Intonation Modeling (based on a lecture by Lesley Carmichael May 30, 2001)

Why is intonation important?
For anyone who doubts the role of intonation in linguistics or its value as an essential communicative tool, I’d like to share the following two examples with you. The question “Would you rather a lion eat you or a bear?” is ambiguous in written form. Intonation features can help to determine scope and intention.

Would you rather a LION eat you or a BEAR?

<table>
<thead>
<tr>
<th>tones</th>
<th>H*</th>
<th>L+H</th>
<th>H-H%</th>
<th>H*</th>
<th>L-L%</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>would you rather a lion eat you or a bear</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breaks</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Would you rather a LION eat YOU or a BEAR?

<table>
<thead>
<tr>
<th>tones</th>
<th>H*</th>
<th>L-L+H*</th>
<th>H*</th>
<th>L-L%</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>would you rather a lion eat you or a bear</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breaks</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
What are the “objects” of intonation?

**Acoustic:** Fundamental frequency (f0) is generally regarded as the primary acoustic cue.

Duration and intensity are included as secondary cues in some models.

**Perceptual:** pitch (correlate of f0), duration, loudness (correlate of intensity)

What is the difference between “intonation” and “prosody”? These terms are often used interchangeably, probably because there is an ongoing dialogue about where the domain of one ends and the other begins.

**Intonation** typically refers to modulation of pitch; it also typically refers to global tonal distribution, a hierarchical (suprasegmental) level of information larger than the word.

**Prosody** typically refers to temporal (duration) and air pressure (intensity) parameters; it typically refers to inherently lexical tonal features.

Observing intonation

Returning to the original sentences I presented, observe the pitch contour with the above information in mind. The prominent information in these two utterances is clearly different just by looking at the fundamental frequency contour. Notice the dramatic “scoop”-shaped change in f0 which occurs on the syllable “lion” in the first utterance. In the second utterance, a dramatic pitch change occurs on the syllable “you”. These two “bitonal” accents (showing a scoop or rise) are often used in English to indicate contrastive stress.

Some people say that the height of f0 or the location of the peak is the indicator of a pitch accent. Recent research outcomes are showing support for the idea that the degree of f0 change over time and the location of the peak relative to the beginning of the accented syllable are perceptually interesting. Also, notice in the first utterance how the high pitch extends from “lion” until the end of the phrase after “eat you”. This is due to a phrase boundary tone, and it helps to segment the speech stream into meaningful clusters.

Intonation is a key to disambiguation in spoken language. Strings which are structurally ambiguous in text format are often disambiguated in speech. Even a structurally ambiguous sentence that does not employ contrastive stress typically uses different intonation patterns for its separate possible meanings.
Notice the pitch dropoff after “woman” in the second sentence. “Woman” is a prominent piece of information in both sentences, but in the first one, the descriptor “with the telescope” is prosodically grouped with woman and the pitch contour shows a more gradual declination over the phrase. The intonation differences are definitely more subtle in cases like these which contain syntactic ambiguity without contrastive information (cf “The man saw the WOMAN with the telescope” [not the NEIGHBOR] or “…the woman with the TELESCOPE” [not the binoculars]). A caveat should be introduced here that the range of intonational expression is large and shows tremendous variety. Spontaneous speech in particular is a beast. The great challenges in investigating intonation are finding regular patterns, determining the contribution of various perceptual cues, and accounting for our ability to access intention, emotion, and focus in sometimes extremely subtle signal features.
How do we model intonation?
According to Paul Taylor (forthcoming), an intonation model should be able to provide a linguistically meaningful representation of the intonation of an utterance. Such representations should be able to be automatically captured from acoustic information, i.e., phonetically. Also, the acoustics of the utterance should be able to be regenerated from the representations drawn from the use of the intonation model. Taylor specifies 5 main properties that a linguistic representation of intonation should have:

1. It should be compact such that one part of derivation should not be derivable from another (i.e., no redundancy).
2. It should cover as many perceptually distinct phenomena as possible.
3. It should be linguistically meaningful.
4. It should allow for automatic synthesis. The f0 can be regenerated from the representations given in the prosodic modeling (i.e., the phonetic information can be reproduced based on the phonological representation).
5. It should allow for automatic analysis. The phonological representation should be automatically derivable from the acoustic signal (phonetic information).

Phonetic vs. phonological intonation modeling
There are two main types of intonation models:
1. Tone sequence models, which make categorical distinctions among prosodic phenomena.
2. Acoustic phonetic models, which consider a hierarchical prosodic structure and identify prosodic phenomena by their raw acoustic features.

Because of the inherent variability and the relative nature of measurable prosodic correlates, the distinction between these two methods for assessing intonation can be critical. Phonological phenomena which are not strongly evidenced by phonetic information illustrate the gap between the two assessments. For example, pitch accents in English can emerge as level tones (have no significant f0 excursion) and thus would show up in a phonological transcription but not a phonetic one (and particularly not an automated transcription).

Confounding behavior in intonation
In addition to the lack of a tidy correspondence between phonological and phonetic descriptions of intonation, other issues complicate the process of intonation analysis. Speech is composed of both voiced and voiceless segments; therefore, f0 can only be evaluated on parts of an utterance (the voiced parts). Voiceless segments (typically about ¼ of connected English speech) have no vocal fold vibration, leaving information gaps throughout the signal. In addition, the voicing of obstruents tends to introduce perturbations into the f0, raising it around voiceless obstruent closures and depressing it around voiced obstruent closures. Vowel quality can also affect the f0: Low vowels have an inherently lower pitch than high vowels. While human listeners seem to have no problem accommodating these acoustic realities, an intonation model that will be used computationally must be able to account for them.
References


Purpose: Synthesis vs. Recognition
Taylor 1995 (RFC): Prototypical pattern algorithms simply describe a notional centroid in the distribution of the acoustic patterns associated with the pitch accent type, and give little or no explanation of how or why the acoustic patterns for that pitch accent vary. For speech synthesis purposes, this is often adequate, as a single acoustic pattern is often sufficient for each pitch accent type. Given the fact that only a single pattern is needed, it makes sense to use a prototypical pattern for each pitch accent type. For speech analysis purposes, this paradigm poses problems because the distribution of a class may become complex and it is not a simple matter to identify the pitch accent type of a section of f0 contour by matching observed patterns to prototypical patterns. The use of this paradigm is made even more difficult because the prototypical patterns given in the literature are often described in ad-hoc ways: no principled (i.e., experimental) reasons are given for why the patterns are as they are, and no account is given for what variations may be expected between speakers, etc.

For part 3: Phonological vs. Phonetic models
Phonetic models describe intonational phenomena observable in the f0 contour (Taylor, 2000). This contrasts with phonological models, which seek to identify the underlying intonational structure of the utterance. Taylor says the distinction is usually not important, except when phonological phenomena are not strongly evidenced by phonetic information, such as level tones (pitch accents with no f0 excursion—perhaps only evidenced by a duration difference—this would show up in a phonological transcription of data but not a phonetic one, or an automatic one, for that matter).

Autosegmental Metrical approach (Pierrehumbert, 1980, Liberman, 1975, Bruce, 1977) Intonation has a phonological level of representation just like the segmental level of speech does. They argue that abstract phonological representations are the best way to describe sound patterns because it eliminates the need to deal with acoustic values directly. Intonation can be compared across speakers when discrete categories of intonational events are used to describe the sound patterns. Taylor (2000?) says this is because the representations are abstract—but not necessarily because they are phonological. (Tilt values are abstract in a sense—but how to account for acoustic effect of subglottal pressure reduction toward end of utterance? ToBI calls an H* an H*; how does Tilt normalize for differences in same type of pitch accent? Or does it matter?)