From: *Proceedings of the 1992 International Conference on Spoken Language Processing*, 12-16 October, Banff.

TOBI: A STANDARD FOR LABELING ENGLISH PROSODY

Kim Silverman[1], *Mary Beckman*[2], *John Pitrelli*[1], *Mari Ostendorf*[3], *Colin Wightman*[4], *Patti Price*[5], *Janet Pierrehumbert*[6], *Julia Hirschberg*[7]

[1]NYNEX Science & Technology, Inc.; [2]Ohio State University; [3]Boston University; [4]New Mexico Institute of Mining and Technology; [5]SRI International, [6]Northwestern University, [7]AT&T Bell Laboratories

1. ABSTRACT

An understanding of prosody is critical in basic research in speech and natural language processing, and in the technology for building high quality speech synthesis and spoken language understanding systems. Sufficient understanding and development of computational models require large amounts of prosodically transcribed speech. Unfortunately there is no single standard for prosodic transcription that is analogous to IPA for phonetic segments. To meet this need, a group of researchers with expertise in a variety of approaches to prosodic analysis and speech technology have developed TOBI: an agreed transcription system which builds on much recent progress in prosodic modelling. In a study with twenty transcribers with varied experience using this system and a total of 20,000 decisions, high inter-transcriber reliability was achieved. We report this and other evaluations of TOBI which document the consistency with which it can be used. We propose this system as a standard for prosodic transcription of large speech corpora.

2. INTRODUCTION

Prosody is central to the interface between speech and natural language processing technologies. It not only accounts for much of the variability in speech signals, but also conveys much of the information that is necessary for recovering the intended meaning of an utterance — information which is unavailable in orthographic transcriptions. Consequently an understanding of prosody — how it relates both to the acoustic speech signal and to text and discourse structure — is crucial as speech understanding and synthesis technologies progress towards the development of complete spoken language systems that accomplish complex real-world tasks.

Significant progress in quantitative computational modelling of prosody requires very large amounts of prosodically transcribed speech. As the DARPA community has demonstrated [1], multisite shared corpora for spoken language research make available larger amounts of data than any single site can generate, promote reproducibility of results, and enable comparative evaluation. These benefits are particularly important when automatic training techniques are used, or when it is desirable to study naturally-occurring (as opposed to laboratory) speech, while still controlling contextual variability.

But large corpora are of little use unless they are annotated in some way that permits retrieval and analysis of similar phenomena. Such annotation for shared corpora demands agreement on labelling standards. There is a well-established agreement concerning use of IPA for segmental transcriptions, and recently agreement has been achieved concerning core syntactic bracketing [2]. However there has been no analogous consensus concerning how to transcribe prosodic structure. To meet this need, a group of researchers with expertise in prosody have developed a transcription system that meets the following four criteria: (1) reliability: agreement between different transcribers must be at least 80%; (2) coverage: sufficiently comprehensive to capture the most important prosodic phenomena in spontaneous speech; (3) learnability in a relatively short time, in order to be used in multi-site data collections, and (4) capability of being related to current approaches to speech recognition, to parser outputs, and to formal representations of semantics and pragmatics.

This paper describes the development and initial evaluations of reliability of this system. It represents agreement forged among contributors from academia and industry who span major different approaches to transcribing prosody, along with those experienced in exploiting prosodic information for speech recognition, speech synthesis, and computational linguistics. We have held two workshops aimed at coming to agreement on prosodic notation: Victor Zue, who saw the need for and possibility of a single agreed prosodic standard, hosted a workshop at MIT in August 1991, and in April 1992 Kim Silverman hosted a second workshop at NYNEX. We expect that the prosodic labelling system that resulted from these two workshops, which is called TOBI (TOnes and Break Indices), will become a standard for prosodic transcription of most varieties of American English because it provides the following features:

- It captures categories of prosodic phenomena, thereby making it possible to retrieve instances of the same type of event from a large corpus.
- It allows transcribers to represent some uncertainty in their transcriptions, thereby avoiding the drawbacks of forcing every decision to be an all-or-nothing choice.
- It allows researchers to transcribe using subsets or supersets of the notation. This makes it particularly adaptable to various transcription requirements.
- It has demonstrated high inter-transcriber agreement among many different transcribers with a wide range of prior experience. We believe that this unique feature of this standard reflects the careful consideration of the goals of the system and the combined expertise of the contributors.
- It defines ASCII formats for machine-readable representations of the transcriptions which are in principle independent of the pitch extraction and signal display available to the researcher. These formats facilitate sharing and comparison of transcriptions across sites and across hardware and software platforms.
- It is equipped with software to support transcription using a widely-used signal processing software system (Entropic WAVES), and with UNIX programs that check transcriptions for internal consistency and diagnostically flag transcriber errors.

We think that this system, and the process by which it was created, can also provide a useful model for creating comparable standards for prosodic transcription of other varieties of English and for other languages.

3. DESCRIPTION OF THE SYSTEM

3.1. General overview

The system consists of parallel tiers, reflecting that prosody has multiple components. Each tier consists of symbols representing prosodic events, accompanied by the associated time in the utterance where these events occur. The tier which most closely resembles traditional intonational analysis is the *tonal tier*. On this tier we transcribe the tune and a two-level prosodic phrase structure.

In addition to tonal structure, utterances can differ in the way words are grouped or separated by non-tonal means. Sometimes pauses or lengthening can occur between adjacent words within the same intonational phrase. Similarly, the disruption of speech that accompanies an intonational phrase boundary is not solely specified by its tonal makeup. In order to complement the tonal information with a representation of the rhythmic structure of speech, the strength of the coherence or disjuncture between all adjacent words is marked in the *break index tier*.

A third category of variability in speech, which is rare in laboratory recordings but a pervasive problem in real-world applications of speech recognition, is hesitations, disfluencies, breaths, laughs, false starts or restarts, and other spontaneous speech effects. These are often parallel to the other components of prosody: for example, laughter can overlay the articulation of an otherwise well-formed intonational phrase. The onset and offset of these effects are marked in the *miscellaneous tier*. A small set of items is defined for this tier, and a format is specified for users to define their own items.

3.2. Tonal tier

The tune in an utterance is transcribed as a linear sequence of pitch events that are sparsely distributed across the text. These are based on Janet Pierrehumbert's intonational phonology [3] which has been particularly influential over the last decade in such areas as speech synthesis, relationships between prosody and discourse, and laboratory phonology. However, the TOBI transcription standard incorporates a small number of modifications to this system to make the tonal elements slightly less abstract, easier to teach, and easier for automatic recognition.

To summarize these changes: there are five pitch accents (pitch movements or configurations that lend prominence to their associated word), rather than the original six. Specifically, **H*+L**, the downstep-inducing version of **H***, has been deleted. More generally, any downstepped high tones are explicitly marked as such (e.g. **!H***), instead of downstep being implicitly triggered by characteristics of the left-hand context. There is only one initial boundary tone (transcribed as **%H**). There are two levels of phrasing, each with its own boundary tone. Each intermediate phrase receives an indication of its pitch range, by a mark at the time point of the highest F0 value in the highest pitch accent within that phrase.

3.3. Break Index tier

Recent work in exploiting prosody for speech understanding [4] has used a system of "break indices" [5]: a seven-point scale from **0** to **6** of the strength of association between adjacent words. The TOBI standard merges the three highest break indices, which represented intonational phrases and groupings of intonational phrases, into a single category. In addition, definitions of break levels **0** and **2** were modified and made more explicit, as described below, to increase labelling consistency.

4. EVALUATIONS

In preparation for the workshop in April 1992, a set of test utterances were solicited from participants. Some were chosen by their contributors to exemplify prosodic phenomena which would be difficult to transcribe using the 1991 draft version of the TOBI system. Others contained phenomena for which transcriptions ought to be straightforward and noncontroversial. Criteria for submission included that these utterances should represent "real communicative speech", rather than less-realistic recorded citations with unnatural prosodic forms.

Twenty-five utterances from these submissions were then distributed to all participants for transcription prior to the second workshop. These represented a wide variety of speaking styles and scenarios. They included extracts from: radio news broadcasts, narratives, interactions with a simulated airline traffic information system (ATIS), recorded calls to telephone operators (Directory Assistance), interviews, and role-playing to demonstrate prosodic variation.

Each participant, and some of their colleagues and graduate students, transcribed the prosody of the utterances using the draft system and sent in their transcriptions electronically. Twenty transcribers took part. Their transcriptions were first checked via software for grammaticality, and on the basis of the output some transcribers were able to correct some "slips of the mouse" and/or misunderstandings of details of the draft system. Only utterances containing grammatically correct transcriptions were included in the subsequent evaluations: about 10% of the 500 transcriptions (25 utterances x 20 transcribers) were excluded for various technical reasons. In total there were 446 utterance transcriptions.

Agreement was calculated across all possible pairs of transcribers for each word of each utterance. For example, 4 labelers (a, b, c, d) would produce 6 possible transcriber pairs (ab, ac, ad, bc, bd, cd). Our agreement criterion is stringent: if 3 of 4 transcribers (a, b and c) agree, only 3 of 6 pairs will match (ab, ac and bc but not ad, bd and cd) and we would report 50% agreement.

4.1. Tonal transcriptions

Table 1 shows the agreement across transcriber pairs for pitch accents. The first row represents four transcribers who were most experienced with the tonal framework from which the TOBI tonal tier was derived. The second row adds two transcribers with extensive experience in intonational analysis, but within a different theoretical framework[1]. The final row contains the agreement over all transcribers: about one third of them rated themselves as having no previous experience in tonal transcription. Within each row, the first figure is the number of (transcriber pairs x words = "pairs") over which the following proportions were calculated. Three proportions then follow. The first is the proportion of these pairs where the transcribers agreed whether or not there was a pitch accent present. This figure is quite high in each row, and matches well with similar proportions reported in the next section.

The final two figures in each row show agreement on pitch accent types for that subset of pairs where transcribers agreed a pitch

[1] The set of 4 was: Mary Beckman, Julia Hirschberg, Bob Ladd, and Kim Silverman. The set of 6 included Rene Collier and Jacques Terken.

accent was present. The first of these is the agreement on accents matching exactly, with the exception that **H*** was allowed to match **L+H***. The rightmost figure also allows accents to match downstepped versions of themselves (e.g. **H*** matches **!H***).

Table 1: Pitch Accent Agreement

			If there is an accent:	
number of transcribers	N pairs	Is a word accented?	what pitch accent?	+/- downstep
4	973	86%	64%	79%
6	2250	88%	67%	78%
20	37908	83%	61%	73%

Consistency on phrase accents (the tone at the right hand end of intermediate, phrases) is shown in Table 2: 91% agreement on the locations of intermediate phrase boundaries, and between 81% and 89% agreement on the type of accompanying phrase accent when a transcriber pair agreed on an intermediate phrase boundary location.

Table 2: Phrase Accent Agreement

number of transcribers	N pairs	Does a word end a phrase accent?	What type of phrase accent?
4	968	91%	87%
6	2243	91%	89%
20	37840	91%	81%

The third set of items in the tonal tier are the locations and types of full intonational phrase boundaries. Table 3 shows 95% agreement on the locations of the boundaries, and around 90% agreement on the accompanying tone when the pair agrees on a full intonational phrase boundary location.

Table 3: Boundary Tone Agreement

number of transcribers	N pairs	Does a word end in a boundary tone?	What type of boundary tone?
4	968	94%	90%
6	2243	95%	91%
20	37840	95%	89%

Many of the causes for disagreement in the tonal tier were addressed at the second workshop, including distinguishing **L+H*** from **H*** and similarly **L*+H** from **L***, how to detect downstepped accents, and the addition of a **H+!H*** pitch accent to the inventory. Consequently we expect TOBI tonal transcription consistency to be even higher than reported here.

Agreement cannot be compared across the above three tables because of different numbers of categories. But within each table an interesting pattern is that agreement decreased only a small amount as inexperienced transcribers were added to the pool. This is evidence that transcribers can learn TOBI tonal transcriptions very quickly. A similar result was found in the other evaluations described below.

4.2. Break Index transcriptions

Table 4 shows the reliability of the break indices, excluding the obligatory utterance-final **4**'s. The first row represents four transcribers possessing the most prior experience with the system from which this tier was derived[1]. Several of the remaining 16 transcribers rated themselves as having no prior experience in transcribing break indices. The evidence that transcription can be quickly learned is even stronger here than in the tonal tier.

Table 4: Break Index Agreement

number of transcribers	N pairs	Exact match	Agreement within +/- 1
4	1452	69%	94%
20	33636	67%	93%

The first proportion shows those break indices that matched exactly. The second proportion relaxes the match criterion to within 1: for example **1** matched **0**, **1** or **2**, but **0** only matched **0** or **1**. Interestingly, the figures for the exact match and near match (67% and 93%) correspond very closely to those found in [6] (68% and 94%), described below.

As in the tonal tier, the causes for disagreement in this tier were addressed in the second workshop. These included defining more explicitly the difference between **0** (definite phonetic evidence of cliticization) and **1** (normal inter-word boundary), addition of diacritics to mark pauses, and a mechanism to resolve conflicts between break indices and tonally-defined phrasing. As a result, we expect that TOBI break index transcription consistency will also be higher than reported here.

4.3. Other evaluations

Other studies have provided evaluations of the reliability of TOBI transcriptions. In [6], two transcribers with no prior experience in intonational transcription and about one day's training each transcribed 72 utterances from a corpus of telephone speech. The agreement proportions for the tonal structure (including all accents and boundary tones) and break indices were 81% and 68% respectively, for a strict match criterion. When the match criterion was slightly relaxed (e.g. a **H*** pitch accent matches **H*?** — i.e. its "less certain" variant; or a **0** break index matches a **1**) the agreement rose to 92% and 94%.

In a recent experiment conducted by Wightman, 8 subjects marked prominences in 6 minutes of spontaneous speech. They were not given any training and were forced to work very quickly, completing the task in ten times real time (i.e. 1 hour), which is faster than often needed for segmental transcriptions. Nevertheless, they achieved 84% agreement. This matches well with the 83% to 88% range for the corresponding metric in Table 1, above. Moreover, merging the 8 sets of labels with a 3-of-8 criterion produced labels having 90% agreement with labels produced by an expert labeler.

[1] Mari Ostendorf, Patti Price, Nanette Veilleux, and Colin Wightman.

Taken together, the above evaluations indicate which components of TOBI exceed, meet, or approach the original criterion of 80% reliability. Overall, there is 89% agreement on whether each category of tonal element (pitch accent, phrase accent or boundary tone) is present. In those cases where a transcriber pair labels a word with the same tonal category, the "exact-match" criterion yields 72% agreement on which element in the category is present. Relaxing the match criterion to allow differences in downstep, we find 79% agreement, or 84% among the six experienced transcribers. Exact agreement on break indices is 67% for the full set of transcribers, or 93% using a common relaxation criterion. Given the very small amount of training that many of the transcribers received (often less than 1 day), this suggests that the initial learning curve for TOBI transcriptions is very favorable. These evaluations were for the draft version of TOBI. A similar evaluation is currently underway for the current version, in which many of the sources of disagreement have been addressed.

5. SOFTWARE TO SUPPORT TRANSCRIPTIONS

Each tier is stored as a separate ASCII file associated with the corresponding utterance. File formats have been defined for use with the Entropic WAVES signal editing package, and other formats have been defined that are independent of any particular software system and are easily created by hand with a text editor. Scripts built around standard UNIX tools have been developed to convert between these file formats. The formats are intended to be easy to generate with any signal processing system. We have developed scripts that utilize the Entropic software to facilitate transcription: these could also be used as models for other packages. The most widely-appreciated script displays a speech waveform and time-aligned fundamental frequency and energy plots, and has mouse-driven menus of the inventory of items in each tier. Its use makes transcription noticeably easier and faster. Other scripts which use standard UNIX tools check transcription files for grammaticality, and support preparation of utterances for transcription.

6. FUTURE PLANS

This paper has described the development and initial evaluation of a prosodic transcription system that we propose as a standard for the annotation of large corpora. The core group which forged the system is now making plans to collaborate in annotating existing corpora and in building new ones to fill various research needs. We urge other researchers to join us in this effort to provide multi-site corpora that can be used to advance basic linguistic research and to train and test the different components of speech synthesis or spoken language understanding systems. The last two subsections of this paper describe development of a training course and partial automation of the transcription process: steps that we are taking to enhance the value and usefulness of TOBI as a standard.

6.1. Develop a training course

A group led by Mary Beckman is now designing a training course, which will be developed and tested over the next year. The course will consist of a set of transcribed utterances and an accompanying textbook. The set of utterances will demonstrate, with progressively more difficult examples, relevant aspects of the transcription on each tier. When fully tested, the training course will be made available to members of the spoken language processing community in two formats. The electronic format will contain the speech files with attendant F0 and other acoustic parameters, along with the transcriber-assisting script, and a checker program to compare the student's transcription with stored label files providing a "teacher's transcription" that is annotated to point out the difficult points being illustrated by a particular utterance. For those without access to the facilities required for this format, we will also make available audio tape, printed traces of the F0 and other parameters, and a printed "teacher's transcription" using the text-based transcription convention.

6.2. Automation of transcription

For TOBI to be widely used for prosodic annotation, researchers will need to be able to generate the annotations consistently and relatively quickly. Labeling large corpora by hand, however, is both painstaking and prone to fatigue-induced errors. These risks can be reduced through the use of automatic labeling tools which can greatly increase labeling speed, and thus improve accuracy.

Annotation of syntactic structure in large corpora has been made a reality through the use of automatic tools whose output is then hand corrected. Similarly, we expect that the use of automatic labeling algorithms, followed by hand correction, will result in sustainable labeling speeds sufficient to annotate extensive corpora. Algorithms to automatically label break indices and prominences in professionally read speech have already achieved reasonably good performance. These algorithms utilize extensible, statistical methods which can, in principle, be extended to produce the full TOBI label set.

7. REFERENCES

[1] Hirschman, L. *et al.* (MADCOW), Multi-Site Data Collection for a Spoken Language Corpus. These Proceedings. 1992.

[2] Black, E; Abney, S; Flickenger, D; Gdaniec, C; Grishman, R; Harrison, P; Hindle, D; Ingria, R; Jelinek, F; Klavans, J; Liberman, M; Marcus, M; Roukos, S; Santorini, B; & Strzalkowski, T. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. *Proceedings of the DARPA Speech and Language Workshop*, 1991.

[3] Pierrehumbert, J. and Hirschberg, J. The Meaning of Intonation Contours in the Interpretation of Discourse. In P.R. Cohen, J. Morgan & M. Pollack (Eds): *Plans and Intentions in Communication and Discourse.* MIT Press. 1990.

[4] Price, P.J.; Wightman, C.W.; Ostendorf, M.; & Bear, J. The Use of Relative Duration in Syntactic Disambiguation. *Proc. 1990 International Conference on Spoken Language Processing,* 1990.

[5] Price, P.J.; Ostendorf, M.; Shattuck-Hufnagel, S.; & Fong, C. The Use of Prosody in Syntactic Disambiguation. *J. Acoustical Soc. Am.*, December 1991.

[6] Silverman, K.E.A.; Blaauw, E.; Spitz, J.; & Pitrelli, J.F. A Prosodic Comparison of Spontaneous Speech and Read Speech. These Proceedings. 1992.

8. ACKNOWLEDGEMENTS

Contributors to this agreement, excluding the authors of this paper, are: James Allen, Gayle Ayers, Ava Berinstein, Rolf Carlson, Lin Chase, Rene Collier, Nancy Daly, Donna Erickson, Cynthia Fong, James Hieronymus, Sheri Hunnicutt, Bob Ladd, Christine Nakatani, Nina Gronnum, Mari Olsen, Alexander Rudnicky, Stefanie Shattuck-Hufnagel, Liz Shriberg, Judith Spitz, David Talkin, Jacques Terken, Nanette Veilleux, Briony Williams, Victor Zue. We are grateful to MIT and to NYNEX for sponsoring the two workshops, and to the institutions with which the contributors are affiliated for supporting these efforts.