IViE - A Comparative Transcription system for Intonational Variation in English

Grabe, E., Nolan, F., and Farrar, K.J.

Department of Linguistics, University of Cambridge, Sidgwick Avenue
Cambridge CB3 9DA

ABSTRACT

In this paper, we offer an alternative to ToBI, the current *de facto* standard for machine-readable labelling of English prosody. We have three reasons for arguing that an alternative is needed. Firstly, the ToBI tone inventory is not maximally constrained. Inter-transcriber agreement for tone labels is relatively low and F0 contours generated from ToBI labels seem to be no closer to the original than those derived from a model which does not distinguish between accent shapes. Secondly, the growing demand for prosodically labelled data from non-standard varieties of English suggests a need for a transparent comparative transcription system. ToBI was not designed for this purpose. Thirdly, the low inter-transcriber agreement scores for ToBI suggest that the system is not as easy to apply as it may at first appear. In the present paper, we describe an alternative: the IViE system (Intonational Variation in English). We describe the structure of IViE and discuss its application with examples.

1. INTRODUCTION

Prosody plays a number of roles in speech, some of which are crucial to speech comprehension. In English, the location of intonation phrase (IP) boundaries is relevant to syntactic processing, the location of pitch accents determines the focus structure of an utterance, and the shape of a pitch accent can signal the difference between a question and a statement [4]. As a result, speech synthesis and recognition systems can benefit from information about prosody.

The current *de facto* standard for the prosodic labelling of large speech corpora is ToBI, the first widely used transcription system of its kind [1,16]. ToBI consists principally of a small set of discrete tone labels which identify intonational categories observed in standard varieties of English, and a set of 'break indices' which mark different degrees of rhythmic discontinuity between words in utterances. ToBI has been relatively successful, especially with respect to the application of break indices, but the tone labels appear to be somewhat problematic. Most importantly, the tone inventory is not maximally constrained; there seem to be more ToBI accents than actual phonological contrasts in the standard varieties of English. Evidence for this observation comes from studies using ToBI labels to predict intonation contours. These studies have been forced to collapse the ToBI tone inventory into a smaller number of classes in order to achieve their results [2,15]. Similarly, in a ToBI evaluation exercise, tone categories were collapsed in more than one case [14]. The need to combine categories suggests a problem with the mapping between phonological categories and their phonetic implementation. Some categories transcribe potentially contrastive patterns, but others transcribe patterns which are somewhat different but unlikely to be contrastive [6]. This impression is supported when one examines the inter-transcriber agreement scores for ToBI tone labels given in [14]. After collapsing a number of labels into new categories, the investigators found that 64.1% of their data points exhibited agreement as to which pitch accent was present. This agreement score is rather low, especially since it was reached only after collapsing categories.

Evidence for shortcomings in the predictive power of ToBI tone labels comes from a study by [5]. When generating F0 contours for data from American radio-speech, the authors found that contours generated from ToBI labels fared little better in predicting the shape of real F0 contours than those derived from an approach to labelling which distinguishes solely between the presence or absence of an accent (TILT). Still, it is possible that [5]'s results were an artefact of the corpus of data [5] used to evaluate ToBI and their one-accent model. If that corpus contained an overwhelming number of pitch accents of the same type (e.g. H*), then [5]'s findings are unsurprising. However, an overwhelming absence of any accent other than H* seems unlikely, especially in radio speech. Alternatively, it is possible that a number of ToBI labels were inappropriately assigned. Considering the size of the ToBI inventory, this explanation appears to be more likely.

A different set of questions about ToBI labels emerges from a cross-varietal study of British English intonation [17]. Dialectal variation in the segmental phonetic structure of English has been extensively described, but dialectal variation in English intonation has largely remained unaccounted for. In response, a cross-dialectal corpus of speech data is being collected at the University of Cambridge. The complete corpus will contain directly comparable speech data from seven English dialects, and the recordings and their prosodic transcriptions will be made available on CD-ROM. In this context, the need for a comparative prosodic transcription system arises. The ToBI system is not obviously suitable, because it was designed to account for the so-called standard varieties of American, British, and Australian English, and cannot be transferred to other varieties of English [11].

Finally, there is the question of learnability. Labellers have been shown to acquire the basics of ToBI relatively quickly. The low inter-transcriber agreement score for tones, however, shows that some aspects of the system are easier to learn than others (i.e. labelling break indices is quite straightforward). Tone labelling is difficult, because some tones appear to represent phonologically contrastive categories, but others indicate phonetic differences between accents belonging to one and the same phonological category. As a result, it is difficult to distinguish reliably between contrasts relevant to speech understanding and the different phonetic shapes which those contrasts map onto in different contexts. A more clearly constrained accent inventory combined with a means of transcribing observed phonetic variation within tonal categories may elicit a higher inter-transcriber agreement score.

Finally, note that transcriptions based on the British Tradition of intonation analysis [3] appear to elicit a somewhat higher inter-transcriber agreement score than ToBI. In their transcription of the Spoken English Corpus, Williams and Knowles reached a score of 68% with an inventory of as many as eleven pitch accents [12]. The agreement score for ToBI tones among four labellers was 64 % [16].

In response to the observations and queries discussed, we have developed an alternative to ToBI: the IViE system (IViE stands for Intonational Variation in English). An outline of the system is given in the following section.

2. THE STRUCTURE OF IVIE

IViE is very obviously different from ToBI in two respects:
(1) IViE has a different tone inventory from ToBI. ToBI is based on [13] and IViE is based on [6,8, and see [9] for an experimental evaluation of [8]]. [6] and [8], are based on the British Tradition of intonation analysis, i.e. systems similar to that used by Knowles and Williams in their transcription of the Spoken English Corpus.
(2) In IViE, the intonational representation is 'unpacked' into three tiers. One tier contains labels for rhythm, another labels for pitch movements surrounding rhythmically prominent syllables, and the third contains labels providing a phonological categorisation of these pitch movements. In ToBI, the intonational representation is unilinear. Unpacking the ToBI tone level into three levels of transcription is advantageous because labellers arrive at a phonological classification step-by-step. In the course of this, they learn to distinguish between phonological categories and differences in the phonetic implementation of these categories.

Note, however, that there are also a number of similarities between ToBI and IViE. These similarities are not surprising, because IViE is based on ToBI. First of all, both systems are based on autosegmental-metrical approaches to intonation analysis [10], and describe intonation with a limited set of categories made up from high and low tones. Both label rhythm, albeit in slightly different ways, and both can be used in conjunction with waves+ in exactly the same way. Transcribers use a time-aligned speech wave, labelling template and FO trace to arrive at their transcriptions. And just like ToBI, IViE can also be used without waves+.

IViE has five tiers arranged as follows:

Comment tier
Phonological tier
Auditory phonetic tier } Prosody
Rhythmic tier
Orthographic tier

On the orthographic tier, labellers transcribe the words spoken. On the rhythmic tier, the location of rhythmically prominent syllables, and the location of rhythmic boundaries is indicated. The label '<' indicates the left edge of the stressed syllable, and '>' the right edge. % indicates the location of an IP boundary.

On the auditory phonetic tier, pitch movements surrounding rhythmically prominent syllables are transcribed. The transcription involves three landmarks: the pitch level on the prominent syllable, the pitch level on the immediately preceding syllable and that on the immediately following syllable. These pitch levels are transcribed relative to each other (i.e. not relative to some absolute value). The transcription is given separately for each prominent (accented) syllable, but auditory transcriptions do not cross IP boundaries. The latter point is relevant especially in spontaneous speech, where IPs are often very short (e.g. a single word). Relationships between accents are not taken into account; such relationships are transcribed at the phonological level.

Three pitch levels are available: high, mid and low. The pitch level on the rhythmically prominent syllable is transcribed with a capital letter (i.e. H, M or L), and the ones on the immediately preceding and following syllables with a small letter (h, m or l). Capitalising the letter on the prominent syllable makes sure that readers of the transcription can keep track of the location of the stressed syllable. l_H_l, for instance, transcribes high pitch on the accented syllable, and low pitch on the syllables before and after. m_H_l transcribes high pitch on the accented syllable, mid pitch on the syllable before it and relatively lower pitch on the one after it.

The tone inventory on the phonological tier is based on the phonological accounts of Southern Standard British English in [6,8]. These accounts combine aspects of the British school of intonation analysis (e.g. all accents are left-headed) and current autosegmental models of intonation (e.g. the use of primitives H and L). Note, however, that unlike the accounts of English in [6] and [8], the IViE tone inventory is not intended to represent any particular intonational system. Rather, the options represent a pool of labels for comparative analysis, and we assume that different varieties of English are characterised by different subsets of labels. By drawing all label subsets from the same pool, we achieve as much comparability as possible. The following labels are available on the phonological tier:

Tone labels

H*	high pitch target
H*L	high target followed by low target
H*LH	fall rise
L*	low target
L*H	low target followed by high target
L*HL	rise-fall

Tone modifiers

^ upstep of a tone
! downstep of a tone
_ precedes tone and indicates displacement of a tone to the right, e.g. H*_L
+ connects tones functioning as phonological unit in a particular variety, e.g. H*+L

Boundary specifications

Phrase-initial	Phrase-final
%H	H%
%0	0%
%L	L%

The tone modifiers '^' and "!" modify the location of tones in the speaker's register. '_' and '+' modify the location of tones in the time domain, and allow us to capture similarities as well differences between related patterns. E.g. the pattern H*+_L is very similar to a pattern H*+L, apart from the L having been shifted away from the high target to the right (H*+_L is described as a falling head in the British Tradition).'+' indicates that tones function as a phonological unit, e.g. H*+L can transcribe a falling nuclear accent, a commonly observed pattern in simple statements in British English.

With respect to H and L boundary tones, IViE does not differ from ToBI, but the 0 (zero) boundary tones are new and have been adopted from [6]. A zero boundary tone means that the pitch on the last syllable in the IP does not differ from the immediately preceding tone. In practise, a 0% means: here is a relevant landmark in the contour; an IP boundary. At this boundary, there is no pitch change, i.e. the pitch level of the last tone preceding the boundary is simply continued. In ToBI, the transcription of such patterns is less transparent. Finally, note that IViE has one level of intonational phrasing: the intonation phrase, but ToBI has two.

Labelling proceeds as follows: labellers begin with the orthographic transcription. Then they determine the location of stressed syllables and rhythmic boundaries. After that, the pitch movement surrounding stressed syllables is transcribed. Finally, i.e. when information about rhythmic structure and auditory phonetic structure has been collected, phonological generalisations are made.

For work on [17], a subset of labels for General Southern British English is has been established, and the development of subsets for other varieties of English (Newcastle and Belfast) is in progress.

3. EXAMPLES ILLUSTRATING SOME ADVANTAGES OF IVIE

3.1. Tone inventory

The 0% option allows for directly comparable transcriptions of cross-varietal differences in boundary specifications. General Southern British English has two boundary options: rising (H%) or level (0%). Belfast English has three: rising (H%), level (0%) and falling (L%). IViE can capture this difference transparently.

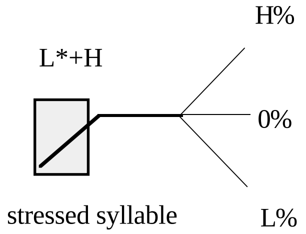

Figure 1: IViE transcription of boundary options in Belfast English.

In ToBI, a rise-plateau L*+H 0% needs to be transcribed as L*+H L-H%. This transcription is less transparent, because the L and the H in L+H represent pitch targets, but the L- and the H% do not - they represent level pitch.

3.2. Auditory phonetic tier

The auditory phonetic tier explicitly incorporates the assumption that there is a one-to-many mapping between phonological generalisations in intonation and their phonetic realisation. This assumption is not explicit in ToBI, rather, ToBI combines phonological and phonetic aspects of intonation. For instance, ToBI transcribes an accent L+H*L- and an accent H*L-. This distinction was collapsed in the ToBI evaluation exercise. In IViE, the difference is captured as one involving phonetic realisation and transcribed on the auditory phonetic tier. In the phonology, both accents are H*+L. Conversely, phonetic effects such as truncation can be retrieved from the auditory transcription [7]. In ToBI, such effects cannot be transcribed in any obvious way.

Finally, information about the auditory implementation of a pattern is of interest to researchers working on speech synthesis or recognition. We can assume that listeners cannot uncover the correct phonological choice and come to an appropriate conclusion as to what the intended meaning of a pattern is without being aware of the possible auditory implementations of that pattern. The auditory phonetic level offers such information.

4. SUMMARY AND CONCLUSION

In the present paper, we have introduced the IViE system for comparative prosodic labelling. IViE was developed primarily as a tool for linguistic research in intonation. Such research is increasingly based on the analysis of large speech corpora. The labelling of such corpora is relatively time-consuming, and labelling is not always carried out by experts. Therefore, it is important that transcriptions can be replicated, and that the motivation for a specific transcription is retrievable. In the IViE system, information about the prosody of an utterance involves tonal generalisations which are flexible enough to account for several varieties in a comparable system; information about the phonetic realisation of those generalisations, and information about the rhythmic structure of utterances. Thus, information about prosody is arrived at step-by-step. As a result, the transcription can be learned relatively quickly, and can be replicated by other researchers. Also, the information contained in an IViE transcription can be used selectively, if required. For instance, IViE tone labels can be combined with ToBI break indices, if required.

In sum, a widely applicable transcription system for English prosody must be easy to learn, transparent, and comparable across different varieties of English. We suggest that the composite nature of IViE can meet these demands.

5. REFERENCES

1. Beckman, M. and Ayers, G. “Guidelines for ToBI labeling, version 2.0”, Linguistics Department, Ohio State University, 1994.

2. Black, A. and A. Hunt, A. “Generating F0 contours from ToBI labels using linear regression”, In *Proc. ICSLP*, Philadelphia, Penn. 1996, 1385-1388.

3. Cruttenden, A. *Intonation*. Cambridge: CUP. 1997.

4. Cutler, A., Dahan, D. and Donselaar, W. van “Prosody in the Comprehension of Spoken Language: A Literature Review”, Language and Speech, 1997, 141-210.

5. Dusterhoff, K. and Black, A. “Generating F0 contours for Speech sysnthesis using the TILT intonation model”, In *Proc. of the ESCA Tutorial and Research workshop on Intonation,* Athens, Greece, 1997. 107-110.

6. Grabe, E. Comparative Intonational Phonology: English and German. Doctoral Dissertation, MPI Series 7, Nijmegen, The Netherlands, 1998a.

7. Grabe, E. “Pitch accent realisation in English and German”, *Journal of Phonetics* 26. 1998.

8. Gussenhoven, C. *On the grammar and semantics of sentence accents,* Dordrecht: Foris, 1984.

9. Gussenhoven, C. and Rietveld, T. “An experimental evaluation of two nuclear-tone taxonomies”, *Linguistics*, 29: 423-49. 1991.

10. Ladd, D. R. *Intonational phonology,* Cambridge: CUP, 1996.

11. Nolan, F. and Grabe, E. “Can ToBI transcribe intonational variation in the British Isles?”, *Proc. of the ESCA Tutorial and Research Workshop on Intonation,* Athens, Greece. 1997. 259-262.

12. Pickering, B., Williams, B., and Knowles, G. Analysis of transcriber differences in the SEC. In G. Knowles, A. Wichmann and P. Alderson (eds). *Working with Speech*, London, Longman. 1996

13. Pierrehumbert, J. B. The phonology and phonetics of English intonation, Unpublished doctoral dissertation, Cambridge, MA: MIT. 1980.

14. Pitrelli, J. Beckman, M.E. and Hirschberg, J. “Evaluation of prosodic transcription labeling. Reliability in the ToBI framework”, *Proc. ICSLP*, Yokohama, Japan, 1994.

15. Ross, K. and M. Ostendorf, M. “A dynamical system model for generating F0 for synthesis”,In *Proc. ESCA workshop on speech synthesis 131-134,* Mohonk, NY, 1994.

16. Silverman, K., Beckman, M. E., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. “ToBI: a standard for labeling English prosody. In *Proc. ICSLP* 2: 867 - 70. Banff, Canada, 1992.

17. ESRC grant R000237145, ‘Intonational Variation in the British Isles’. Award holders F. Nolan and E. Grabe; research associate K.J. Farrar, University of Cambridge, 1997-2000.