the presence of a beat if both coders indicated (a) a beat (X, X?, or x), or (b) no beat (x? or no label). Moreover, two coders were said to agree on the strength of beat if both coders indicated (a) a strong beat (X or X?), (b) a weak beat (x), or (c) no beat (x? or nothing).

3. Results

Table 2 reports agreement related to labeling phrasal boundaries and pitch accents in ToBI and RaP. Agreement is reported in terms of transcriber-syllable-pairs (TSP) and a kappa statistic (Kappa). Table 3 reports agreement for the presence and strength of a beat on a syllable in terms of TSP and Kappa; since only the RaP system permits the labeling of speech rhythm, no values for ToBI are reported.

4. Discussion

High agreement was obtained for both ToBI and RaP for two metrics of inter-transcriber reliability for presence and type of phrasal boundary, and for presence and type of pitch accent.

Moreover, the agreement numbers for ToBI observed here are comparable to those in previous studies [5,6,10], indicating that study participants were proficient coders. In addition, RaP demonstrated somewhat higher agreement than ToBI for presence and type of phrasal boundary. This may be because phrasal boundaries in RaP are based entirely on perceived disjuncture, while phrasal boundaries in ToBI are based on both perceived disjuncture and tonal labels. The two systems perform comparably with respect to presence and type of pitch accent. Finally, the results show that RaP permits reliable coding for speech rhythm.

Table 2. *Agreement for pitch accent and phrasal boundary labels in ToBI and RaP.*

	TSP		Kappa	
	ToBI	RaP	ToBI	RaP
Presence of a pitch accent	87%	86%	0.71	0.71
Type of pitch accent	80%	80%	0.68	0.65
Presence of a phrasal boundary	88%	92%	0.66	0.74
Type of phrasal boundary	76%	84%	0.40	0.61

Table 3. *Agreement for speech rhythm labels in RaP.*

	TSP	Kappa
Presence of a beat	90%	0.80
Strength of a beat	79%	0.65

5. Conclusions

The present paper examined inter-transcriber agreement for two prosodic labeling systems, the ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch) systems. These results demonstrate high agreement for both systems, with somewhat better performance of RaP for labeling phrasal boundaries. Finally, it was demonstrated that RaP permits reliable coding of speech rhythm.

6. Acknowledgements

We gratefully thank Alejna Brugos, Stefanie Shattuck-Hufnagel, and Nanette Veilleux for participating in the ToBI training of the labelers, and for useful discussions during the design of the project and preparation of this paper. We wish to acknowledge Laurie Maynell, Kiwako Ito, and Allison Blodgett for providing helpful feedback on different versions of this paper. We would also like to thank Jennifer Ford, Tess Diduch, Caroline Rubin, and Nakul Vyas for their labeling efforts, as well as Vivek Rao and Serenus Hua for help in piloting the RaP system. Thanks also to Meredith Brown for creating word- and syllable-aligned Praat textgrids for both ToBI and RaP.

7. References

[1] Silverman, K., Beckman, M., Pitrelli, J. Ostendorf, M. Wightman, C. Price, P., Pierrehumbert, J. & Hirschberg, J., "ToBI: A standard for labeling English prosody," *Proc. of the Intl. Conf. on Spoken Lang. Proc*., Banff: Canada, 867-870, 1992.

[2] Beckman, M., and Ayers Elam, G. *Guidelines for ToBI Labelling*. V. 3.0, The Ohio State University, 1997. www.ling.ohio-state.edu/research/phonetics/E_ToBI/.

[3] Dilley, L. and Brown, M. *The RaP Labeling System*, v. 1.0, ms., 2005. http://faculty.psy.ohio-state.edu/pitt/dilley/rap-system.htm.

[4] Pitrelli, J., Beckman, M. & Hirschberg, J., "Evaluation of prosodic transcription labeling reliability in the ToBI framework," *Proc. of the Intl. Conf. on Spoken Lang. Proc*, Yokohama: Japan, 123-126, 1994.

[5] Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M., "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," *Proc. of the Intl. Conf. on Spoken Lang. Proc.,* Nara: Japan*,* 2729-2732, 2004.

[6] Wightman, C., "ToBI or not ToBI?" *Proceedings of the International Conference on Speech Prosody*, Aix-en-Provence: France, 2002.

[7] Dilley, L., "The phonetics and phonology of tonal systems," Ph.D. diss., MIT, 2005.

[8] Ladd, D. R., & Schepman, A., "`Sagging transitions' between high accent peaks in English: experimental evidence," *J. Phonetics*, 31, 81-112, 2003.

[9] Pierrehumbert, J.B., "The phonology and phonetics of English intonation," Ph.D. diss., MIT, 1980.

[10] Nazzi, T., and Ramus, F., "Perception and acquisition of linguistic rhythm by infants," *Sp. Comm*. 41, 233-243, 2003.

[11] Cutler, A., and Norris, D., "The role of strong syllables in segmentation for lexical access," *J. Exp. Psych.: Human Perception and Performance*, 14, 113-121, 1988.

[12] Syrdal, A. and McGory, J., "Inter-transcriber reliability of ToBI prosodic labeling," *Proc. of the Intl. Conf. on Spoken Lang. Proc*., Beijing: China, 235-238, 2000.

[13] Calhoun, S., "Phonetic dimensions of intonational categories: the case of L+H* and H*", *Proc. of Speech Prosody*, Nara: Japan, 2004.

[14] Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C., "The use of prosody in syntactic disambiguation," *J. Acoust. Soc. of Am*., 90(6), 2956-2970.

[15] Arvaniti, A., Ladd, D. R., & Mennen, I., "Stability of tonal alignment: the case of Greek prenuclear accents," *J. Phonetics*, 26, 3-25, 1998.

[16] Ladd, D. R., Faulkner, D., Faulkner, H., & Schepman, A., "Constant `segmental anchoring' of F0 movements under changes in speech rate," *J. Acoust. Soc. of Am*., 106(3), 1543-1554, 1999.

[17] Dilley, L., Ladd, D. R., and Schepman, A. "Alignment of L and H in bitonal pitch accents: Testing two hypotheses," *J. Phonetics*, 33(1), 115-119, 2005.

[18] Ostendorf, M.F., Price P. J., Shattuck-Hufnagel S. "The Boston University Radio News Corpus," Technical Report No. ECS-95-001, Boston University, 1995.

[19] Linguistic Data Consortium "CALLHOME American English Speech Corpus," 1997. http://www.ldc.upenn.edu.