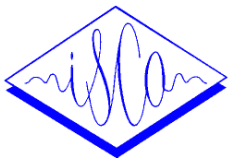

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

EVALUATION OF PROSODIC TRANSCRIPTION LABELING RELIABILITY IN THE ToBI FRAMEWORK

John F. Pitrelli[1], *Mary E. Beckman*[2], *and Julia Hirschberg*[3]

[1]NYNEX Science & Technology, Inc., 500 Westchester Ave., White Plains, NY 10604, U. S. A.;
[2]Ohio State University; [3]AT&T Bell Laboratories

ABSTRACT

A diverse group of speech scientists and engineers has developed the ToBI (TOnes and Break Indices) prosodic transcription system and materials to teach it to transcribers. ToBI consists of parallel tiers reflecting the multiple components of prosody, the most important being a tone tier, for intonational analysis, and a break index tier, for indicating strength of coherence or disjuncture between adjacent words. To assess the system, we measured inter-transcriber agreement on utterances representative of the varied types of speech important to researchers, employing a diverse set of transcribers ranging from experts to newly-trained users. Consistency was measured in terms of number of transcriber pairs agreeing on the labeling of each particular word, a stringent metric. Using this metric, we observe 88% agreement on the presence or absence of a particular category of tonal element, and 81% agreement on the exact label for a tonal category. For break indices, agreement to within one level occurs 92% of the time. We conclude that the ToBI standard and its training materials have been refined to the point that they can be used fruitfully for large-scale annotation of prosodic phenomena in speech databases.

1. INTRODUCTION

Building computational models of prosody is critical both to basic research and to the development of spoken language systems. However, it requires much prosodically-transcribed speech, which in turn requires a transcription system that meets the needs of a diverse group of researchers and technology developers. Such a group has convened over the last three years to develop the ToBI (TOnes and Break Indices) system. ToBI consists of parallel tiers reflecting the multiple components of prosody. In addition to the recorded speech and a direct electronic or paper representation of the fundamental frequency contour, a ToBI transcription of an utterance consists minimally of four tiers of symbolic labels: an orthographic tier, for indicating the words in the utterance; a tone tier, for indicating the basic contrastive elements in the intonational contour; a break index tier, for indicating strength of coherence or disjuncture between adjacent words; and a miscellaneous tier, for indicating various spontaneous-speech effects, such as laughter or hesitations, which are necessary to interpreting the elements on the other tiers.

Analysis on the tone tier assumes a hierarchy of intonational phrases containing one or more intermediate phrases. Three types of tonal events are marked:

1. An **H%** or **L%** boundary tone must mark the end of each well-formed intonational phrase. An optional **%H** can mark the beginning.
2. An **H-** or **L-** phrase accent must go after the last pitch accent in the intermediate phrase. This tone then fills the space until the end of the phrase or the boundary tone.
3. Pitch accents are associated with the stressed syllables of prominent words. There must be at least one (the "nuclear accent" on the word with most prominence in the phrase) in each intermediate phrase. The five types of pitch accent are listed below, with examples of contours in which they commonly occur as nuclear pitch accents:

H*	simple high	(canonical declarative)
L*	simple low	(yes-no question)
L+H*	rising to high from low	(contrastive focus)
L*+H	"scooped" late rise	(pragmatic uncertainty)
H+!H*	fall onto stress	(pragmatic inference)

In addition to the above, there are also elements such as **!H*** and **!H-**. These are the counterparts of **H***, **H-**, etc., with "downstep" or compression of the pitch range marked by "**!**" diacritic.

Analysis on the break index tier is in terms of a hierarchy of perceived disjuncture between words. For the most part, the break index hierarchy and intonational grouping coincide, with:

- **0** between words that have been closely grouped phonetically by the application of fast-speech processes such as flapping of /t/.
- **1** between two different prosodic words
- **3** for an intermediate phrase boundary
- **4** for a full intonational phrase

The exception is break index **2**, which indicates:

- a strong disjuncture marked by a pause or virtual pause, but with no tonal marks; *i.e.* a well-formed tune continues across the juncture

OR

- a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary.

Break indices **1**, **2** and **3** can also take a "**p**" diacritic, indicating various sorts of disfluency. For example, **1p** indicates an abrupt cutoff before a repair or restart, and **3p** indicates a hesitation pause or pause-like prolongation in the portion of an utterance where there is a phrase accent on the tone tier.

Transcribers may express uncertainty by use of "**?**" diacritics on certain tonal labels and "**-**" (minus) on break indices.

Other site-, user-, or task-specific tiers can be added freely, making ToBI flexible and extendible. A description and preliminary assessment of an earlier version of ToBI was presented at ICSLP '92 [1]. In the intervening months, we have refined the transcription system and guidelines for its use, and have developed stand-alone training materials [2] for new transcribers to use in learning the system.

The quantity of data typically required to support much of cutting-edge speech research and the development of robust spoken language systems is growing well beyond what is feasible for labeling at a single site. Therefore, a crucial requirement for a transcription system such as ToBI is that it be learned easily and used consistently by different transcribers with varied backgrounds, working in different locations. Fur-

thermore, because of the varied types of speech being worked on, it is necessary that such inter-transcriber consistency be established across several different databases, representing read speech, spontaneous speech from human-human and from human-machine interaction, and so on.

This paper reports on an experiment in which we measured agreement among 26 subject-transcribers with a variety of backgrounds, working independently at diverse sites. They transcribed a common set of utterances produced in several American English dialects, representing a wide range of speaking styles, from read speech to very informal spontaneous speech in an interview between friends. The results show that ToBI can be used consistently in developing multi-purpose databases.

2. METHODOLOGY

2.1. Subjects

To obtain a pool of diverse subject-transcribers, we solicited volunteers from sites engaged in prosody research. New students of prosody were especially encouraged to participate as well as experts, so that we could determine the sufficiency of the training materials for conveying consistent understanding of the uses of the ToBI symbols and conventions, and not just measure consistency among unrepresentatively expert subjects. Table 1 describes our subject pool. The 26 transcribers were from 14 sites. The subjects span a variety of levels of experience with prosody and experience with ToBI ranging from absolute beginners to contributors to its development. The table also shows that the on-site expertise available to a transcriber for consultation varied greatly from site to site, from three experts in the ToBI system available to the new users at sites A and B, to no expert peer for the new user at site N. Such variations in labeler expertise and availability of expert peers are representative of the challenges to labeling consistency faced by the typical multi-site speech database annotation project.

Table 1: Number of transcribers at each site at each level of experience.

Site	# Experienced with ToBI	# Experienced with Prosodic Transcription but New to ToBI	# New to Prosodic Transcription
A	3	0	2
B	3	0	1
C	1	0	1
D	1	0	0
E	1	0	0
F	0	2	0
G	0	2	0
H	0	2	0
I	0	1	1
J	0	1	0
K	0	1	0
L	0	1	0
M	0	1	0
N	0	0	1
Total	9	11	6

2.2. Training of Subjects

Each transcriber was provided with a document describing the ToBI standard [3], and the ToBI training materials [2]. The training materials contain a short tutorial explaining each of the labels in ToBI, along with recorded examples of transcribed utterances for listening at key points in the tutorial narrative. Interspersed in the tutorial are lists of untranscribed utterances similar to the examples, which the transcribers could use to practice the labels described up to that point in the text. Transcribers were encouraged to discuss these examples with others; however, the training materials are designed to be self-paced, so that the user need not have an expert on site.

2.3. Database

To show that results are applicable to a variety of types of speech, we had the transcribers label a common set of utterances from a variety of speech styles. In order to illustrate the immediate applicability of the results, we chose utterances from widely-used standard databases, whenever such databases were available for a given speech style. We chose 34 utterances totalling 489 words and lasting 148 seconds. Our database consists of four sections: read general-text sentences drawn from the Wall Street Journal database [4], spontaneous utterances from "Wizard of Oz" simulated human-machine interactions with an air-traffic information system (ATIS) [5], spontaneous elicited human-human dialog from the TRAINS database [6], and spontaneous monologue (the "SAILOR database") [7]. A summary of database details is provided in Table 2.

Table 2: Database Components.

Database	# Utterances	# Words	Total Duration (seconds)	# Speakers
WSJ	8	119	40	4
ATIS	9	85	36	7
TRAINS	7	81	25	7
SAILOR	10	204	47	1
Total	34	489	148	19

For each utterance, transcribers were provided with a waveform sampled at 8000 Hz, a F0 contour with an analysis rate of 100 Hz, a time-aligned orthographic transcription, and a dummy break-index file consisting of placeholder Xs at each word boundary except for the obligatory 4 at the end of a non-truncated utterance.

2.4. Transcription Procedure

Transcribers were required to work alone without consultation on the utterances used for the experiment, though they continued to be permitted to discuss other utterances, such as those in the training materials, during the experiment. Transcriptions were done using Entropic Research Laboratory's Waves+ speech analysis software including the `xlabel` attachment. Transcribers were provided with a software tool to check the grammaticality of their transcriptions. The function of this checker was to provide error messages about illegal transcriptions, such as a word boundary without a break index, or an intonational phrase without any pitch accent. This enabled transcribers to correct some "slips of the mouse" and misunderstandings of the details of the system.

2.5. Measuring Inter-Transcriber Consistency

The basic unit for measuring agreement is the *transcriber-pair-word*; that is, a comparison of the labels that two particular transcribers placed on one particular word or at one particular word boundary in the database. The measure of inter-transcriber consistency is then the percentage of transcriber-pair-words exhibiting agreement on a particular element in the transcription. For example, consider the following labelings from four transcribers for an utterance:

Orthography:	Book	the	first		flight	for	me.		
Tr. 1: Tones:	H*		H*		H*		L-	L%	
Breaks:	1	1	1		1	1			4
Tr. 2: Tones:	H*		H*		H*		L-	L%	
Breaks:	1	1	1		1	1			4
Tr. 3: Tones:	H*		H*		!H*		L-	L%	
Breaks:	1	1	1		1	1			4
Tr. 4: Tones:	H*		H*		H* L-		H*	L-	L%
Breaks:	1	1	1		3	1			4

Four transcribers yield six transcriber pairs, times six words in the utterance, equals 36 transcriber-pair-words. For pitch accents, two words, "the" and "for," exhibit total agreement on a lack of pitch accent, and two more, "book" and "first," exhibit total agreement on an H* pitch accent, yielding 24 transcriber-pair-words showing total agreement. For "flight" and "me," three of the four transcribers agree with each other but disagree with the fourth. In each case, the three pairs drawn from the three transcribers who agree are counted as agreements, and the three pairs which include the disagreeing transcriber are counted as disagreements. Therefore, we have a total of 30 agreements and six disagreements, scoring 30/36 or 83% agreement. Analyzing similarly for phrase accents, we find five words in total agreement, but one disagreeing transcriber on the word "flight", generating three transcriber-pair-words of disagreement, so the agreement score is 33/36 or 92%. For break indices, there is agreement at every word boundary other than "flight for", for which one transcriber disagrees with the other three, so agreement again appears to be 33/36. However, the final **4** is obligatory (except when the recording is excerpted from an utterance which continues on), and so we do not count the six transcriber-pair-words corresponding to the final **4**s, and therefore we would report break index consistency as 27/30 or 90%.

Note that transcriber-pair-word agreement is a stringent metric: when three of four transcribers agree on a label, agreement on that label is reported to be just 50% because only three of the six pairs drawn from the set of four transcribers agree.

3. RESULTS

A total of 10,754 independent transcriptions of words was collected. This count can be reasonably interpreted as the number of independent data points available for analysis within each element of the transcription;[1] therefore, we have 10,754 observations for analysis of the tonal elements, and, excluding obligatory **4**s, 10,220 break indices. This is somewhat fewer than the maximum possible (26 transcribers times 489 words) because a portion of the monologue was optional and so was not completed by all transcribers, and because a few submitted utterance transcriptions were disqualified due to ungrammaticalities flagged by the checker but not repaired by the transcriber. In any case, we have a data set of reasonable size for drawing statistical inferences.

Assembling all transcriber-pair-words from these transcriptions, we have 117,035 for each tonal element and 110,584 for break indices. All figures below are based on these data set sizes.

3.1. Agreement on Tonal Elements

Of the 117,035 transcriber-pair-words, 68.3% exhibited overall agreement on pitch-accent labeling, meaning that both transcribers marked the same pitch accent or both did not mark a pitch accent. It is informative to break this result down into two types of disagreement: (1) disagreements on whether there is or is not a pitch accent on the word, and (2) disagreements about the pitch accent type when both transcribers agreed there is a pitch accent. Analyzing this way, there was 80.6% agreement as to whether or not a pitch accent is present. Of those agreements, 42.3% were agreements that there is a pitch accent (the remaining 57.7% were agreements that there is not a pitch accent). When there was agreement that a pitch accent is present, 64.1% of the data points exhibited exact agreement as to which pitch accent was present.[2] While this is a high disagreement rate, the majority of these disagreements fall into a few categories which represent relatively fine distinctions. The largest is ambiguity in application of the downstep diacritic, accounting for 33.5% of the disagreements. Equating across these distinctions, we reach 76.1% agreement when both transcribers agreed that there is a pitch accent, and 72.4% overall pitch accent labeling agreement.

Analyzing phrase accents similarly, we found 85.0% overall agreement on phrase accent labeling (85.3% when we relax the downstep distinction). There was 89.8% agreement on whether or not a phrase accent should be placed; 20.5% of these were agreements that there is one. Of the agreements that there is a phrase accent, 72.9% were agreements as to which one (74.5% when we relax downstep).[3]

Boundary tones showed 90.9% overall agreement. Presence or absence of boundary tones was agreed upon 93.4% of the time; of these, 12.4% were for presence. Of the agreements that a boundary tone is present, 78.8% were also agreements as to which one.[4]

Pooling across the three types of tonal elements, we measure 81.4% overall agreement, or 82.9% relaxing downstep. This breaks down to 88.0% agreement on whether an element of each type was placed. When both transcribers agreed to place a given type of element, they agreed as to which one 69.1% of the time; relaxing downstep raises this agreement to 76.0%.

3.2. Agreement on Break Indices

The analysis of break index labeling consistency is simplified by the fact that there is not an issue of whether or not to place one. Of the 110,584 transcriber-pair words without obligatory **4**s, 66.6% showed exact agreement in the labeling. Relaxing the presence or absence of the diacritics "**p**" and "**-**", we found 70.4% agreement. A standard break index consistency criterion is agreement within +/- 1 level; 74.6% of the disagreements were such close mismatches, so the near-agreement rate is 92.5%.

---

[1] While the 10,754 data points for one element of the transcription are independent of each other, they are not all independent of the corresponding points from another element; for example, the presence or absence of a boundary tone is highly correlated with break index **4**. Therefore, we have 10,754 independent observations rather than four times that many (separating pitch accents, phrase accents, boundary tones and break indices).

[2] The three categories indicating uncertainty or underspecification, *, *? and X*?, have been merged with the most common pitch accent, H*, for this analysis. Also, L+H*, a minor variant of H*, has been merged with H*; similarly, the downstepped counterpart L+!H* has been merged with !H*.

[3] Again, the underspecified label "-" has been equated with the most common label L-.

[4] Again, the underspecified label % has been equated with the most common label L%.

3.3. Consistency Across Transcribers

The above results do not distinguish whether the disagreements were evenly scattered across the transcriber pool or focused on a particular few "outliers" who might have had difficulty learning ToBI. To address this question, for each pair of transcribers we measured percent agreement on tonal elements, thus assembling an agreement matrix for the transcribers. Then, for each transcriber, we averaged the 25 cells reflecting his or her agreement with each of the other 25 transcribers, generating an *agreeability* measure for each of the 26 transcribers. The same analysis was repeated for break indices. In each case, the maximum relaxation discussed above was employed; downstep distinctions, break index diacritic discrepancies, and differences of one break index were not counted as mismatches.

Table 3: Percent Agreeability of each Transcriber. ID = Transcriber's identification number; TA = Tone tier % agreeability, and BA = Break index tier % agreeability.

ID	TA	BA	ID	TA	BA	ID	TA	BA
1	86	93	10	84	93	19	84	94
2	83	92	11	85	93	20	84	93
3	84	92	12	84	93	21	85	93
4	85	93	13	85	94	22	85	93
5	83	92	14	85	94	23	79	90
6	79	94	15	82	90	24	84	94
7	84	93	16	84	94	25	82	93
8	83	93	17	78	89	26	80	91
9	84	94	18	81	89			

Results are shown in Table 3. The relatively small variation among the transcribers' agreeability measures indicates that the training materials are sufficient for building competence in the transcribers; if they were not, the nine ToBI developers would have outstandingly higher agreeability measures simply due to agreeing with each other better than the new users, who merely had the training.

4. DISCUSSION AND CONCLUSIONS

These results show that ToBI now exhibits levels of consistency comparable to or better than those provided in previous transcription systems for prosody. For example, a comparison of consistency between only two transcribers for the considerably smaller set of labels in the prosodic transcription system for the Lancaster/IBM corpus showed 72% agreement for "stress" labels (corresponding roughly to the ToBI tone tier elements) [8] (cited in [9]). Reyelt's [10] more stringent study of intertranscriber consistency for the ten pairs of transcriptions from five transcribers yielded agreement rates of 66%-79% for presence versus absence of "phrase accent" (a category corresponding roughly to the presence of the last pitch accent before a break index **3** or **4** in our system) and rates of 32%-44% for "secondary accent" (roughly the presence of a pitch accent earlier in the phrase).

The levels of inter-transcriber agreement in our study also compare favorably with those of segmental transcription systems, which have served the speech community well. For example, Eisen [11] reports complete agreement among three transcribers to be at about 50% for labelling ten broad categories of segments such as "voiced plosive" in a narrow phonetic transcription. (It is interesting that when the transcribers were asked to transcribe only segments that deviated from the dictionary-predicted transcription, consistency improved to about 85% overall, which is closer to the 70-90% inter-investigator agreement levels reported for example in the literature on child language acquisition, where transcription is a standard investigative tool -- *e.g.*, [12][13].) Given the diversity of databases and transcribers on which our transcriber-consistency statistics were obtained, and the fact that a stringent evaluation metric was used, we conclude that the ToBI convention and its training materials have been refined to the point that they can be used fruitfully for large-scale annotation of prosodic phenomena in speech databases.

REFERENCES

1. Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "TOBI: A Standard for Labeling English Prosody," *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP)*, Banff, Alberta, Canada, October 13-16, 1992, v. 2, pp. 867-870.
2. Beckman, Mary E., and G. Ayers, "Guidelines for ToBI labelling, version 2.0," Manuscript and accompanying speech materials, Ohio State University, 1994. [Obtain by writing to tobi@ling.ohio-state.edu.]
3. Beckman, Mary E., and Julia Hirschberg, "The ToBI Annotation Conventions", Manuscript, Ohio State University, 1994. [Obtain by writing to tobi@ling.ohio-state.edu.]
4. Paul, Douglas B., and Janet M. Baker, "The design for the Wall Street Journal-based CSR Corpus," *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP)*, Banff, Alberta, Canada, October 13-16, 1992, v. 2, pp. 899-902.
5. Hirschman, L., *et al.*, "Multi-Site Data Collection and Evaluation in Spoken Language Understanding," in Bates, M., ed., *Proceedings of the ARPA Human Language Technology Workshop*, March, 1993 (Morgan Kaufmann Publishers, Inc., 1993).
6. Gross, D., J. Allen, and D. Traum, "The TRAINS 91 Dialogues," TRAINS Technical Note 92-1, Computer Science Department, University of Rochester, July, 1993.
7. Interview recorded by Monica Crabtree as part of a sociolinguistic survey of Ohio dialects; no published reference.
8. Alderson, P. and G. Knowles, *Working with Speech* (London: Longman, in press).
9. Bagshaw, P. C. and B. J. Williams, "Criteria for labelling prosodic aspects of English speech", *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP)*, Banff, Alberta, Canada, October 13-16, 1992, v. 2, pp. 859-862.
10. Reyelt, M., "Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German," *Proceedings ESCA Workshop on Prosody 1993*, pp. 238-241.
11. Eisen, B., "Reliability of speech segmentation and labelling at different levels of transcription," *Eurospeech '93: Proceedings of the 3rd European Conference on Speech Communication and Technology*, Berlin, Germany, September, 1993, v. 1, pp. 673-676.
12. Vihaman, M. M., M. A. Macken, R. Miller, H. Simmons, and J. Miller, "From babbling to speech: a re-assessment of the continuity issue," *Language*, v. 61, pp. 397-445, 1985.
13. Davis, B. L., P. F. MacNeilage, "Acquisition of correct vowel production: a quantitative case study," *Journal of Speech and Hearing Research*, v. 33, pp. 16-27, 1990.

ACKNOWLEDGEMENTS

We gratefully acknowledge the help of the following contributors, without whose help this study would not have been possible: James Allen, Gayle Ayers, Paul Bagshaw, Ken deJong, Laura Dilley, Donna Erickson, Esther Grabe, Martine Grice, Peter Heeman, Mary Hendrix, Rebecca Hwa, Christine Nakatani, Francis Nolan, Mari Ostendorf, Janet Pierrehumbert, Scott Prevost, Peter Roach, Inge Rogers, Ken Ross, Stefanie Shattuck-Hufnagel, Kim Silverman, Lisa Stifelman, David Talkin, Nanette Veilleux, Jennifer Venditti, Julie Vonwiller, Colin Wightman, Briony Williams, and Susan Zlotkin. We are grateful to the institutions with which the contributors are affiliated for supporting these efforts. The development of the ToBI standard, and particularly of the training materials, was supported in part by NSF under grant No. SBER-9222685 to Mary Beckman.