1. Overview and some basics

Spoken language conveys not only words, but a wide range of other information about timing, intonation, prominence, phrasing, voice quality, rhythm etc. that is often collectively called spoken prosody. These aspects of an utterance are sometimes called supra-segmental, because they can span regions larger than a single phonemic segment (consonant or vowel). The ToBI (Tones and Break Indices) transcription system is a method for transcribing two distinctive aspects of the prosody of spoken utterances:

a) accents, which contribute to a word's relative prominence in an utterance and
b) phrasing, which creates a grouping of words.

These prosodic aspects of spoken language can convey distinctive semantic, syntactic or even morphological facts, and can be useful for tasks such as mapping prominence and grouping patterns to meaning differences, understanding the effects of prominence and grouping on the pronunciation of words, and synthesizing prosodically natural-sounding speech. Speech scientists are interested in annotating the prosodic structure of large numbers of utterances in order to study these phenomena, and ToBI was developed to provide a common transcription system that could be used by many different laboratories, making it easier to share data.

An example of how prosodic prominence and grouping can convey differences in structure and meaning can be seen in two different pronunciations of the word string *It broke out in Washington* (Price et al. 1992, Lehiste 1987). On the one hand, this string can be produced with a prominence on *broke* and a phrase boundary just after that word, corresponding to the orthographic representation *It BROKE, out in WASHington*. <broke1> On the other hand, it can be produced with a prominence on *out* and a phrase boundary just after that word, corresponding to *It broke OUT, in WASHington*. <broke2>.

In this case, the two different patterns that correspond to two different syntactic structures and two different meanings can be easily captured by orthographic conventions and punctuation, i.e. upper case typeface and commas. But other differences, like the distinction between prominence signaled by a high pitch vs. one signaled by a low pitch, or between a prominence signaled with an early pitch peak vs. a later one, cannot be easily captured by conventional orthography. The ToBI system, like other transcription systems, was developed to capture prosodic differences of all types, not just the ones that are easy to write down in English orthography. As you will see, the ToBI framework assumes that the differences that matter are phonological differences, i.e. that the job of the transcription system is to capture the differences between prosodic categories that correspond to differences in function and in meaning, but it is not yet entirely clear how this mapping between prosodic categories and their meanings and functions works. One advantage of developing an explicit system for transcribing the categories is that, when it is applied to samples of natural speech, it provides both a stringent test of the theory behind the transcription system, and a laboratory tool for determining how the theory needs to be extended and revised.

A challenging aspect of transcribing prosody is that there is a substantial level of variability. For example, one important cue to prominence and phrasing is the intonation of an utterance, i.e. changes in the pitch that are caused by changes in the frequency of vibration of the vocal folds, often called f0. An f0 that is high in a speaker's range on a salient syllable can mark a pitch

accent, but so can a lower (but still high-for-this-speaker) f0 on another salient syllable. Moreover, a high f0 might mark a pitch accented word or it could mark a phrase boundary tone, as in the pitch rise often heard on the final syllable of certain kinds of questions in English, like *Is it raining?* In order to determine the appropriate ToBI transcription, the entire utterance must be parsed, that is, understood as a whole, to determine how the high and low tones are implemented. This tutorial will begin by introducing utterances with relatively unambiguous ToBI annotations and will gradually move to utterances for which even experienced labellers might disagree about how to transcribe some regions. This will be less disturbing if one recognizes that ambiguous cues are also present in other types of speech annotation. For example, the acoustic cues for a canonical /t/ might not be present, e.g. in some renditions of the word *butter*. However, the existence of a well-established lexicon allows annotators to recognize tokens produced with a flap (sounding more like a /d/) and tokens produced more carefully as variations of a single lexical item, and to label them consistently. At this writing, the nature of variation for prosodic categories (and thus for ToBI labels) is not fully understood, and this results in some reasonable disagreement in prosodic parses in some renditions of some utterances, particularly in spontaneous speech. ToBI labelling is a new endeavor compared to, say, phonetic transcription, and as the nature and range of prosodic variation becomes better understood and documented, we expect the transcription ambiguity to lessen. In the meantime, the system provides a number of tools for recording your uncertainty, to help us understand better where the system can be improved. In fact, although the development of a ToBI system for a particular language requires a certain degree of prior understanding of how prominence and grouping work in that language, a ToBI system can also be viewed as a tool for exploring the prosodic system of a language in greater detail. In this tutorial we introduce you to the ToBI system that has been developed for Mainstream American English; for information about ToBI systems for other languages, and about the general ToBI framework within which these systems have been developed, see Jun (2005), *Prosodic Typology*.

1.1. The basic parts of ToBI[1]

A ToBI transcription of an utterance consists minimally of a recording of the speech, its fundamental frequency contour, and (in the transcription proper) symbolic labels for prosodic events. The transcription proper is usually arranged in four time-aligned parallel horizontal panels or tiers, so that the symbolic labels can be easily matched with the corresponding f0 track and speech waveform. (Other tiers can be added for the needs of particular sites.) The four labelling tiers each appear in their own window:

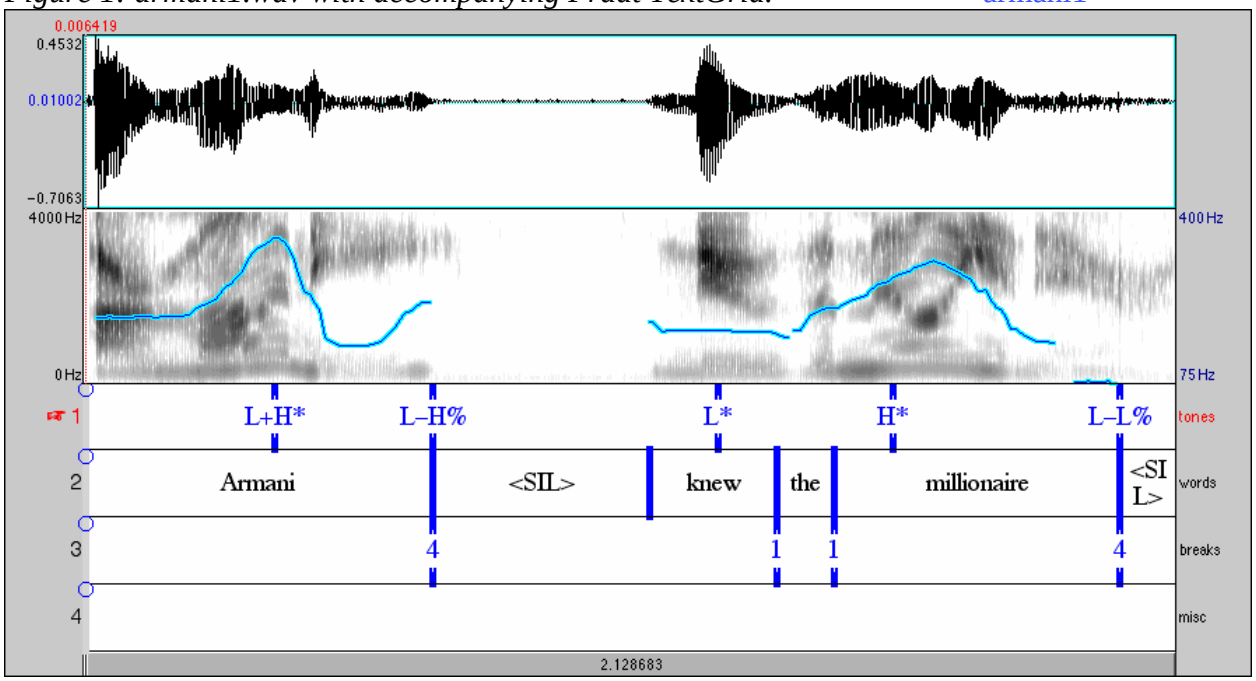
(1) the Tone tier, for transcribing tonal events
(2) the Orthographic tier, for transcribing words
(3) the Break-Index tier, for transcribing boundaries between words
(4) the Miscellaneous tier, for recording additional observations

In addition, two new tiers have been suggested for labellers who find them useful. These are not used in the examples in the beginning of this tutorial, but will be explained in Section 2.12:
(5) an Alternative Tier, for transcribing alternative labels in the case of ambiguity
(6) a Discussion tier, for recording data for selected research issues

[1] Parts of this section and section 1.2 have been adapted from the Guidelines for ToBI Labelling, version 3.0 (1997).

One popular program for labelling and displaying ToBI transcriptions is Praat (available at http://www.praat.org). Here is an example using Praat to display the utterance waveform, the f0 contour, the spectrogram and the first four tiers for the utterance *Armani knew the millionaire.*

Figure 1: armani1.wav with accompanying Praat TextGrid. <armani1>



The topmost window in this display shows the waveform of the recorded utterance; later, when you learn how to expand the horizontal time scale, you will see more clearly the vertical striations that indicate the individual pitch pulses made by the vocal folds as they vibrate. The horizontal axis corresponds to time and the vertical axis to the amplitude of the vibration. You can see in the waveform that the amplitude varies roughly with the syllable structure of the utterance: it is large for vowels (where the mouth is open relatively wide), smaller for consonants (which have a constriction in the vocal tract), and zero (or nearly zero) when there is no speech signal.

The rate of vibration of the vocal folds is what we hear as the pitch, and this is represented in the second window as a semi-continuous blue line superimposed on a different representation of the speech signal, the spectrogram (see “What is the spectrogram?” in the grey box below). When you play this utterance, notice where you hear a higher pitch. Do these words and syllables correspond to the places where the F0 contour (also called the pitch track or the f0 track, although f0 and perceived pitch are not exactly the same thing) is higher, as indicated by the blue lines?

What is the spectrogram?

The gray and white pattern in the background behind the f0 track is the spectrogram which is another way of displaying information about the speech signal that you see in the wave form above it. Like the wave form, it shows time on the horizontal axis, but the vertical axis here is frequency. Among other things, the spectrogram shows the frequencies that resonate especially well in the vocal tract as black horizontal bands that appear and disappear, and rise and fall, as the speech articulators move around inside the vocal tract, changing the size and shape of the resonant cavities inside the mouth, nose and throat. This is much like the change in sound that happens when a jug is filling up with water from a faucet: the pitch of the noise rises as the air cavity above the water level gets smaller). The spectrogram often shows the change from a consonant to a vowel especially clearly; notice the sharp boundaries between the /m/s and /n/s and their surrounding vowels in Figure 1.

Experienced ToBI labellers often use the spectrogram to help find the location of successive sounds, syllables and words. This in turn helps them to keep track of where changes in the pitch track occur across the words and syllables of the utterance. Learning how to use this information may take some time, so at the beginning you may want to rely more on listening to make these judgments.

Underneath the spectrogram with the blue pitch contour is the set of four thinner white boxes that make up the four tiers in the ToBI labelling window. Unlike the speech displays, these boxes are text writeable and this is where you will type in the ToBI transcription labels. The top white box is the Tone tier, and the third box is the Break Index tier. These two tiers represent the core ToBI analysis. The Tone tier is the part of the transcription that corresponds most closely to a phonological analysis of the utterance's intonation pattern. It consists of labels for distinctive pitch events, transcribed as a sequence of high (H) and low (L) tones marked with diacritics indicating their intonational function. Tones function either as prominence markers called pitch accents, as parts of pitch accents, or as boundary-related events called phrase accents and boundary tones, that mark the edges of two types of phrases. These categories are based on the work of Janet Pierrehumbert (1980) and joint work by Mary Beckman and Janet Pierrehumbert (1986, 1988).

The Break-Index tier captures the prosodic grouping of the words in an utterance by labelling the end of each word for the subjective strength of its association with the next word, on a scale from 0 (for weakest perceived boundary/strongest perceived conjoining, as in *doncha* for *don't you*) to 4 (for the most disjoint boundary, i.e. at the end of the highest-level intonationally marked phrase). These categories of association strength or "break indices" are based on work by Mari Ostendorf, Patti Price, Stefanie Shattuck-Hufnagel, and their associates (Price et al., 1991). The two highest break indices (3 and 4) are equated with two levels of prosodic groupings (phrases) that are marked intonationally; additional higher-level groupings of these Intonational Phrases are not marked in the ToBI system.

The Orthographic tier is the third white box. It contains a straightforward transcription of all of the words in the utterance, in ordinary English orthography. The word transcriptions are aligned with their locations in the speech waveform. For labellers using Praat or a similar labelling computer application, the convention is to place the orthographic label for a word between two marks that delineate the approximate time interval in the signal that corresponds to the utterance of that word and placing <SIL> to mark silence between words, if any.[2] The orthographic tier is arguably not part of any core prosodic analysis, except inasmuch as the labels on this tier can be used to interface the transcription to dictionary entries which do indicate such things as which syllable is likely to be most stressed in each word, prosodic information which is not otherwise included in the ToBI system (more on this below). This tier also helps the labeller to keep track of which time regions in the wave form, spectrogram and f0 track correspond to which words in the utterance.

The Miscellaneous tier is the bottom white box in this display. It is essentially a 'comment' tier that can be used to mark events such as breaths, coughs, laughter, long silences and other non-speech events. These are traditionally marked with angle brackets (e.g. <cough>). Like the orthographic tier, it can include events that are arguably not part of prosody per se. However, many events that are typically marked on the Miscellaneous tier are important for interpreting the analyses on the Tone tier and Break-Index tier, because they disrupt the smooth rhythm of the utterance or interrupt the intonation contour. Labels on this tier usually mark the beginning and end of an event interval; one exception is the label ‘disfl’, which often stands alone to flag the occurrence of a perceived disfluency of some type.

1.2. Guiding principles

As the preceding discussion shows, ToBI does not try to transcribe all aspects of prosody, or even all aspects that are amenable to symbolic transcription. In deciding what to include and what to leave out, we are guided by three principles. First, we want to be able to distinguish in our transcription all of the categorically distinct intonation patterns and prosodic units of the language (in this case, Mainstream American English (MAE), see Jun 2005 for ToBI systems for other languages and dialects). Second, we do not transcribe aspects of prosody which are more amenable to continuous-valued quantitative measures than to the categorical divisions of a symbolic transcription, such as the slope of the changing f0 curve. Finally, we do not want to squander the user's energies in transcribing even categorical aspects of prosody which are predictable from other parts of the transcription or from auxiliary tools, such as dictionaries, that can be used to determine the location of lexical stress within words.

The categorical aspects of prosody which we try to capture completely (according to the first principle) are of two types. The first is the prosodic structure -- the alternating rhythm of more and less prominent words and syllables and the grouping of words into prosodic constituents of various sizes -- and the second is the intonation pattern – the sequence of contrastive pitch events

[2] In older transcriptions using xwaves ™, the orthographic label had to be aligned to the end of the word rather than spanning the whole word interval. Other popular programs for displaying and labelling speech are wavesurfer (http://www.speech.kth.se/wavesurfer) and emu (http://emu.sourceforge.net/emu-tobi.shtml).

that we call pitch accents, phrase accents, and boundary tones, and that determines the f0 contour of the utterance. A basic assumption of the ToBI approach is that both of these goals can be met using the same inventory of elements.

[The next two paragraphs contain further discussion of what is not captured by a ToBI transcription; read them if you are curious about this question. Otherwise, skip directly to section 2.0 below.]

An example of the non-categorical aspects of prosody which we leave out (in accordance with the second principle) is the local tempo of each word in the utterance, which we feel could be more accurately and directly captured by some quantitative measure such as normalized segment duration (e.g., Campbell, 1992) than by any symbolic transcription such as an arbitrary division into, say, categories `1', `2', and `3' (for `slow', `medium', and `fast' tempi).

A categorical aspect of prosody which we leave out (in accordance with the third principle) because it should be fairly predictable is the marking of the lexically stressed and unstressed syllables within each word. By this level of stress we mean the word-internal alternation between more stressed and less stressed syllables, where the relative prominence of any pair of syllables is fairly fixed and can be thought of as inherent to the word's dictionary entry.