The Alternatives (Alt) Tier for ToBI: Advantages of Capturing Prosodic Ambiguity

Alejna Brugos‡, Nanette Veilleux†, Mara Breen, Stefanie Shattuck-Hufnagel***

‡Boston University, †Simmons College,
*University of Massachusetts, **Massachusetts Institute of Technology
abrugos@bu.edu, veilleux@simmons.edu, mbreen@psych.umass.edu, stef@speech.mit.edu

Abstract

Using the alternatives (alt) tier in ToBI transcriptions allows labellers to capture annotation ambiguities explicitly; this labelling innovation allows researchers to address several open research questions concerning prosodic phonology. Furthermore, the standard alt notation allows this data to be shared among researchers and to be machine-readable for the examination of large labelled corpora. Finally, it is anticipated that having an alt tier to record competing labels will facilitate faster and more reliable hand annotation. This paper reports on the use of the alt tier in a large speech sample labelled by 4 expert annotators.

1. Introduction

Mainstream American English (MAE) ToBI (for Tones and Break Indices) [2] is a system for transcribing the intonation and prosodic constituent structure of spoken utterances in MAE. Based on the intonational theory of Pierrehumbert [7] and Pierrehumbert and Beckman [8] and the break index method for labelling constituent structure in Price et al. [11], it specifies both the phonological intonation targets presumed to govern the f0 contour (e.g. High, Low and complex pitch accents, High and Low phrase tones, and High and Low boundary tones), and 5 levels of constituent structure.

A ToBI transcription consists minimally of a recording of the speech, an estimate of its fundamental frequency contour, and (in the transcription proper) symbolic labels for prosodic events. The transcription is usually arranged in four time-aligned parallel horizontal panels or tiers (Fig 1), indicating correspondence between the symbolic labels and the speech waveform. The four labelling tiers are: 1) the Tone tier (tones), for transcribing tonal targets, 2) the Orthographic tier (words), for transcribing words, 3) the Break-Index tier (breaks), for transcribing boundaries between words, and 4) the Miscellaneous tier (misc), for recording additional observations. The Miscellaneous Tier has been used for everything from noting non-speech events to commenting on labelling difficulties. Because much of the notation in this tier has not been standardized, it has been of limited use for drawing conclusions from large labelled corpora.

The growing use of the ToBI framework, for MAE and in developing systems for other dialects and languages [5,3], has revealed several important issues that must be addressed if the goals of the ToBI development community are to be met, e.g. the creation of "a common standard for transcribing an agreed-upon set of prosodic elements, in order to be able to share prosodically transcribed databases across research sites in the pursuit of diverse research purposes and varied technological goals." [2] One set of issues relates to labeller uncertainty, including: 1) Pockets of unreliability: Utterances often contain regions for which there is more than one plausible transcription; when different users select different candidate labels, reliability scores go down, 2) Incomplete capture of information: When a region of an utterance is ambiguous, the labeller is usually considering only two competing analyses, rather than a multitude; when the labeller must specify just one of the two, we lose valuable information about the ambiguity, and 3) Labeller dissatisfaction: The sense of losing information, of making somewhat arbitrary decisions among competing candidate transcriptions, and of spending disproportionate amounts of time on just a few locations is sometimes discouraging to a labeller. The original MAE-ToBI provides some mechanisms for transcribing uncertainty (e.g. X*?; see below for discussion), but these mechanisms have a distinct disadvantage: they do not explicitly capture the alternatives. Instead, they force the labeller to choose between a) marking uncertainty without specifying the alternatives, or b) selecting one of the alternatives without indicating the uncertainty.

This paper introduces a fifth labelling tier, the alternatives (alt) tier, which uses standard machine-readable notation to explicitly capture the alternative transcriptions considered for an ambiguous region of an utterance. The alt tier is designed to ameliorate many of the practical difficulties that labellers have encountered in using ToBI. Even more importantly, regular use of the alt tier provides data that allow researchers to examine locations where labellers considered competing labels, and the alternatives they considered. It distinguishes regions of ambiguity from regions where even controversial labels are clearly appropriate. This makes it possible to address issues that are of importance to the theory of prosodic phonology, such as peak alignment for pitch accents [13] and tone-duration mismatches at phrase boundaries [2].

2. Background

Critics of the ToBI annotation system, and of prosodic annotation systems in general, note that these systems are difficult to learn, slow to use and challenging to use consistently. Wightman [16] summarized these criticisms succinctly: "It 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 labelling 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 labellers working under ideal circumstances, full ToBI labelling typically takes 100 to 200 times real time [15]."

Furthermore, it is claimed that ToBI can be difficult to use in a variety of contexts where listeners perceive ambiguity in the signal. The ambiguity problem was recognized by the original ToBI development community, which provided

standard mechanisms for labelling uncertainty. However, the use of these mechanisms often leads to an underspecified annotation. For example, it is sometimes difficult in compressed pitch ranges to distinguish between a pitch accented syllable and a syllable that merely carries main lexical stress. ToBI provided the *? label which signifies: "I think there's a pitch accent here but I'm not entirely sure". However, this does not capture the labeller's judgment of which type of pitch accent is present, if there is one. In other regions the choice of a label is dependent on preceding labels, so uncertainty about one label may lead to uncertainty about a following label (see section 3.1, below), which cannot be straightforwardly indicated with existing labels.

Ambiguity often leads to labeller disagreement. To avoid this problem, speech engineers interested in including prosodic information in their models to improve speech recognition and synthesis have developed laboratory-specific annotation systems. These systems may be based on or related to elements of ToBI system, but collapse across pairs of often-competing labels, such as L+H*/H* [17, 6, Ostendorf and Shattuck-Hufnagel, p.c.]. This practice reflects the widely-shared intuition that some labels are "closer" to each other than others, and that confusing close labels is less egregious than confusing other less close labels [10].

Clustering prosodic elements into such super-classes may be sufficient for the task at hand, but does not further the original goals of ToBI. Annotations based on local simplification schemes cannot be used by the larger speech community. Because they collapse across phonologically significant contrasts which are often (though not always) clearly distinct, and lose information about the nature and frequency of ambiguities, they obscure data that can help to evaluate and improve prosodic theories. In sum, coarse-grained transcriptions may facilitate faster labelling and improve some measures of reliability, but the loss of detail eliminates important research opportunities, e.g. determining where the L+H*/H* distinction is reliably perceived vs. not. The alt tier is designed to address these lost opportunities by providing a well-defined mechanism to capture explicit information about prosodically ambiguous regions.

3. The Alt Tier

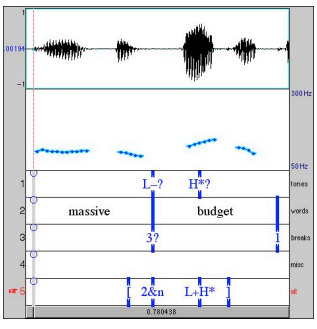
Discussion at the ToBI workshop held at Simmons College in 2004 acknowledged that much of the ambiguity in ToBI occurs when several alternative labels seem plausible, suggesting that prosodic theory has not yet accounted for all possible prosodic contours. Recording significant detail about this type of ambiguity is a critical step for further research. Workshop participants noted that many problem examples came from speech tokens for which two experienced labellers disagreed. Researchers suspected that such inconsistencies do not arise in a uniform distribution over all possible labels; instead, a small number of ambiguities tend to arise again and again. It was noted that these relatively few contexts occupied a disproportionately large amount of time in labelling, and contributed notably to user frustration. As an alternative to collapsing across confusable phonological categories, an additional labelling tier was suggested, to capture the alternatives under consideration. This approach has several practical and theoretical advantages. It allows labellers to specify what they perceive as the most likely category, but also record what they see as a competing hypothesis. They can do so in a way that is conventionalized, allowing them to move on past the troublesome point. Significantly, these alternative labels are not limited to single pairs of tonal targets: using the alt tier, a labeller could indicate that an entire sequence of tones and breaks has an alternative annotation.

3.1. Alt tier Mechanics and Notation

The alt tier uses a standard notation based on established ToBI labelling conventions, introducing minimal changes to the system. Use of the alt tier consists of labels in two locations: in the main label tiers (tones and breaks) and in the alt tier. Building on the established question mark (?) diacritic denoting uncertainty with regard to pitch accent type (X*?) or presence (*?), the alt tier conventions prescribe using a question mark after labels in the tones or breaks tiers (eg. H*? or 3?) for which an alternative (e.g. L+H* or 1) is listed in the alt tier at a point that is time-aligned to the main tier label. The question mark indicates labeller uncertainty and also signals that the alt tier has an entry. When considering alternatives, the labeller lists a first choice in the main label tier, and the second choice in the alt tier. One can list more than one alternative in the alt tier, though this is dispreferred.

Conventions for use of the alt tier have been established (available at www.tobihome.org) for 3 main types of labelling uncertainty: a) single-label uncertainty, affecting a label on only one tier, b) uncertainty affecting labels on both the breaks and tones tiers for a single point, and c) uncertainty for a region or sequence of labels. The conventions also dictate that labellers using the *? notation specify in the alt tier the pitch accent type they would choose if they were certain about the presence of the pitch accent.

In Figure 1, a labeller has indicated a region for which she considered two distinct sequences of labels. The word *massive* ends in a low f0, followed by a sharply rising f0 in the first syllable of the word *budget*. In the tones and breaks tiers, the labeller has indicated that the low f0 could be interpreted as a Low phrase accent and a 3 break, and that the following rise could be seen as a High pitch accent (H*) on the word *budget*. In the alt tier, the labeller has indicated an alternative parse, whereby the low-high sequence is annotated as the bitonal L+H* pitch accent, and the break between the words *massive* and *budget* is annotated as less strong than an intermediate phrase. The square brackets [ ] in the alt tier indicate that any labels between those brackets should be considered part of a related sequence of labels; an alternative sequence to all labels in the tones and breaks tiers in the region delineated. Figure 2 shows the use of an alt tier label indicating uncertainty about a single point; namely, about which phrase accent label to use before the boundary tone on the whispered words *thank you*.



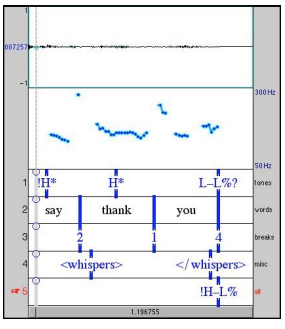


Figure 1 (left) *shows a sequence of alternative labels and* Figure 2 (right) *a single point alternative label.*

4. Data

This paper examines use of the alt tier in a study comparing inter-labeller consistency for two annotation systems, ToBI and RaP [4].The data comprise label files for approximately 6 minutes of speech, produced by 7 speakers, in roughly equal parts read and spontaneous speech. All 6 minutes of speech were labelled by four expert ToBI labellers.

5. Results

In examining the use of the alt tier by these four labellers, several questions were posed:

1) How was the alt tier used?
2) What useful information is captured by using the alt tier that is not captured by standard ToBI notation?
3) Does the alt tier facilitate labelling? (e.g., speed, ease and reliability)

5.1. How was the alt tier used?

Labellers used the alt tier 388 times in a corpus that contained 6 minutes of speech containing 1073 words and 1559 syllables for each of four annotators. This included 47 labels that were listed as part of 23 sequences or 12% of alt tier uses for all four labellers. There were 125 places (or 32%) where more than one labeller used the alt tier. The 4 labellers used the alt tier from 42 to 123 times (average use: 97 times). The alt tier was used 256 times with regard to pitch accent, 61 times with respect to phrase accent or boundary tone, and 71 times with respect to break size.

One concern about adopting an alt tier was that labellers would use it frequently and would list a large set of alternative labels in difficult contexts. The results show that explosive use of the alt tier does not occur: it was used only 388 times in a corpus with 7144 main tier tone and break labels, and generally contained only one alternative.

5.2. Is useful information captured using the alt tier?

The alt tier documents the regions where labellers are able to envision more than one prosodic parse and are dissatisfied with assigning a single parse. Rather than forcing an arbitrary choice between alternatives, creating 'pockets of unreliability', the time-aligned labels in the alt tier document where (and in what context) labellers find more than one reasonable alternative, and this is a fertile area for research in prosodic phonology. Tables 2 and 3 show the confusion matrices (across all 4 labellers) in this study. The main tone tier label is listed vertically, with its corresponding alt tier label when the alt tier was used. By examining where the alt label was used, one can discover what syntactic, prosodic and other acoustic contexts might play a role in this uncertainty. The data in Table 2 support earlier claims that certain labels are commonly confusable, such as L+H* and H*. However, the data also indicate that labellers CAN distinguish between these two labels: in 811 other H* labels (as shown in Table 1), the alt tier was not used, indicating certainty about this label.

Table 3, on the other hand, shows data that run counter to an assumption made in earlier ToBI labelling reliability studies (and also embedded in many condensed labelling schemes), i.e. that break index disagreements that are 'off by one' are less egregious because they involve confusable categories [10]. However, the current data suggest that confusable categories do not always involve contiguous break indices. Specifically, labellers indicated 1 or 3 as the alternative to the 2 break index (or v.v.) only 11 times, but indicated 3 as an alternative to a 1 (or v.v.) 27 times. This suggests that labellers often perceived a mis-match between 3-cues and 1-cues, and when given the means of the alt tier to express this ambiguity, they did so. On the other hand, labellers also used the 2 break index without recourse to the alt tier, suggesting that in some cases they may have perceived a boundary intermediate between 1 and 3 [14].

Table 1: *Number of prosodic pitch accent labels used and used with the alt tier.*

Tone	H*	L+H*	!H*	H+!H*	L+!H*	L*	*?	X*?
Tone tier labels	848	393	306	132	54	29	141	6
Alt labels	41	33	14	15	2	5	140	6

Table 2: *Confusion matrix showing the distribution and the number of specific alternatives listed in the alt tier. In some cases, no Pitch Accent was given as an alternative.*

Alt Tone / Tones tier	H*	L+H*	!H*	H+!H*	L*	L*+H	L+!H*
H*	x	21	3	0	1	0	0
L+H*	31	x	1	0	0	0	1
!H*	4	2	X	2	1	0	3
H+!H*	3	0	10	x	1	0	0
L*	0	0	4	0	x	0	0
L*+H	0	0	0	0	0	x	0
L+!H*	1	0	1	0	0	0	x
*?	77	2	46	7	6	0	0

Table 3: *Confusion matrix: the distribution of break indices with alternative labels, and their alternative labels*

Alt / Breaks	0	1	2	3	4
0	X	0	0	1	0
1	0	x	1	9	2
2	0	1	x	6	0
3	0	19	3	x	10
4	0	0	0	9	x

5.3. Does the alt tier facilitate labelling?

Overall, labellers in this study reported increased satisfaction when using the alt tier, and most self-reported faster and

easier labelling. Although direct comparison is not possible given different labellers and corpora, recorded labeller rates were significantly faster than labelling speeds reported elsewhere: the labelling rate of approximately 50 times real time, on average, for the present experiment is much less than the 100-200 times real time reported in [15]. Design considerations prohibit attributing this rate increase solely to the alt tier, but the data are suggestive.

The extent to which the alt tier can contribute to measures of reliability has not been fully explored in this study. One approach to quantifying the boost to agreement that the alt tier offers is, in a pair-wise reliability calculation, to consider a pair to be in agreement if either the primary label or the alt tier label agree. In this dataset, this method provides a 20% increase in agreement on pitch accent type for those syllables for which at least one labeller indicated an alternative label, leading to an estimated 3% improvement on agreement on pitch accent type for the entire dataset, if disagreement is distributed evenly throughout the corpus. Even leaving reliability calculations aside, however, use of the alt tier provides valuable information towards the evaluation of reliablilty: the context and nature of disagreement, and the difference between disagreement and near agreement.

6. Discussion

ToBI annotation is a comparatively new field, and there are a number of open research questions, both about the set of contrastive categories in the language and about the constraints on how they can be realized. Several research areas were identified at the 2004 ToBI workshop: the placement and shape of F0 peaks in High pitch accents, the apparent mis-match between tonal and durational cues at some phrase junctions and the role of relative strength in pitch accents. These areas relate directly to the reported confusion of L+H* and H*, the use of the 2 break index and the difficulty in distinguishing a small accent from a lexically main-stressed syllable in some contexts. These and other fertile research areas lie directly in those regions where labellers hesitate between two alternative labels, resulting in time-consuming and ultimately inconsistent primary annotations. Using the alt tier captures the alternatives that the labeller is considering,, allowing the labeller to stop agonizing (as Wightman put it) and move on. More importantly, it documents the type and location of ambiguities, allowing researchers to investigate the causes and contexts of uncertainty rather than regarding it simply as noisy data.

7. Conclusions

The alt tier addresses three of the main issues that critics as well as users of the ToBI annotation system have noted: pockets of unreliability, incomplete capture of information, and labeller dissatisfaction. We believe that this extension of the ToBI system will allow researchers to identify frequently-confusable patterns and to survey the contexts where these confusions occur, and that using the alt tier will provide greater labeller satisfaction, leading to wider adoption.

8. Acknowledgments

We gratefully acknowledge the contributions of Laura Dilley and Ted Gibson, co-designers (with the third author) of the experiment that provided these data; the participation in alt tier convention development by Nakul Vyas, Caroline Rubin, Tess Diduch, Marti Bolivar, Sun-Ah Jun, Mary Beckman, Adam Albright, Jennifer Cole, Jennifer Venditti and other participants in the August 2004 ToBI Workshop at Simmons College, Boston MA; and additional contributions by John Kraemer, Marti Bolivar, and Meredith Brown.

9. References

[1] Beckman, M. E. & Pierrehumbert, J. (1986), Intonational structure in Japanese and English. *Phonology Yearbook*, 3: 255-309.

[2] Beckman, M. & Ayers-Elam, G. (1997), *Guidelines for ToBI labelling, Version 3*. Ohio State University

[3] Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005), The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) , op.cit.

[4] Breen, M., Dilley, L., Kraemer, J., Bolivar, M., & Gibson, E. (in preparation), Inter-transcriber agreement for two systems of prosodic annotation: ToBI and RaP.

[5] Jun, S. (2005), Editor. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford Univ. Press.

[6] Kompe, R. (1997), *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer-Verlag, Berlin. (Verbmobil)

[7] Pierrehumbert, J. (1980), *The phonology and phonetics of English intonation*. PhD Thesis, MIT Departement of Linguistics

[8] Pierrehumbert, J. & Beckman, M. (1986), *Japanese Tone Structure.* Cambridge, MS: MIT Press

[9] Pitrelli, John F. (2004), ToBI Prosodic Analysis of a Professional Speaker of American English. *Proceedings of Speech Prosody 2004*, Nara, Japan.

[10] Pitrelli, J. F., Beckman, M. E. & Hirschberg, J. (1994), Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. In *Proceedings of ICSLP, Yokohama, Japan*: 123-126

[11] Price, P., Ostendorf, M., Shattuck-Hufnagel S. & Fong, C. (1991), The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America 90:* 2956-2970

[12] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M.; Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J., (1992), TOBI: A Standard for Labelling English Prosody. In *Proceedings of ICSLP, Banff, Alberta, Canada*: 867-870.

[13] Shattuck-Hufnagel, S., Dilley,L., Veilleux, N., Brugos, A. & Speer, R. (2004), F0 peaks and valleys aligned with non-prominent syllables can influence perceived prominence in adjacent syllables. In *Proceedings of Speech Prosody 2004*: 705-708.

[14] Shilman, M. (2007), Levels fo the Prosodic Hierarchy in English. *Proceedings of ICPhS XVI*: 973-996.

[15] Syrdal, A. K., Hirschberg, J., McGory, J. & Beckman, M. (2001), Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody. *Speech Communication*, 33(1-2): 135-151.

[16] Wightman, C. W. (2002), ToBI or Not ToBI. In *Proceedings of Speech Prosody*, Aix-en-Provence, France.

[17] Yoon, T., Chavarría, S., Cole, J. & Hasegawa-Johnson, M. (2004), Intertranscriber Reliability of Prosodic Labeling on Telephone Conversation using ToBI. *ICSA, ICSLP, INTERSPEECH 2004*: 2722-2732.