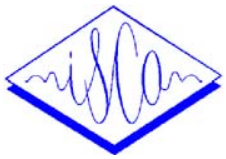
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Intertranscriber Reliability of Prosodic Labeling on Telephone Conversation Using ToBI

Tae-Jin Yoon[1], *Sandra Chavarría*[1], *Jennifer Cole*[1] *& Mark Hasegawa-Johnson*[2]

Department of Linguistics[1]; Department of Electrical and Computer Engineering[2]
University of Illinois at Urbana-Champaign, U.S.A.
{tyoon; chavarri; jscole; jhasegaw}@uiuc.edu

10.21437/Interspeech.2004-659

Abstract

Two transcribers have labeled prosodic events independently on a subset of Switchboard corpus using adapted ToBI (TOnes and Break Indices) system. Transcriptions of two types of pitch accents (H* and L*), phrasal accents (H- and L-) and boundary tones (H% and L%) encoded independently by two transcribers are compared for intertranscriber reliability. Two commonly used methods of reliability measurement, 'transcriber-pair-word' comparison and kappa statistic, are used for comparison with previous reports on the intertranscriber consistency. The results obtained from transcriber-pair-word comparison are: The overall agreement on the presence or absence and choice of pitch accent is 86.57%. The agreement on the presence or absence and the choice of phrasal accent is 85.63%. The presence and choice of boundary tone is 89.33%. When both transcribers agreed that there is at least a phrasal tone, the agreement on the choice of the type of either phrasal accent or boundary tone is 73.86%. The kappa coefficient of agreement (K) of 0.7 to 1 indicates the degree of reliability to be from good to perfect. A kappa coefficient of 0.75 is obtained for agreement on the presence or absence of pitch accents, 0.67 for the presence of phrasal accents, and 0.61 for the strength of disjuncture between phrasal accent and boundary tone. Comparison of the present results with those of previous reliability studies [1][2][3][4] suggests that some higher agreement rates for this study may result from our adoption of fewer labeling distinctions in the transcription of pitch accent events. The results for phrase boundary labeling suggest that spontaneous speech of the type found in the Switchboard corpus is harder to code for the degree of disjuncture between prosodic domains than is read speech.

1. Introduction

Prosodic events that mark phrasal prominence or disjuncture may be encoded variously by phonetic properties such as pitch, intensity, or duration. The non-uniform acoustic expression of prosody presents a challenge for the development of computer speech synthesis and recognition systems, and for fundamental scientific inquiry into the nature of prosody. To fulfill the need for a standard prosodic transcription system and large scale prosodically-labeled speech corpora, the ToBI system has been developed by speech scientists and engineers over the last decade [1][5][6]. Since then, the ToBI system for standard American English has been widely adopted as a standard prosody transcription system. It also has been adapted to other languages like German [7], Japanese [8] and Korean [4], or other variants of English such as Glasgow English [3]. The advantages of using the ToBI system are its reliable intertranscriber consistency [1][3] due to the relatively simple labeling conventions, and its applicability to both prosody-dependent speech recognition requiring large corpora [9] and fundamental research on prosody and spoken language [1][10][11]. Though favorable intertranscriber reliability results have been reported for ToBI-labeled corpora of mainly read speech produced in a laboratory setting or by professional announcers, only a few intertranscriber reliability tests have been reported for large scale spontaneous speech corpora including numerous speakers.

When considering a speaker-independent application of prosody-dependent automatic speech recognition (ASR), a large scale database of multi-speaker, prosodically transcribed corpora is still in demand. In addition, potential differences between spontaneous speech and read speech will diminish the effectiveness of a speaker independent recognition system trained solely on labeled corpora of read speech. In order to build a model of prosody-dependent ASR for spontaneous speech that can make use of high-level (pragmatic) linguistic information, we have undertaken work on the ToBI transcription of spontaneous telephone conversation speech.

In this paper, we report on the intertranscriber reliability of transcriptions produced in our lab by two linguistics graduate students' ToBI labeling on a subset of the Switchboard corpus.

2. Methodology

2.1. Corpus

Switchboard is a corpus of spontaneous informal telephone conversations [12][13]. Prosodic events have been transcribed for files from the WS97 subset of the Switchboard corpus which are segmented by conversational turn and have word- and phone-aligned transcriptions. Two transcribers, graduate students in Linguistics with training in acoustic phonetics and phonology, independently labeled 181 files, containing utterances from 79 different speakers and around 1600 words. The mean duration of the speech files is 3.6 seconds with standard deviation of 2.4 seconds. The overall duration for all files is approximately 9 minutes.

Table 1 shows the number of words, transcribers, and transcriber-pair-word comparison pairs labeled on the Switchboard subset for the present study (ToBI-Swb), along with the corresponding numbers taken from previous studies ([1], [2], [3], and [4], respectively).

2.2. Transcription Procedure

The basic structure of the ToBI system for prosodic transcription consists of 4 separate tiers: the tone tier, the orthographic tier, the break index tier, and the miscellaneous tier. The tone

Table 1: *Number of tokens for reliability test*

	Words	Transcriber	Pairs
ToBI on Swb	1594	2	3188
Original ToBI	489	26	12714
ToBI on Read Speech	644	6	3864
Glasgow ToBI	273	7	1911
K-ToBI	153	21	3213

tier denotes the main prosodic events in terms of pitch accents to mark the perceived prominence of words, and phrasal tones to mark perceived juncture within the utterance. Both pitch accents and phrasal tones are marked with either low (L) or high (H) tone features. Pitch accents are distinguished from phrasal tones by the star diacritic (*) next to L or H. Phrasal tones are further divided into phrasal accents that mark the end of an intermediate phrase, and boundary tones that mark the end of an intonational phrase. The two types of phrasal tones are differentiated by appending to L or H a dash (-) for phrasal accents and percentage sign (%) for boundary tones.

The basic tone elements can be combined using diacritics to denote a complex or more detailed prosodic event. For example, the combination of L and H as L*+H marks a scooped late rise usually found in the context of pragmatic uncertainty [1]. In the present study, pitch accents have been labeled only for the starred tone, collapsing pitch movements with specified leading or trailing tones.[1] Though losing some detail, the labeling of only the starred tone for pitch accents has facilitated labeling with less confusion on the choice of pitch accents. In this regard, it is interesting to observe that Syrdal and McGory [2] report that the distinction between H* and L+H* was the most often confused, accounting for 50% of the disagreements involving either or both of these accents.

Another basic tier of the ToBI system is the orthographic tier. The WS97 subset of the Switchboard corpus includes a word-aligned transcription for the orthographic tier, with additional information like [BREATH], [CROSSTALK], [LAUGH], etc.

Break indices, which comprise another basic tier of the system, have not been labeled. In general, the labeling of break indices is redundant because the information can usually be inferred from the tone tier [1][14].

While not the focal concern of this study, one of the commonly observed phenomena in informal spontaneous speech is disfluency. The transcribers labeled disfluencies in the Switchboard subset, marking the reparandum, editing phase, and alteration on the miscellaneous tier. Disfluencies frequently interrupt the otherwise coherent phrasing in a way that affects the realization of a boundary tone. When an apparent boundary interruption is observed, %r is labeled on the tone tier. The miscellaneous tier can also be used for labeling non-prosodic events such as breathing, cross-talk, or laughter, but these events were already marked in the WS97 transcription for the orthographic tier.

[1]The transcribers optionally used complex bitonal pitch accents when the observed pitch contours were otherwise hard to describe. However, the complex tones were collapsed into either H* or L* for the purpose of the reliability measurements reported here. When perceived prominence was elusive between H* and L*, X* was conservatively used.

3. Reliability Measurement

The measurement of intertranscriber consistency in this study follows as much as possible that of previous studies to facilitate comparisons between studies. The two most commonly reported measurements are transcriber-pair-word comparison and the kappa statistic.

3.1. Pairwise Transcriber Agreement

Since the first study of reliability of ToBI by Pitrelli et al. [1], "a comparison of the labels that transcribers assigned to a word or word boundary in the corpus," called transcriber-pair-word, has been the basic unit of reliability measurement. This pairwise analysis compares the labels of each transcriber against the labels of every other transcriber for each prosodic unit on a word or word boundary, and offers a stringent measure of transcriber agreement. For instance, if three out of four labelers marked a word H*, and one labeler did not mark any pitch accent on that word, the level of agreement is considered to be 3 agreements out of 6, not 3 out of 4. Since there are only two transcribers in our study, the pairwise transcriber agreement results represent only a single comparison pair for each word in the corpus.

3.2. Kappa Statistic

Mayo (1996) [3] states that "while pairwise agreement used by Pitrelli et al. is relatively reliable in that it produces one figure which sums reliably over all coder pairs, it does not take into account the number of possible categories available to the transcriber at any one time." Thus, when considering the number of categories available to the transcriber, the kappa statistic, as in (1), is commonly used as an alternative evaluation of intertranscriber reliability.

$$K = \frac{P_o - P_c}{1 - P_c} \tag{1}$$

where P_o is the percent agreement measured between transcribers and P_c is the agreement that would be predicted by chance.

When the value obtained from the kappa statistic is greater than 0.7, the level of agreement between transcribers is considered to be good.

4. Results

Table 2 shows the distribution of the presence or absence of pitch accent (PA) and the type of pitch accent for the two transcribers in the present study of speech from the Switchboard corpus.

Table 2: *Agreement Matrix of Pitch Accents (Column headings indicate labels assigned by labeler A and row headings are labels assigned by labeler B)*

	H*	L*	X*	No PA	Total
H*	612	14	2	73	700
L*	16	24	0	17	57
X*	9	0	0	4	13
No PA	70	9	0	744	823
Total	707	47	2	838	1594

The choice each transcriber can make on each of 1594 words is 1 out of 3 categories (H*, L*, and no pitch accent).

X* is not counted as a full choice here, as labelers were advised to use this label only in the face of extreme uncertainty. So, the agreement by chance is around 33%. The agreement on whether or not there is a pitch accent, regardless of its type, is 89.14%. The agreement becomes 86.57% when we consider whether both transcribers marked the same pitch accent or both did not mark any pitch accent. The agreement on the pitch accent type when both transcribers agreed that there is a pitch accent is 94.6%.

The agreement obtained in this study is higher than the agreement made in the original ToBI reliability study [1]. In that study, the agreement on the presence or absence of pitch accent is 80.6%, the agreement on the presence and choice of pitch accent is 68%, and agreement on the choice of pitch accent when pitch accent is present is 64.1%.

Table 3 shows the distribution of the presence or absence of phrasal accent (PhA) and the type of phrasal accent for two transcribers.

Table 3: *Agreement Matrix of phrasal accents (Column headings indicate labels assigned by labeler A and row headings are labels assigned by labeler B)*

	H-	L-	No PhA	Total
H-	27	26	14	67
L-	18	298	72	388
No PhA	7	92	1040	1139
Total	52	416	1126	1594

As with pitch accent analysis, the choice each transcriber can make for phrasal accents on each of 1594 words is 1 out of 3 categories (H-, L-, and no phrasal accent). The agreement on the presence or absence of phrasal accent is 88.39%. The agreement on the presence or absence and the choice of phrasal accent is 85.63%. And the agreement on the choice of phrasal accent when both transcribers agreed that there is a phrasal accent is 88.07%.

Though the present agreement results for pitch accent are higher, the results for phrasal accent are comparable to the original ToBI results. In the original ToBI transcriber-pair-wise comparison, the agreement on the presence or absence of phrasal accent is 89.8%, the agreement on the presence and type of phrasal accent is 85%. And the choice of phrasal accent when transcribers agreed on the presence of phrasal accent is 72.9%.

Table 4 shows the distribution of the presence or absence of boundary tone (BT) and the type of boundary tone for the two transcribers.

Table 4: *Agreement Matrix of Boundary Tones (Column headings indicate labels assigned by labeler A and row headings are labels assigned by labeler B)*

	H%	L%	%r	No BT	Total
H%	35	12	0	10	57
L%	2	83	1	30	116
%r	1	1	15	15	51
No BT	29	36	14	1291	1370
Total	67	132	30	1365	1594

A similar analysis of boundary tone to pitch accent and phrasal tone shows the following results: an agreement of 90.4% is obtained for the presence or absence of boundary tone. The presence and choice of boundary tone is 89.33%, and when both transcribers agree that there is a boundary tone, the agreement on the choice of boundary tone is 88.7%.

In the original ToBI reliability study, the agreement rate for the presence or absence of boundary tone is 93.4%. Overall agreement on the presence or absence and choice of boundary tone is 90.9%. When transcribers agreed that a boundary tone is present, the agreement on the choice of boundary tone is 78.7%.

When comparing the agreement strength between phrasal accent and boundary tone, the agreement on the presence or absence of phrasal tone, regardless of its type, is 87.4%. The agreement on the presence or absence and the choice of the phrasal tone is 80.9%. When both transcribers agreed that there is a phrasal tone, the agreement on the level of phrasal tone is 73.86%.

In addition to the transcriber-pair-word analysis, the Kappa statistic is also obtained, as shown in Table 5.

Table 5: *Kappa statistic*

	Kappa coefficient
Pitch accents	
Presence of pitch accent	0.75
Choice of pitch accent	0.51
Phrasal accents and Boundary tones	
Presence of phrasal accent	0.67
Choice of phrasal accent	0.48
Presence of boundary tone	0.58
Choice of boundary tone	0.79
Strength of phrasal tone	0.61

The presence of pitch accent and the choice of boundary tone each have a kappa statistic of over 0.7, which indicates that those categories are reliably labeled in general. However, the agreement on which label is assigned within a pitch accent or phrasal tone category is lower, as is the agreement on the presence of boundary tone, as shown by the relatively smaller kappa coefficients for these measures.

5. Comparison with Previous Studies

The results of the current study have interesting implications when we compare the level of agreement reported above to agreement levels from the previous studies, in particular when we consider the number of categories available to the transcribers.

First, when compared to the level of consistency reported for the original ToBI reliability study, the overall agreement rates on the presence or absence and choice of phrasal accent and boundary tone are quite similar. However, when the overall agreement rate on the presence or absence and choice of pitch accent is compared, our result is much higher than the result from the original ToBI study (86.57% for our result versus 68% for the result from the original ToBI). Our higher agreement rate for pitch accents is almost certainly due to the smaller number of accents distinguished in the modified ToBI labeling performed by our transcribers. While the number of distinct phrasal accents and boundary tones is the same for both studies, the number of distinct pitch accents is limited to two (H* and L*) for our study, but the original ToBI study is based on a transcription that distinguishes more than 6 categories (H*(L+H*), L*+H, !H*(L+!H*), H*+!H*, L*, L*+!H*), even after the merger of L+H* into H* and L+!H* into !H*. The

smaller set of pitch accent choices for the labelers in our study certainly contributes to our higher agreement rate.

Second, our agreement rates may be considered as a measure of the difficulty of the labeling task for telephone conversational speech, and as such can be compared with two of the previous reliability studies with regard to agreement on boundary types. The Syrdal and McGory study [2] reports the reliability of the ToBI labeling on read speech by professional announcers in a near optimal condition, and Mayo's study [3] reports the reliability of a ToBI transcription for spontaneous speech.

The kappa statistics reported by Syrdal and McGory [2] for the agreement on phrasal accent and on boundary tone type are 0.84 for female and 0.76 for male speakers. The corresponding kappa statistics reported in Mayo [3] and in the current study are only 0.7 and 0.61, respectively. These differences suggest that decisions about boundary strength are more difficult for spontaneous speech than for read speech.

6. Conclusion

The intertranscriber agreement for prosodic transcription of spontaneous speech reported in this study exhibits comparable and, for one category, higher levels of reliability when compared with other agreement studies. Differences in the type of speech transcribed, and in the number of categories available to the transcribers, must be taken into account when comparing agreement results.

As stated in the introduction, an efficient and reliable method for prosodic transcription of speech is needed for linguistic research on prosody, in general, and for the development of prosody-dependent ASR systems in particular. Prosodic events are known to condition acoustic (and articulatory) variation in read speech, but less is known about the effects of prosody on spontaneous speech. The results reported here confirm that reliable prosodic transcriptions can be manually produced for spontaneous speech, which in turn opens the door to the future development of automatic prosody transcription using machine learning techniques.

Our study of prosody in the Switchboard corpus demonstrates that the subjective judgment of prosodic events are shared across labelers. In related work, we report on acoustic correlates that differentiate prosodic categories in the Switchboard corpus annotated with our prosody transcription [15], and on the improved performance of prosody-dependent ASR systems trained on prosodically-transcribed speech corpora [9][16].

7. Acknowledgements

This work was funded through the University of Illinois Critical Research Initiative.

8. References

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

[2] Syrdal, A., and McGory, A., "Inter-transcriber reliability of ToBI prosodic labeling", *Proceedings of the International Conference on Spoken Language Processing*, Beijing: China, 235-238, 2000.

[3] Mayo, C.J., *Prosodic transcription of Glasgow English: an evaluation study of GlaToBI*, MSc in Speech and Language Processing, University of Edinburgh, 1996.

[4] Jun, S., Lee, S., Kim, K., and Lee, Y., "Labeler agreement in transcribing Korean intonation with K-ToBI", *Proceedings of the International Conference on Spoken Language Processing*, Beijing: China, 211-214, 2000.

[5] Beckman, M.E., and Ayers, G., *Guidelines for ToBI labelling* (version 3.0), ms., The Ohio State University, 1997.

[6] Beckman, M.E., and Hirschberg, J., *The ToBI annotation conventions*, ms, The Ohio State University and AT&T Bell Telephone Laboratories, 1994.

[7] Grice, M., Reyelt, R., Benzmüller, R., Mayer, J., and Batliner, A., "Consistency in transcription and labelling of German intonation with GToBI", *Proceedings of International Conference on Spoken Language Processing*, Philadelphia, USA. 1716-1719, 1996.

[8] Venditti, J., *Discourse structure and attentional salience effects on Japanese intonation*, PhD Dissertation, The Ohio State University, 2000.

[9] Chen, K., and Hasegawa-Johnson, M., "How prosody improves word recognition", *ISCA International Conference on Speech Prosody 2004*, Nara: Japan, 583-586, 2004

[10] Pierrehumbert, J., and Hirschberg, J., "The meaning of intonational contours in the interpretation of discourse", In *Intonations in Communication*, Cohen, P., Morgan, J., and Pollack, M.E. (eds.), Cambridge, Mass.: MIT Press, 271-311, 1990.

[11] Cole, J., Choi, H., Kim, H., and Hasegawa-Johnson, M., "The effect of accent on the acoustic cues to stop voicing in radio news speech", *Proceedings of International Conference on Phonetic Sciences*, Barcelona: Spain, 2665-2668, 2003.

[12] Godfrey, J.J., Holliman, E.C., and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development", *Proceedings of the International Conference on Audio, Speech and Signal Processing*, 517-520, 1992.

[13] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G., "Stochastic pronunciation modelling from hand-labelled phonetic corpora", *Speech Communication* 29, 209-224, 1999.

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

[15] Chavarría, S., Yoon, T., Cole, J., and Hasegawa-Johnson, M., "Acoustic differentiation of ip and IP boundary levels: Comparison of L- and L-L% in the Switchboard corpus", *ICSA Proceedings of the International Conference on Speech Prosody 2004*, Nara: Japan, 333-336, 2004.

[16] Hasegawa-Johnson, M., Cole, J., Shih, C., Chen, K., Cohen, A., Chavarría, S., Kim, H., Yoon, T., Borys, S., and Choi, J., "Speech recognition models of the interdependence among syntax, prosody, and segmental acoustics", *HLT/NAACL Workshop on Linguistic and Other Higher Level Knowledge in Speech Recognition and Understanding*, Boston: USA, 2004.