ToBI Or Not ToBI?

Colin W. Wightman

Department of Computer & Information Sciences
Minnesota State University, Mankato
Colin.Wightman@mnsu.edu

Abstract

In the decade that has passed since the introduction of the ToBI system for the transcription of prosody, speech technology has moved out of the laboratory and into commercial applications on several fronts. However, virtually none of the commercial products have made large-scale use of prosody. Nevertheless, researchers in both recognition and synthesis continue to agree that better utilization of prosody is essential to improving the performance and acceptability of commercial systems. In this paper, we review the current state of prosody in commercial systems, and examine how the ongoing discussions related to what and how to transcribe with respect to prosody have simultaneously advanced and inhibited the field. In particular, we argue that, in hindsight, the ToBI system contains several flaws that have limited its acceptance and application.

1. Introduction

In 1992, the ToBI (**To**nes and **B**reak **I**ndices) model for standard American English prosody [19] was introduced. Since then, the ToBI framework has been applied to at least a half-dozen other languages and been used by numerous researchers for work ranging from linguistic research to systems engineering.

The rapid acceptance and widespread application of ToBI occurred for two principal reasons: First, prior to 1992, there were no widely accepted systems for the transcription of prosody that addressed both intonation and phrasing in an integrated way. The second reason was the growing emphasis on computational methods, which were driving dramatic improvements in speech recognition and synthesis technology.

In general, computational methods require the automated analysis of large speech corpora, as compared to the direct observation of a much smaller corpus typical of classical linguistics research. The use of automated analysis tools however, requires that the corpus of interest be consistently annotated with a standard label set. This need, in combination with the requirement for vastly larger corpora, drove the creation of ToBI. The question is, has ToBI really met those needs, or are some changes in order?

In the next section, we will briefly examine some of the few commercial applications that make use of prosody and the lessons that the development of such systems hold for prosody researchers. In the following section, we review the ToBI system, its original vision and the compromises that led to its creation. This is then followed, in section 4, by an examination of some of the prosody transcription systems that have been developed to adapt ToBI to other languages, to cover other phenomena, or to simplify it. Finally, we attempt to identify the flaws in ToBI that have become apparent during the past decade and to answer the question of how best to transcribe prosody for use in technology development.

2. Applications of Prosody

Despite more than a decade of research, and a general consensus that better utilization of prosody will be essential to improving commercial systems, current commercial applications make only minimal use of prosody.

2.1. Recognition

In general, the current commercial applications of automatic speech recognition (ASR) treat prosody as a noise source: one more type of variability that needs to be modeled to improve the accuracy of the acoustic models being used. This is not at all the case in ASR research, however: recent years have seen a great deal of work in this area (cf. [2], [17] and [25]).

Why haven't the research efforts to make better use of prosody in ASR systems had more of an impact commercially? The primary reason is, perhaps, timing: commercial applications have only recently begun to move beyond command, control, and transcription systems. And yet, when developers do want to move beyond recognition of words to also recognize some prosodic attributes that might be useful in some higher-level application, they are confronted by the stark question of exactly what it is that they should be recognizing.

2.2. Synthesis

Commercial deployment of Text To Speech (TTS) systems has accelerated considerably in the past few years, driven largely by dramatic improvements in the synthesized voice quality. Yet even in TTS, where prosodic control has a very measurable effect on perceived quality [23], the provisions for prosodic control are minimal.

Generally, commercial TTS systems provide for either high or low level control of prosody. Low-level control allows (requires) the input text to be annotated to specify the pitch and durational attributes of the text, leaving the question of how these parameters should be set as an exercise for the user. In contrast, high-level control permits the text to be tagged with prosodic events, such as a phrasal boundary or prominence. Developers of TTS systems with high-level control of prosody, however, confront the same question as the ASR developers: what are the appropriate prosodic events?

2.3. Lessons from the commercial systems

If we examine the needs of those who are developing commercial speech technology, several lessons become clear:

- *The need for standardization.* Commercial developers have learned the value of standards: they allow markets to develop rapidly by allaying user fears of being committed to the “wrong” proprietary system. A standard way of describing prosody, both as an ASR output or a TTS input, could allow system components from different venders to be used together and would ensure that materials created for the development of one system could be used in the development of others as well.

- *The need for speed.* With commercial development cycles shrinking from years to weeks, even while the size of the corpora needed to achieve market-leading performance is steadily growing, there is a tremendous need for very rapid labeling of new corpora. If prosody is to be included in new systems, we must be able to label it quickly.
- *The value (and cost) of highly trained individuals.* Graduate-level researchers trained in speech processing, and prosody in particular, are neither plentiful nor low cost. To occupy them primarily to annotate large corpora is difficult to justify: If prosody is to be annotated by humans, it must be done in such a way that non-specialists can be readily trained to do it.
- *The need for real data.* Much of the improved robustness of commercial systems has been due to the use of more realistic data (*i.e.* telephone bandwidth, spontaneous speech, disfluent speech, *etc.*) for system development. However we choose to label prosody, the system must at least allow for, if not explicitly label, the phenomena of real speech such as disfluencies, interruptions, back-channel speech, *etc.*

These lessons are not new: The same observations were made, or anticipated, more than a decade ago. Indeed, it was these observations that motivated the some of the participants involved in the creation of ToBI.

3. The ToBI Transcription System

In 1991, motivated by the need to rapidly label prosody in a standardized way that would allow sharing of annotated corpora, researchers from several academic and commercial organizations began the collaboration leading to development of the ToBI system. Initially focused on standard American English, ToBI has been extended into a more general framework applicable to other languages, as discussed below.

In the following subsections, we will examine the original design goals for ToBI and the ways in which they were met. This is not meant to provide a detailed description of ToBI itself: Such a description, as well as training materials, may be found at http://www.ling.ohio-state.edu/~tobi/ or in [3].

3.1. ToBI Design Goals

Several design goals were established during the initial discussions that led to the creation of ToBI. Of these, the most fundamental were:

- *Transcribe prosody.* Initially, this was a very broadly defined goal but it was gradually focused on stress (prominences) and phrasing.
- *Use “Theory friendly”, machine-readable notation.* This goal was motivated by the desire to be able to share annotated corpora between researchers who may have differing theories of prosody.
- *Transcriptions should be reproducible with good inter-transcriber agreement.*
- *Notation should be extensible to other languages and/or phenomena.*

In addition, two other goals emerged during the development process:

- *Transcribe intonation.*
- *Transcriptions should be “Tool independent”.*

At the time, software tools such as XWAVES™, which could compute pitchtracks and show them time-aligned with the speech waveform and labels, as well as the hardware on which these programs were run, were costly and thus not uniformly available. Not wanting to limit ToBI to those who had access to such computing resources, transcribing the intonation contour in sufficient detail emerged as an additional design goal.

3.2. Transcribing Intonation

Considering the design goals, the transcription of intonation needed to simultaneously fill two roles: (1) It needed to capture the “meaning” of intonational events such as prominent, continuation, final, *etc.*, and (2) it had to describe the shape of the pitchtrack.

To achieve both of these goals simultaneously, the notation developed by Pierrehumbert [14] was adapted. Pierrehumbert’s notation describes the intonation as a series of pitch accents and boundary tones each of which can be either low (L), or high (H). Accents are distinguished by appending a star (*), whereas tones are distinguished by appending either a percentage sign (%) or a minus sign (-), denoting boundary and phrase tones, respectively. By tagging individual syllables with these labels, it became possible to identify perceived prominences and major phrase boundaries by * and %, respectively, while the H and L portions of the labels described the shape of the pitchtrack. The pitchtrack was further described by the use of the ! diacritic to indicate downstepping, and the inclusion of the HiF0 label to mark the location of the peak F0 value in each major phrase.

While the adaptation of Pierrehumbert’s notation met the need for simultaneous transcription and description, as well as the design goal of machine-readability, it inevitably brought some of the associated theoretical assumptions into ToBI, putting the claim to “theory friendly” in some jeopardy. While ToBI is not a direct instantiation of Pierrehumbert’s theory, it does carry some vestiges of that heritage and these have led to some interesting issues.

3.3. Transcribing Phrasing

The transcription of prosodic phrasing is based on the system developed by Price, *et al.* [16]. They proposed a seven-level labeling scheme in which each word boundary is tagged with a *break index* ranging from 0 for a cliticized boundary, to 6 for the strongest boundaries. The emphasis of their system was on capturing the listener’s perception of the phrasal structure: forcing each boundary to be labeled actually seemed to reduce the difficulty of the labeling process, and the seven levels seemed to offer sufficient resolution. Indeed, the seven levels could generally be mapped to various prosodic constituents that had been proposed in the literature.

Nevertheless, during the development of ToBI, as in the literature, the questions of how many levels should be labeled, and what they corresponded to, produced considerable discussion. Eventually, the seven levels were reduced to five as follows:

- 0 – A cliticized boundary.
- 1 – A default prosodic word boundary.

- 2 – A boundary between perceived word groups within an intermediate phrase.
- 3 – An intermediate phrase boundary (one terminated by a phrase accent).
- 4 – An intonational phrase boundary (one terminated by both a phrase accent and a boundary tone).

The evolution of the break indices is interesting for two different reasons. First, it resulted in a move away from the perceptual experience of the listener. That is, the subjective opinion of the labeler that one boundary was stronger than another was de-emphasized in favor of the identification of a specified set of prosodic phrasal constituents. Secondly, as this evolution took place, the description of the phrasing labels began to include reference to the intonational events such as boundary tones. This created a linkage between the phrasal and intonational tiers of the transcription.

3.4. Linkage Rules

The linkage between tiers is described in the labeling guidelines as redundancy [3]. Thus, for example, the presence of a break index of 4 is redundant with the occurrence of a boundary tone label. Indeed, the ToBI guidelines *require* that a break index of 3 and an intermediate phrase accent only be used together. Likewise, the break index of 4 and the intonational phrase boundary tone labels can only be used together.

The restriction that certain break indices can be used only in combination with specific intonation labels is one of the most controversial aspects of the ToBI system. The linkage between the tiers further de-emphasizes the perceptual experience of the listener: The ToBI guidelines even suggest that, once either the tonal or phrasal labels have been produced by the listener, that the redundant labels be inserted automatically to save time and increase inter-transcriber agreement. Indeed, with the 0 break index being reserved for boundaries with clear evidence of clitization, break indices of 3 and 4 restricted to duplicating the phrase accent and boundary tone labels, and break index 1 serving as the default word boundary, there would seem to be little flexibility left with which the labeler can record their perceptual experience.

3.5. Inter-Transcriber Agreement

As described above, one of the original design goals for the ToBI system was a high degree of inter-transcriber agreement. That is, two people labeling the same utterance should produce essentially the same labels. The extent to which this was possible was, in fact, used to help guide the development of ToBI [15]. More recently, this study has been supplemented by [20], in which the labeling done by five graduate students and one postdoctoral researcher, all working in the same lab with the same training, was carefully analyzed. There have also been studies of inter-transcriber agreement in other languages, notably German [7].

The reliability studies have shown remarkable consistency in their results: Even though different researchers have conducted them with different labelers and on different speech materials, pairwise agreement between labelers on the presence versus absence of an edge tone (either phrase accent or boundary tone) has ranged from 85% in [15], to 86% in [7], and 92% in [20]. Similarly, pairwise agreement on the presence versus absence of a prominence (pitch accent) has ranged from 81% in [15], to 87% in [7], and 91% in [20].

However, emphasizing the very high agreement on the presence versus absence of edge tones and pitch accents obscures the much lower agreement on the type of intonational label that should be assigned to each of these events. In [20], for example, pairwise agreement on the specific pitch accent label failed to exceed 50% for six of the eight label types. Likewise, agreement on the specific edge tone label failed to exceed 50% for six of the nine label types. Note that this was under the almost ideal circumstances of very highly and uniformly trained labelers working in laboratory conditions with full access to time-aligned pitchtracks, spectrograms, and waveforms.

It thus appears that, while ToBI is often regarded as having good inter-transcriber reliability, the high levels of agreement are only for a subset of the labeling scheme and that, when the full set of labels is considered, the agreement is really much lower. Moreover, using the full ToBI label set is agonizingly slow: Even for highly trained labelers working under ideal circumstances, full ToBI labeling typically takes 100 to 200 times real time [21].

Consequently, we must question the extent to which ToBI meets its design goal of reproducible transcriptions with good inter-transcriber reliability. More importantly, we must acknowledge that these results are in direct conflict with the needs of commercial systems: the need for speed and low cost, less-highly trained individuals.

4. Post-ToBI Transcription

Since the initial proposal of the ToBI system, it has become extremely widespread. Indeed, even a casual search of the literature will identify hundreds of studies using ToBI. A closer examination of those studies however, reveals that many of the authors are not directly using the full ToBI system *per se*, but are instead using variant systems based on ToBI.

These variants generally fall into one of two categories: The first category contains adaptations designed to adapt ToBI to other languages (e.g. German [8], Chinese [1], Japanese [22], and Korean [11]). The second category, more relevant to the current paper, contains variant systems that are based on reducing the ToBI symbol inventory and/or removing the linkage between the tonal labels and the break indices. Of course, there is also a third category: studies in which the authors have used transcription systems that are not based on ToBI at all (*e.g.* [9]).

The variant systems that use a reduced, ToBI-like label inventory generally make use of the observed inter-transcriber agreement results discussed above and merge several of the label categories. For example, in the German VERBMOBIL project (cf. [13] and [12]), labelers tagged clause boundaries and accented words. The actual tagging was more involved, but the emphasis was on *capturing the listener's perception* of the speech, rather than identifying a fixed inventory of prosodic constituents. As a result, labelers were able to proceed much more quickly and with far less training. Similarly, Greenberg, *et al.* have used a three-level measure of *perceived* accent in their study pronunciation variation [6]. Likewise, Wightman and Rose used labelers with no linguistic background to mark only *perceived* major phrase breaks and strong prominences with similar reported benefits [24]. Additional theoretical support for these merged systems comes from [4], in which it is argued that downstepping is a statistical artifact and should be removed from the phonological inventory of English.

The studies that use completely different systems for labeling prosody generally do so because their authors have found that ToBI does not meet their needs in more fundamental ways. This may be because they reject some part of ToBI for theoretical reasons such as the limited number of phrasing levels or the forced linkage between the tonal and phrasal levels. Other researchers, such as Shriberg, et al. [18], wish to transcribe phenomena not included in ToBI such as disfluencies. Alternatively, many researchers prefer to work with computational models such as those developed by Fujisaki et al. [5] or Holm & Bailly [10].

5. Observations and Conclusions

Researchers have not yet demonstrated a commercially significant advantage to the explicit use of prosody, although the VERBMOBIL project has almost done so [13]. Still, to a technical manager, prosody continues to look like a messy quagmire: The field appears to be fractured into different camps, each with their own way of describing prosody. ToBI, although originally proposed as a standard system, has instead become the standard starting point for developing variant systems.

Nevertheless, a more careful examination of the work being done in the field reveals some important convergence that holds out some hope for resolving the situation. During the past decade, two critical developments have occurred:

- An increasing number of researchers have been rejecting the descriptive nature of ToBI in favor of systems that capture the listener's perceptual experience of an utterance.
- The dramatic reductions in cost for both hardware and software have obviated the need for descriptive labeling: virtually anybody can now get time-aligned waveform, pitchtrack, and spectrogram displays.

Thus the ToBI design goal that prompted the use of Pierrehumbert-inspired labels is no longer important and can be discarded.

Freed of the need to describe the shape of the pitch track, we can now break the forced connection between the tonal and phrasal labels and re-examine what those labels ought to be. To this end, we offer the following guideline:

"Label what you hear."

That is, don't label things that can be gotten for free by simply analyzing the acoustics (like pitchtracks). And don't label things that are required by theory unless a listener clearly hears them as distinctive.

Labeling what we actually hear is the only defensible engineering choice: If we are designing a TTS system, for example, why would we want to develop a system in which two different input tags produced no discernable difference in the output speech? This is what we would be doing if we were to label things that we don't actually hear as different. Likewise, if we use different labels for things that we do not reliably hear as distinctively different, we are likely to develop ASR systems in which we have highly confusable models, likely to be to the detriment of recognition accuracy. If you can't hear a difference, don't label a difference.

Only labeling what we hear also has the potential to significantly increase the speed with which prosody can be labeled in large corpora while simultaneously reducing the training required of the people who will do the labeling, and hence the cost of the labeling. This has already been demonstrated to some extent by, for example, [13] and [24]. Moreover, if we are only labeling real, perceptually distinct phenomena, we are likely to have greater success in developing automatic prosody recognition systems that can reduce labeling costs still further, either through fully automatic labeling as in [23], or through an automatic first-pass followed by manual correction as in [21].

In summary, we have reviewed the original design goals for the ToBI system and how they resulted in the existing framework. We also looked at several lessons from the development of commercial systems and observed that ToBI appeared not to have fully learned those lessons. In particular, we observed that the original design goal of providing a description of the intonational contour has compromised the ToBI system to the point that an increasing number of researchers are creating variant labeling schemes. Finally, we argue that the need for the descriptive component of ToBI no longer exists and that we should only be labeling what we hear: our perceptual experience.

Increasingly, those wishing to label prosody in speech corpora have had to struggle with the question of what labeling system to use, ToBI, or not ToBI? The time has come to ask a different question: what do you hear?

6. References

[1] Aijun, L., 2002. Chinese prosody and prosodic labeling of spontaneous speech. In this volme.

[2] Batliner, A.; Möbius, B.; Möhler, G.; Schweitzer, A.; Nöth, E., 2001. Prosodic models, automatic speech understanding, and speech synthesis: toward the common ground. *Proc. EUROSPEECH.* Aalborg, Denmark, vol. 4, 2285-2288.

[3] Beckman, M.; Elam, G., 1997. Guidelines for ToBI labeling, version 3.0. The Ohio State University Research Foundation.

[4] Dainora, A., 2001. Eliminating Downstep in Prosodic Labeling of American English. In *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding.* Red Bank, New Jersey, USA, 41-46.

[5] Fujisaki, H.; Sudo, H., 1971. Synthesis by rule of prosodic features of connected Japanese. In *International Congress on Acoustics.* Budapest, Hungary, 133–136.

[6] Greenberg, S.; Carvey, H.; Hitchcock, L., 2002. The Relation Between Stress Accent and Pronunciation Variation in Spontaneous American English Discourse. In this volume.

[7] Grice, M.; Reyelt, M.; Benzmüller, R.; Mayer, J.; Batliner, A., 1996. Consistency in Transcription and Labeling of German Intonation with GToBI. In *Proc. Int. Conf. On Spoken Language Processing,* Philadelphia, PA, USA, vol. 3, 1716-1719.

[8] Grice, M.; Baumann, S.; Benzmüller, R.,(to appear). German Intonation in Autosegmental-Metrical Phonology. In: Jun, Sun-Ah (ed.) Prosodic Typology, Oxford University Press.

[9] Hirst, D.; Di Cristo, A.; Espesser, R., 2000. Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (ed) Prosody: Theory and Experiment. Dordrecht, Kluwer, 51-87.

[10] Holm, B.; Bailly, G., 2002. Learning the Hidden Structure of Intonation: Implementing Various Functions of Prosody. In this volume.

[11] Jun, S., 1999. K-ToBI (Korean ToBI) Labeling Conventions. In *Speech Sciences*, vol.7, 143-170.

[12] Kay, M.; Gawron, M.; Norvig, P., 1994. Verbmobil: A Translation System for Face-to-Face Dialog. CSLI Publications, Stanford University, ISBN 0937073962.

[13] Niemann, H.; Noth, E.; Kiessling, A.; Kompe, R.; Batliner, A., 1997. Prosodic Processing and its Use in Verbmobil. In *Proc. ICASSP '97*, Munich, Germany, 75-78.

[14] Pierrehumbert, J., 1980. The Phonology and Phonetics of English Intonation. PhD Thesis, Massachusetts Institute of Technology. Distributed by the Indiana University Linguistics Club.

[15] Pitrelli, J.; Beckman, M.; Hirschberg, J., 1994. Evaluation of Prosodic Transcription Labeling Reliability in the ToBI framework. In *Proc. Int. Conf. On Spoken Language Processing*, Yokohama, Japan, vol. 2, 123-126.

[16] Price, P.; Ostendorf, M.; Shattuck-Hufnagel, S.; Fong, C., 1991. The Use of Prosody in Syntactic Disambiguation. *The Journal of the Acoustical Society of America*, vol. 90, 2956-2970.

[17] Shriberg, E.; Stolcke, A., 2001. Prosody Modeling for Automatic Speech Understanding: An Overview of Recent Research at SRI. *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding.* Red Bank, New Jersey, USA, 13-16.

[18] Shriberg, E.; Stolcke, A.; Baron, D., 2001. Can Prosody Aid the Automatic Processing of Multi-Party meetings? Evidence From Predicting Punctuation, Disfluencies, and Overlapping Speech. *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding.* Red Bank, New Jersey, USA, 13-16.

[19] Silverman, K.; Beckman, M.; Pierrehumbert, J.; Ostendorf, M.; Wightman, C.; Price, P.; Hirschberg, J., 1992. ToBI: A Standard Scheme for Labeling Prosody. *International Conf. Spoken Language Processing*. Banff, Canada, 867-869.

[20] Syrdal, A.; McGorg, J., 2000. Inter-Transcriber Reliability of ToBI Prosodic Labeling. In *Proc. Int. Conf. On Spoken Language Processing*, Beijing, China, vol. 3, 235-238.

[21] Syrdal, A.; Hirschberg, J.; McGory, J.; Beckman, M., 2001. Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody. In *Speech Communication,* vol. 33, 135-151.

[22] Venditti, J., 2000. Discourse Structure and Attentional Salience Effects on Japanese Intonation. PhD Thesis, Ohio State University.

[23] Wightman C.; Syrdal, A.; Stemmer, G.; Conkie, A.; Beutnagel, M., 2000. Perceptually Based Automatic Prosody Labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-To-Speech Syntheis. In *Proc. Int. Conf on Spoken Language Processing*, Beijing, China, vol. 2, 71-74.

[24] Wightman, C.; Rose, R., 1999. Evaluation of an Efficient Prosody Labeling System for Spontaneous Speech Utterances. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Keystone, CO, USA, vol. 1, 333-336.

[25] Wu, S., 1998. Incorporating Information From Syllable-length Time Scales into Automatic Speech Recognition. *Technical Report TR-98-014*, University of California, Berkeley, USA.