

INTER-TRANSCRIBER RELIABILITY OF ToBI PROSODIC LABELING

*Ann K. Syrdal*¹

*Julia McGory*²

¹AT&T Labs - Research, Florham Park, NJ, USA [syrdal@research.att.com]

²Dept. of Linguistics, Ohio State University, Columbus, OH, USA [jmcgory@ling.ohio-state.edu]

ABSTRACT

The goal of this study was to evaluate the reliability among transcribers of a standard prosodic labeling system under relatively optimal conditions of training, supervision, facilities, procedures, and extent of speaker familiarity. The ToBI (Tones and Break Indices) model for standard American English[7][1] was used in the study; break indices indicate the degree of junction between words, pitch accents designate word prominence, and edge tones mark phrase boundaries. The American English speech corpora were read by a female professional speaker and by a male professional speaker, and were composed of several types of texts to ensure prosodic variety. Each of four experienced transcribers independently labeled each corpus. For each corpus, word level agreement in break indices, pitch accents, and edge tones between all possible pairs of transcribers was analyzed, and various statistics were calculated. Agreement among labelers was generally higher than that reported in previous studies[6][3] of larger and more diverse groups of labelers. Agreement was high for some prosodic categories, but low for others. The extent of reliability for various prosodic distinctions has important implications for refining the ToBI model and for limitations in the use of prosody in speech technologies.

1. INTRODUCTION

Large annotated speech corpora are widely used both in linguistic research and in the development and testing of speech technologies. Prosodic transcription of large corpora has become a widely accepted research tool in academic research for the purpose of understanding the distribution and the phonetic realization of phonologically distinct tonal categories and their relationship to the interpretation of utterances in spoken communications. The naturalness and comprehensibility of text-to-speech (TTS) synthesis systems are strongly affected by accurate prosody prediction from text input and by the audible realization of prosody in synthetic speech output. The EToBI annotation tool provides a means for representing prosodic structure, testing theories of English intonation, and extending this knowledge to multiple varieties and styles of spoken English. To accomplish these research goals and to conduct collaborative research, it is important

to have one standard system of analysis consistently used by labelers.

1.1. The ToBI System

The tagging schema used in our study to describe prosodic phenomena is the ToBI model for standard American English, EToBI, roughly following Pierrehumbert's[5] description. A fuller description of the ToBI systems may be found in the ToBI conventions document and the training materials available at <http://ling.ohio-state.edu/phonetics/E-ToBI>. The EToBI system consists of annotations at minimally three time-linked levels of analysis, including an orthographic tier of time-aligned words; a break index tier indicating the amount of disjuncture between words; and a tone tier, where pitch accent and edge tone labels denote the contrastive H (high) and L (low) elements in the intonational contour. The tone tier includes three types of tonal events: L- and H-phrase tones that mark the end of minor or "intermediate" phrases, L% and H% boundary tones that mark the end of major or "intonational" phrases, and L*, H* and complex pitch accents that identify prominent words within an utterance. In addition, the H-phrase tone and several pitch accents can be produced within a "downstepped" compressed pitch range. Break indices ranging from 0-4 are used to identify the amount of perceived disjuncture between words. Index 0 corresponds to minimal, and 4 to maximal disjuncture.

1.2. Previous Studies of Reliability

Two previous evaluations of reliability have measured inter-transcriber consistency for ToBI-based annotation systems: EToBI transcriptions of English utterances [6] and GToBI transcription of a German corpus [3]. These studies are similar to one another in that a corpus (489 words in the English corpus and 733 words in the German database) representing a variety of speakers, dialects, and speech styles was annotated from multiple sites. A large number of transcribers with a variety of backgrounds and experience labeled from different sites (26 transcribers in the English corpus and 13 transcribers in the German). The goal of these studies

was to assess whether the annotation standard and its training materials were adequate for use in large-scale annotation projects.

Transcribers in this study were more uniformly trained and supervised than in the prior studies, and they were more familiar with the single speaker in each corpus. Because of the more optimal conditions of the present study, it may be viewed as measuring transcriber reliability under best-case circumstances.

2. METHOD

2.1. Speech Data

Two subsets from a larger corpus of American English utterances, one spoken by a female professional speaker, and the other by a male professional speaker, were each ToBI transcribed by four labelers. The utterance subsets were read from identical text, which consisted of 149 words from business news articles and 495 words from interactive prompts used in telephone services. Various types of texts were included to ensure prosodic variety in the test set.

2.2. Labeler Experience and Training

Transcribers were five linguistics graduate students and one postdoctoral researcher. Before labeling the files included in the inter-transcriber consistency tests, all labelers worked through the “Guidelines to ToBI Labeling” [1]. In addition, the most experienced ToBI labeler (JM) supervised the project and was assigned to train each of the other labelers. Training took place iteratively during the first two weeks of a labeler’s assignment to the project. After multiple sessions of individualized training, each worked alone on several files, before the next tutorial sessions, during which JM checked labels and answered questions regarding the files. After this initial training, labelers met as a group every week throughout the course of the investigation to discuss problematic files. After nine months of prior labeling experience with the female speaker and eight weeks of prior labeling experience with the male speaker, a set of utterances were labeled for inter-transcriber comparisons.

3. RESULTS

3.1. Data Analysis

The measurement of inter-transcriber consistency followed as much as possible that of the two previous studies [6][3] so that comparisons between them were possible. A comparison of the labels that two transcribers assigned to a word or word boundary in the corpus (the “transcriber-pair-word”) was the basic unit of analysis. However, on the rare occasions in the current study when more than one pitch accent was assigned to a polysyllabic word, its constituent metrical feet were used as the units of analysis. Results from the female and male corpora for pitch accents, edge tones, and break

indices were analyzed separately.

Since there were four transcribers for each corpus studied, there were six possible pairwise comparisons among them. If three of four labelers agree on a label for a word, transcriber-pair-word agreement is only 50%. Similarly, if two of the four transcribers agree on the transcription of a word, pairwise agreement is 17%.

The following statistics were calculated for pitch accents, edge tones, and break indices:

Pairwise transcriber agreement was calculated for each label category and for presence versus absence of tones.

Most common confusions were reported.

Cohen’s kappa [2] values test the agreement observed against the degree of agreement expected by chance. Agreement data alone may be misleading because higher agreement would be expected by chance for more frequently assigned categories. Observed κ values are classified according to [4].

Pearson’s Chi-square tests compared female versus male distributions of both rate of incidence (frequency of occurrence) and pairwise transcriber agreement. In all cases, the two speakers’ distributions showed statistically reliable differences.

Individual labeler consistency was calculated between each labeler and the other three.

3.2. Pitch Accents

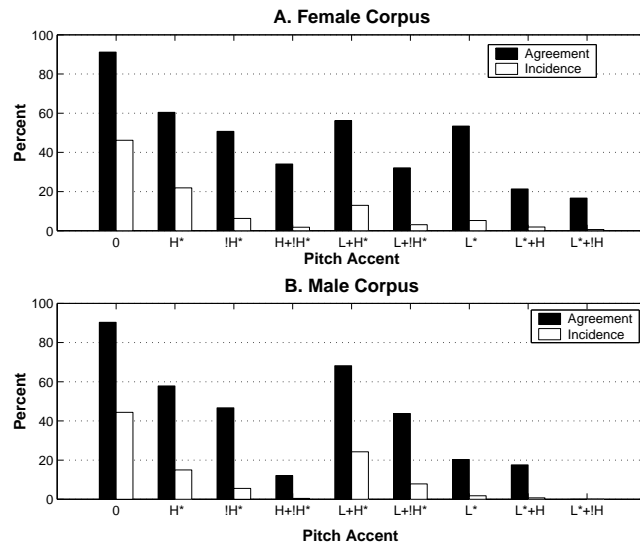


Figure 1: Percent Pairwise Transcriber Agreement (dark bars) and Incidence of Occurrence (light bars) of Pitch Accent Types in Female (A) and Male (B) Corpora

Figure 1 shows pairwise agreement and incidence of occurrence for each pitch accent for the female and male cor-

pora. Overall agreement (the percentage of the total number of transcriber-pair-words in the corpus for which there was agreement) was 71% for the female speaker, and 72% for the male. The corresponding result for the previous EToBI[6] study was 68%, and for the GToBI[3] study, 71%. Agreement for presence versus absence of pitch accent (pooled across accent types) was 92% for the female corpus and 91% for the male corpus. For the previous EToBI study, the corresponding value was 81%, and for the GToBI study, it was 87%.

The more frequently occurring pitch accent types also exhibited higher levels of agreement between transcriber pairs. Cohen’s κ was 0.69 for female pitch accents and 0.67 for male. These values of κ fall within the range classified as ‘substantial,’ meaning that inter-rater agreement was substantially higher than that expected on the basis of agreement on the frequency of category use.

The most common confusion observed among pitch accents was between the two most frequently assigned accents, H* and L+H*, which accounted for one fourth of all pitch accent confusions and half of the confusions involving either or both of these accents.

3.3. Edge Tones

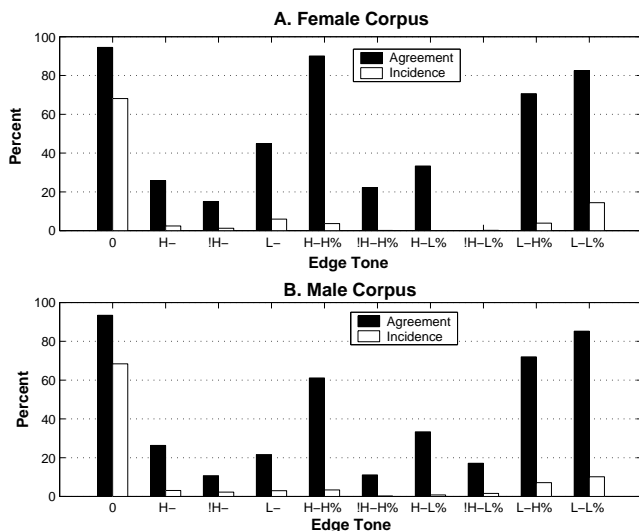


Figure 2: Percent Pairwise Transcriber Agreement (dark bars) and Incidence of Occurrence (light bars) of Edge Tone Types in Female (A) and Male (B) Corpora

Figure 2 shows pairwise agreement and incidence of occurrence for each category of edge tones in the female and male corpora. Overall transcriber-pair-word exact agreement for edge tones (i.e., agreement about the phrase accent and the boundary tone, if any) was 86% for the female speaker and 82% for the male. Cohen’s κ was 0.84 (‘almost perfect’) for the female and 0.76 (‘substantial’) for the male corpus. Overall agreement for edge tones was 85% in the previous EToBI study and 86% in the GToBI study. Agreement on

the presence versus absence of edge tones was 93% for the female speaker and 91% for the male. Presence versus absence agreement for edge tones was 90% in the previous EToBI study, but was not reported in the GToBI study.

The most common confusion among edge tones was between the two most common categories, L-H% and L-L%. It accounted for 11% of all edge tone confusions for the female corpus, and 6% for the male, and for 43% of the confusions involving either or both accents for the female speaker, and 32% for the male. The most common confusion for each of the isolated (non-terminal intermediate phrase) phrase accents was with the case of no edge tone assignment.

Phrase accents and boundary tones were also analyzed individually without regard to their context. That is, pairwise agreement was calculated for phrase accents (0, H-, !H-, and L-) without regard to the presence of or identity of an associated boundary tone, and also for boundary tones (0, H%, L%) regardless of preceding phrase accent. Considered individually, overall agreement for phrase accents was 90%, and for boundary tones, agreement was 97%.

3.4. Break Indices

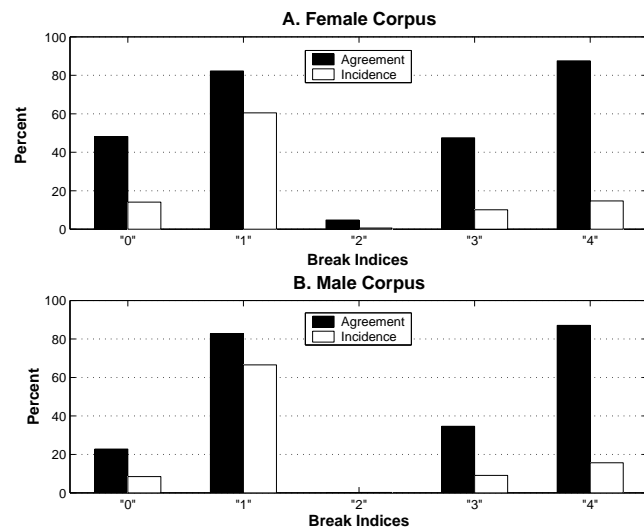


Figure 3: Percent Pairwise Transcriber Agreement (dark bars) and Incidence of Occurrence (light bars) of Break Indices in Female (A) and Male (B) Corpora

Figure 3 shows pairwise agreement and rate of incidence for break indices for the female and the male corpora. Following [6], obligatory (sentence-final) *4s* were deleted from the analyses. Overall transcriber-pair-word agreement for breaks was 74% for the female and 74% for the male speaker. In the previous EToBI study, break agreement was 67%. No break agreement data were reported in the GToBI study.

Cohen’s κ was 0.65 for the female and 0.62 for the male corpus. This indicates that transcriber agreement was ‘sub-

stantially' higher than what would be expected on the basis of chance, given the frequency of category use.

The most common confusion was between *0* and *1*, which accounted for about half of all break confusions and about three fourths of the confusions involving either or both categories. Another common confusion was between breaks *1* and *3*. It accounted for one fourth of all break confusions for the female and one third for the male, and for over half of the confusions involving either or both of these break indices.

3.5. Consistency of Transcribers

FEMALE CORPUS				MALE CORPUS			
ID	TA	BA	RB	ID	TA	BA	RB
CH	90.6	75.4	92.6	CH	88.8	78.9	91.7
JM	89.4	75.7	93.2	ES	90.3	65.3	88.4
KB	91.3	72.5	91.9	JM	88.8	76.2	91.4
LM	89.3	73.3	92.3	TF	90.6	74.8	89.7
\bar{X}	90.2	74.2	92.5	\bar{X}	89.6	73.8	90.3
SD	1.0	1.6	0.5	SD	1.0	5.9	1.5

Table 1: Consistency of Individual Transcribers: Percent agreement of each transcriber (ID) with the other three for combined Tonal Labels (TA = Tone Agreement), Break Indices (BA = Break Agreement), and Break Indices with Relaxed Criteria (RB = Relaxed Break Agreement)

Table 1 lists, in the TA and BA columns, the percent of exact label agreement of each individual labeler with the other three labelers for the tonal tier and for the break index tier in the female and male corpora. Group means (\bar{X}) and standard deviations (SD) are also listed. The GToBI study reported 85% mean exact tone label agreement, and higher overall tone agreement among the three expert GToBI developers (89%) than among the 10 less experienced transcribers (84%). The consistency of individual transcribers for tonal elements in the present study is similar to the GToBI experts, and compares favorably to the previous EToBI[6] tone results (83% mean), even though in that study, the downstep distinction was relaxed (thus increasing agreement).

Transcriber consistency for breaks in the previous EToBI study[6] (92% mean) also used relaxed criteria; differences of one break index (e.g. *0* and *1*) were not counted as mismatches. The RB columns in Table 1 list break tier agreement with equivalent relaxed criteria.

4. CONCLUSIONS

ToBI prosodic transcription reliability was generally higher under the relatively optimal conditions of the present study than in previous EToBI and GToBI studies with larger and more diverse groups of labelers. Across pitch accents, edge tones, and break indices, transcriber-pair-word agreement was substantially higher than would be expected on the basis of chance. Pairwise agreement on the presence versus

absence of a tone was consistently over 90%, which is very high, especially considering the stringent metric of pairwise agreement used for analysis.

Agreement was not evenly distributed among prosodic categories. Pairwise agreement reached or exceeded 50% for only two to four of eight pitch accents, three of nine edge tones, and two of five break indices used in EToBI. Thus, while transcribers agree very well on whether or not a word is prominent or whether or not a phrase boundary follows it, they often do not agree on the identity of the specific tone involved. Manually labeled speech corpora may not be sufficiently consistent for successful training or modeling for recognition or TTS systems. The study provided valuable information about the salience, confusability, and similarity of prosodic categories, which is useful in refining the ToBI system and for using prosody in speech technologies.

5. ACKNOWLEDGMENTS

We gratefully acknowledge Mary Beckman for advising the labelers and overseeing the labeling laboratory, and the team of transcribers: Craig Hiltz, Ken Bame, Laurie Maynell, Liz Strand, and Tim Face.

6. REFERENCES

1. M. E. Beckman and G. A. Elam. Guidelines for ToBI labeling. Guidelines version 3.0, The Ohio State University Research Foundation, 1997.
2. J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.
3. M. Grice, M. Reyelt, R. Benzmuller, J. Mayer, and A. Batliner. Consistency in transcription and labelling of German intonation with GToBI. In *Proc. 4th Internat. Conf. Spoken Language Processing*, volume 3, pages 1716–1719, Philadelphia, 1996. ICSLP.
4. J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.
5. J. Pierrehumbert and J. Hirschberg. The meaning of intonation contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Plans and Intentions in Communications*, pages 271–312. MIT Press, Cambridge, 1990.
6. J. Pitrelli, M. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proc. 3rd Internat. Conf. Spoken Language Processing*, volume 2, pages 123–126, Yokohama, 1994. ICSLP.
7. K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg. ToBI: A standard scheme for labeling prosody. In *Proc. 2nd Internat. Conf. Spoken Language Processing*, pages 867–879, Banff, October 1992. ICSLP.