3 A Multilevel, Multilingual Approach to the Annotation and Representation of Speech Prosody

Daniel Hirst

3.1 Introduction: Prosodic Annotation

Instinctively, when we listen to an utterance, we make a difference between what is being said, and how it is said, that is, the prosody of the utterance. There is, generally, a fairly large consensus among linguists (and nonlinguists) on the annotation of what is said. Interestingly, though, there is a remarkable lack of consensus on the annotation and representation of how it is said. This difference can partly be explained by the fact that most writing systems provide a fairly reliable annotation of what is said, whereas there is very little written indication of how an utterance is spoken, apart from a limited number of abstract prosodic characteristics that are indicated by punctuation.

The result is that there has been a bewildering variety of annotation systems for speech prosody. Until quite recently, practically every researcher has developed their own annotation system: it was quite rare for anyone to use someone else's system. Even for the British School, which had a fairly consistent approach to intonation patterns, there have been quite considerable differences among the annotation used by different authors such as Palmer (1924), Armstrong and Ward (1926), Jassem (1952), Kingdon (1958), O'Connor and Arnold (1961), Faure (1962), Halliday (1967), Crystal (1969), Couper-Kuhlen (1986), or most recently Wells (2006).

There have been at least two partial exceptions to this general tendency. One exception was the tones and break indices (ToBI) annotation system (Silverman et al. 1992), developed as a standard annotation system for American English and later applied to several different languages (Jun 2005). Another exception was my International Transcription System for Intonation (INTSINT), described in more detail in section 3.6.1. The original version of the system (Hirst 1987) was based on an inventory of minimal pitch contrasts found in published descriptions of intonation patterns of several different languages. The aim was to provide a tool for the systematic description of these intonation patterns, something on the lines of the International Phonetic Alphabet. The INTSINT system was later used for nine of the twenty-two intonation systems described in Hirst and Di Cristo (1998).

3.2 Levels of Representation

I have suggested (Hirst, Di Cristo, and Espesser 2000) that one of the reasons for this lack of consensus is that different annotation systems are in fact annotating different levels of representation. A similar point is made by Taylor (2009) who notes that different authors have completely different approaches, phonological, phonetic, acoustic:

> These different approaches are of course quite common in all areas of language study; what is of note is that intonation seems to be one of the few areas where engineering and scientific approaches still have enough commonality to constitute a single field. (244)

The prosody of an utterance can be described at several different levels. In particular, I have argued that it is important to make a distinction between prosodic forms and prosodic functions. *Prosodic forms* are how the prosody affects what the utterance sounds like. *Prosodic functions* are what the prosody does, and how it contributes to the interpretation of an utterance. I mentioned that the only written correspondence to speech prosody is punctuation, which encodes a limited number of abstract prosodic characteristics. The difference between a full stop, a comma, and ellipsis, a question mark, and an exclamation point, for example, corresponds to different ways in which the utterance is intended to be interpreted, for example:

(1) a. OK. b. OK, c. OK . . . d. OK? e. OK!

Here, the first is intended as a statement, the second as an unfinished statement, the third as a tentative statement, the fourth as a question, and the fifth as an emphatic statement. These are clearly examples of prosodic functions, and the forms by which these functions are encoded may vary considerably from language to language and even from dialect to dialect. The example of *OK* is particularly convenient for this, because the expression is used in numerous languages throughout the world, so the representations in (1) are not necessarily specific to a particular language.

This distinction between prosodic forms and prosodic functions[1] is, in my view, one of the sources of the disparity among prosodic annotation systems. In a ToBI annotation, such as, for example:

(2) a. OK
b. % H*L-L% //

the symbols H and L are clearly annotating prosodic form, while the symbols *, -, and % seem much closer to indicating prosodic functions: whether a syllable is prominent, whether it is separated from what follows. Interestingly, Wightman (2002) reported that there was considerably higher intertranscriber agreement over the place of accents and boundaries than over the nature (H or L) of those accents and boundaries. In terms of our distinction between prosodic forms and prosodic functions, this can be interpreted as showing that listeners are much more proficient at identifying prosodic functions than prosodic forms. This seems a reasonable result, because it is well known that speakers and listeners are far more proficient at performing linguistic tasks (i.e., interpreting utterances) than metalinguistic tasks, such as describing the forms of utterances.

It is also an interesting result in the light of the fact that for computers, the results are exactly the opposite: machines are excellent at describing the forms of utterances but rather poor at interpreting them.

Prosodic forms, however, do not constitute a single level of representation. In Hirst, Di Cristo, and Espesser (2000), I suggested that we need to distinguish at least three levels of prosodic form. Most abstract would be a level of underlying phonological representation, which would be the level at which the prosodic form of an utterance is linked to its prosodic function. Between this underlying phonological representation and the acoustic signal, I believe that we can usefully distinguish two more levels: a phonetic representation and a surface phonological representation.

The term *phonetic representation* has been used to cover a number of different phenomena that really need to be distinguished. Phonetics is sometimes used as a synonym for the acoustics and physiology of speech. It should be clear that I wish to distinguish these levels of analysis. *Phonetics* is also sometimes used as a synonym for *surface phonology*, as in the terms *phonetic transcription* or the *International Phonetic Alphabet*. My use of the term *phonetic* follows that of Trubetzkoy (1949), for whom the distinction between phonology and phonetics is one between discrete and continuous phenomena. In this sense, then, a "phonetic transcription" would more appropriately be termed a "surface phonological transcription."

The level of surface phonology is a level of distinctive discrete categories with which we can describe surface phenomena cross-linguistically. The level of phonetics is the level of continuously variable phenomena from which we have factored out universal constraints on the production and perception of sounds. There has been much recent discussion about the interface between phonetics and phonology. Therefore, it would seem rather that phonetics is itself an interface: the interface between cognitive representations (phonology) and the physical manifestations of sound, perception, and articulation.

Neither of these two levels of representation, phonetics and surface phonology, should necessarily be thought of as cognitive representations. It is quite possible that in a complete theory of prosodic representation, neither of these levels would be necessary. Nevertheless, both levels appear to be useful as heuristic descriptive tools, particularly in the task of describing the prosody of languages that have not yet been adequately studied.

In the rest of this chapter, I first give a brief sketch of a functional prosodic annotation system that I developed many years ago for English intonation (Hirst 1977). I then discuss the relationship of this annotation system to the notion of prosodic structure. After this, I work back up from the acoustic signal via the levels of phonetic representation and surface phonological representation, before giving some consideration to the nature of underlying phonological representation of prosody.

Most of the examples I give are in English and a few in French. These are two languages that are geographically close but prosodically rather different. There is nothing, however, in the framework I describe that is specific to either of these two languages, although it remains an empirical question whether the framework could be applied to other languages without substantial modification.

3.3 Prosodic Functions

In Hirst (1977), a revised and translated version of my unpublished PhD thesis (Hirst 1974), I proposed a system of functional annotation to account for some of the ways in which intonation contributes to the meaning of an utterance. The book, written just a few years after Chomsky and Halle's *Sound Pattern of English* (1968), proposed a functional description of English intonation by means of a set of five distinctive features for which I coined the term *intonative features*: [±STRESS; ±CENTRE; ±EMPHASIS; ±BOUNDARY; ±TERMINAL]. These features were transcribed using the following symbols:

(3) a. ‘no b. °no c. <u>no</u> d. /no, no / e. no+ f. no //

The symbol in (3d) was used either at the beginning of an utterance or where the terminal/ nonterminal nature of the boundary was not relevant, while (3e) and (3f) correspond, respectively, to [+BOUNDARY; –TERMINAL] and [+BOUNDARY; +TERMINAL].

These features were each justified by examining minimal pairs where only a difference of the value of the feature in question corresponded to a clear difference in meaning. In most cases, the ambiguity is not complete, and contexts can be imagined where the interpretations given here could be reversed, but it stills seems that, at least for speakers of British English, the most obvious interpretation of each utterance corresponds to the one given here and that this is triggered by the feature described.

The feature [±STRESS],[2] for example, was justified by the minimal pair:

(4) a. he ‘ate a little ‘pudding (= he ate a small quantity of pudding)
b. he ‘ate a ‘little ‘pudding (= he ate a small pudding)

The feature [±CENTRE] (also known as *intonation centre*, *nuclear stress*, *primary sentence accent*, *tonic syllable*, etc.) can be justified by the minimal pair:

(5) a. it’s a ‘good ‘job °too. (= It’s a good thing too)
b. it’s a ‘good °job, °too (= The job is also a good one)

The feature [±EMPHASIS] can be justified by the pair:

(6) a. if you °give it to me, I’ll °mend it (= I’ll mend it for you)
b. if you <u>°give</u> it to me, I’ll °mend it (= I’ll only mend it if you let me keep it)

The feature [±BOUNDARY] (i.e., intonation unit boundary) can be justified by the following pair:

(7) a. she ‘should have °phoned / her ‘mother was °worried (= her mother)
b. she ‘should have °phoned her / ‘mother was °worried (= our mother)

Finally the feature [±TERMINAL] can be justified by the pair:

(8) a. / ‘would you like °tea+ or °coffee+ (= or something else?)
b. / ‘would you like °tea+ or °coffee// (= which?)

My conclusion, in the presentation of these “features,” and which I still believe today, was that:

> the five intonative features we have described are sufficient to account for all the syntactic ambiguities we have so far come across and which are disambiguated by intonation, including a considerable number of ambiguities which up to now have always been treated from a semantic, “attitudinal” point of view. (Hirst 1977, 44)

This does not mean that these five features exhaust the number of different intonational meanings that can be conveyed by prosody. They do, however, to my mind, constitute a minimal “core” system of annotation for prosodic functions, which any theory of intonational meaning should be able to account for.

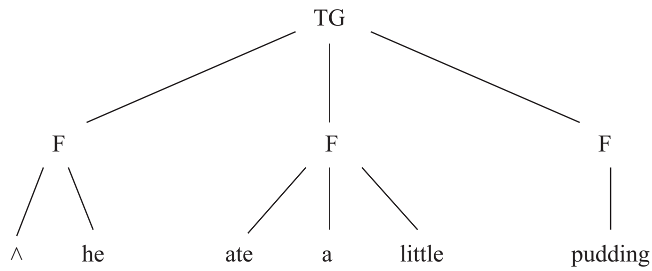
3.4 Prosodic Structure

Since I wrote my thesis and book, a lot of research has been done, particularly in the framework of autosegmental-metrical phonology, on the nature of prosodic structure (Liberman and Prince 1977, Selkirk 1984, Beckman and Pierrehumbert 1986, Nespor and Vogel 1986, Goldsmith 1990, Ladd 1996). It has today become a common point of view that both stress and boundaries are nothing but the reflection of a prosodic structure that is, at least to some extent, independent from syntactic structure. This seems a more convincing approach than a distinctive feature analysis.

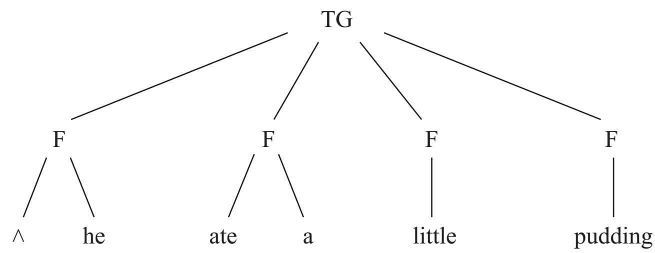
3.4.1 The Stress Foot

Halliday (1967), following Abercrombie (1964), had already suggested that an English utterance is a hierarchical phonological structure, containing four layers: the *tone group*, the *foot*, the *syllable*, and the *phoneme*. With this model, examples like (4) could be represented phonologically (omitting the level of the syllable and phoneme) as:

(9) a.



b.

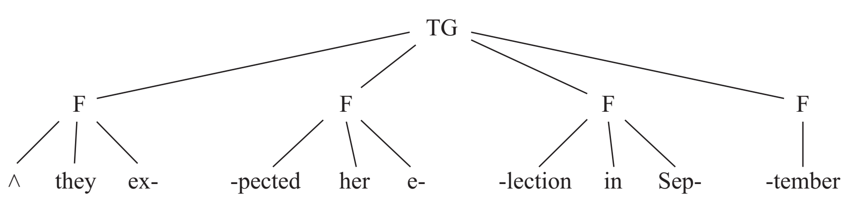


The symbol [^] in the first foot is what Halliday calls a "silent ictus," corresponding to a rhythmic beat occurring before the first syllable of the foot, just as a rest is used in musical notation for a similar purpose.

The Abercrombie/Halliday foot, often called the *stress foot* to distinguish it from the more restricted phonological unit of autosegmental-metric phonology, can be defined for speech as a sequence of syllables beginning with an accented syllable or with a silent beat at the beginning of a sentence, and continuing up to (but not including) the next accented syllable or the end of the *tone group*.

In (9), the foot boundaries all correspond to word boundaries, but this is not always necessarily the case. In (10), none of the foot boundaries corresponds to a word boundary, except the last.

(10)

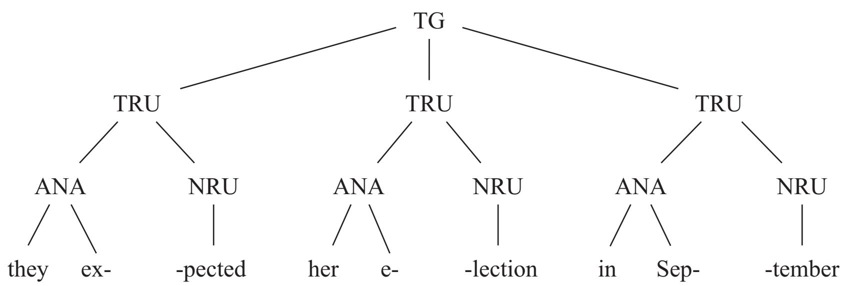


3.4.2 Tonal Units and Rhythm Units

Unlike Abercrombie and Halliday, who use the same unit to describe both melody and rhythm, Jassem (1952), more than ten years earlier, had made a clear distinction between what he called the *tonal unit*, which is conceived of as the domain of occurrence of local pitch movements in English, on the one hand, and the *narrow rhythm unit* and the *anacrusis*, on the other hand, conceived of as the domain of segmental timing and rhythm. The tonal unit, in Jassem's model, was in fact identical to the unit which was, thirteen years later, called the *foot* by Abercrombie, as defined in section 3.4.1.

The narrow rhythm unit (NRU) is similar to the foot, except for the fact that it does not usually cross word boundaries, except in cases of enclitics, which are treated prosodically as belonging to the previous word. Any syllables that are not part of a NRU form an anacrusis (ANA), in which, according to Jassem, the syllables are "pronounced extremely rapidly" (40). The ANA, together with the following NRU, together make up what Jassem termed the *total rhythm unit* (TRU). Example (10) in this analysis would look like (11):

(11)



Jassem gave several examples where his representation allowed subtle distinctions of rhythm to be made that could not be captured without the notion of ANA. In the following minimal pair, for example:

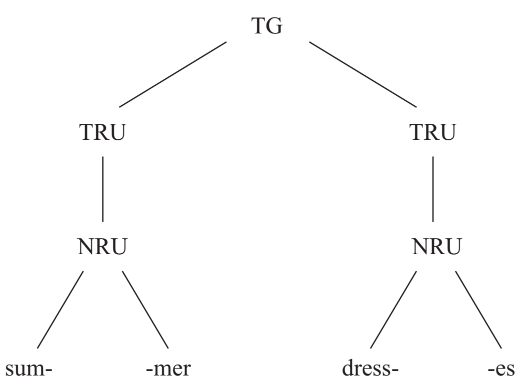
(12) a. Summer dresses b. Some addresses

Jassem notes that although the phonemes and stresses are identical, there is a subtle difference of rhythm between the two: the first syllable of (12a) is shorter than that of (12b), whereas the second syllable is longer in (12a) than in (12b). He attributes this difference to the fact that the first two syllables of (12a) constitute a single NRU, whereas in (12b) the first syllable constitutes an NRU on its own and the second syllable constitutes an ANA. He proposed to represent this in the phonetic transcription by the simple device of a space after each NRU, as in:

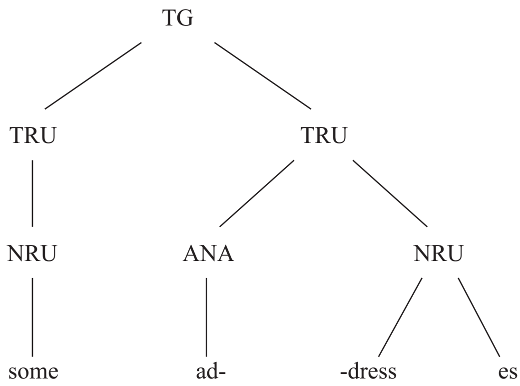
(13) a. /ˈsʌmə ˈdresiz/ b. /ˈsʌm əˈdresiz/

corresponding to the tree representations:

(14) a.



b.



In a recent collection of articles published to celebrate Wiktor Jassem's ninetieth birthday, I suggested (Hirst 2012) that Jassem's TRU does not actually play any phonological role. Instead we can combine the tonal units and the rhythm units (ANAs and NRUs) into a single representation. In fact, if we do that, we can note that there is no longer any need to make a formal distinction between ANA and NRU, because the NRU will always coincide with the beginning of a tonal unit: both can simply be characterized as rhythm units. It is here that the distinction between stress and accent plays a crucial and, I believe, particularly interesting role.

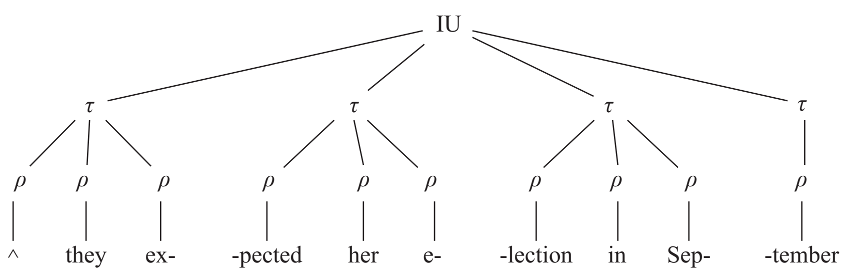
Following Bolinger (1958), I interpret *accent* as referring to the physical manifestation of prominence in an utterance and *stress* as referring to the syllable or syllables of a word that are lexically marked and normally carry this prominence when the word is accented.

In my interpretation of Jassem's model, then, the *tonal unit*, which I label [τ], begins with an accented syllable, or at the beginning of the *intonation unit*, which I label [IU] (= Jassem's tone group) and continues until the next accented syllable or until the end of the intonation unit.

Deviating slightly from Jassem's original formulation, but not, I hope, from the spirit of his analysis, I now define a *rhythm unit* (which I label [ρ]) as beginning either at the beginning of a word or with a stressed syllable and ending at the end of the word or before a stressed syllable. Once again, some cases of enclitics should be considered as forming part of the preceding word.

With this formulation, we can now combine the two levels of representation into a single prosodic tree as in (15):

(15)



Jassem's categories can be derived from this representation: an NRU is the leftmost [ρ] in a [τ], provided it does not begin with [^], and all other consecutive [ρ]s can be grouped to form Jassem's ANA.

It can be seen that the rhythm unit [ρ] and the tonal unit [τ], thus defined, form an interface between the lexical representation of the word and the actual pronunciation of the utterance, or between stress and accent as defined by Bolinger (1958).

In a study of the segmental duration of a large corpus of English, Hirst and Bouzon (2005) found that, as predicted by Jassem, but contrary to Halliday's model, word boundaries do play an important role in the rhythmic structure of English. Strong, negative correlations were found between the duration of a segment and the number of phonemes in the stress foot, in the NRU, and in the word, but no similar effect was found either in the syllable or in the ANA. Moreover the correlation was greater for the NRU than either the stress foot or the word, thus confirming Jassem's predictions.

3.4.3 Emphasis and Terminality

It seems, then, that the prosodic functions corresponding to stress/accent, centre, and boundary can be more adequately accounted for by an explicit model of prosodic structure such as I have just outlined.

The remaining two functions I mentioned, *emphasis* and *terminality*, do not look as if they can be accounted for by a difference in prosodic structure. One possibility, which I have tentatively suggested (Hirst 1983, 1998), is that these features could be accounted for as prosodic morphemes, whose function is similar in many ways to sentence particles, as found in many languages. Wakefield (2009, 2012) develops this idea more explicitly with a direct comparison between Cantonese sentence particles and English intonation patterns.

I will assume here, for the sake of argument, that English contains (at least) two abstract prosodic morphemes, which I will represent as [E] for an emphatic particle and [Q] for a question particle, because one of the major functions of [±TERMINAL] is expressing prosodically a distinction between a statement and a question. If we want to distinguish between the intonation of questions and that of unfinished statements, a debatable issue (as I discuss in Hirst 1998), we could suppose a third particle for that.

Question particles are well known in the description of many languages, both tonal and nontonal, for example, *ma* in Mandarin Chinese and *li* in Russian. Emphatic particles are also common. In Bambara (Mali; Bird Hutchinson, and Kante 1977), a sentence such as:

(16) Muso fila bɛ́ Sísé fɛ. (*Sissé has two wives*)

can be modified by the emphatic particle *de*, giving:

(17) Muso de fila bɛ́ Sísé fɛ. (*Sissé has two WIVES*)
(18) Muso fila de bɛ́ Sísé fɛ. (*Sissé has TWO wives*)

or

(19) Muso fila bɛ́ Sísé de fɛ. (*SISSE has two wives*)

where each time, the emphatic particle *de* immediately follows the emphasized word.

In section 3.8, I return to the question of how these abstract particles may be interpreted prosodically.

3.5 Phonetic Representation

3.5.1 Models of Fundamental Frequency

The search for an appropriate scale for measuring fundamental frequency has been one part of a systematic attempt, in particular by researchers from Holland ('t Hart, Collier, and Cohen 1990), to develop a model of the way in which pitch is perceived. This was done by stylizing raw fundamental frequency patterns as a sequence of straight lines, such that when the stylized frequency is used to resynthesize the utterance, the result is judged to be perceptually equivalent to the original intonation pattern.

Another approach to modeling pitch has been to attempt to model the way in which pitch is produced by speakers. In particular, work by Fujisaki (1991) and his colleagues has applied a model of pitch production to a large number of languages, including several tone languages, analyzing an intonation pattern as the superposition of three underlying components: a global baseline component, a sequence of phrasal components, and a sequence of shorter accent components. These three components are added in the logarithmic domain to produce a raw fundamental frequency curve.

A third approach has been to develop acoustic models that are neither directly models of speech perception nor of speech production but are compatible with both. This is the approach that I follow in this presentation.

3.5.2 Micromelodic Effects

Fitting a raw F0 curve with a mathematical model is not a simple, straightforward problem due to the fact that fundamental frequency curves are not always continuous: unvoiced portions of the utterance have no associated F0. Even when the curve is continuous, it is often not smooth, and this type of irregularity is hard to model simply.

If we look at a two-second extract of an utterance, corresponding to the words, "More news about the Reverend Sun Myung M(oon),"[3] we notice (see figure 3.1a) that the beginning and end of the F0 curve are in fact fairly continuous and smooth. The reason for this becomes obvious when we look at the phonemes associated with the curve as in figure 3.1b.

The smooth portion at the beginning of the extract corresponds to the phonemes /mɔːnjuː/, whereas that at the end of the extract corresponds to /ʌnmjʌŋm/. All of these phonemes are fully sonorant, either vowels, semivowels, or nasals. The discontinuity and irregularity of the F0 curve are due to the presence of obstruents in the utterance: stops and fricatives, which either interrupt the curve (for voiceless obstruents) or make it irregular (for voiced obstruents). The effect of these consonants has been called *micromelodic*, as distinct from the *macromelodic* characteristics of larger pitch movements associated with accents and intonation patterns. Micromelodic effects, then, are caused by the aerodynamic characteristics of the articulation of different phonemes (Di Cristo and Hirst 1986). Phonemes like vowels and sonorants, which hardly obstruct the airflow, have virtually no micromelodic effect, whereas stops and fricatives disturb or interrupt the flow.

The raw fundamental frequency curve then can be thought of as the interaction between two components: a *micromelodic* component, which is conditioned by the segmental nature of the individual speech sounds, and a *macromelodic* component, which corresponds to the underlying laryngeal gesture. This corresponds to the observation by Nooteboom (1997) that we do not perceive the observable discontinuities of raw pitch patterns unless they are longer than about two hundred milliseconds, as if

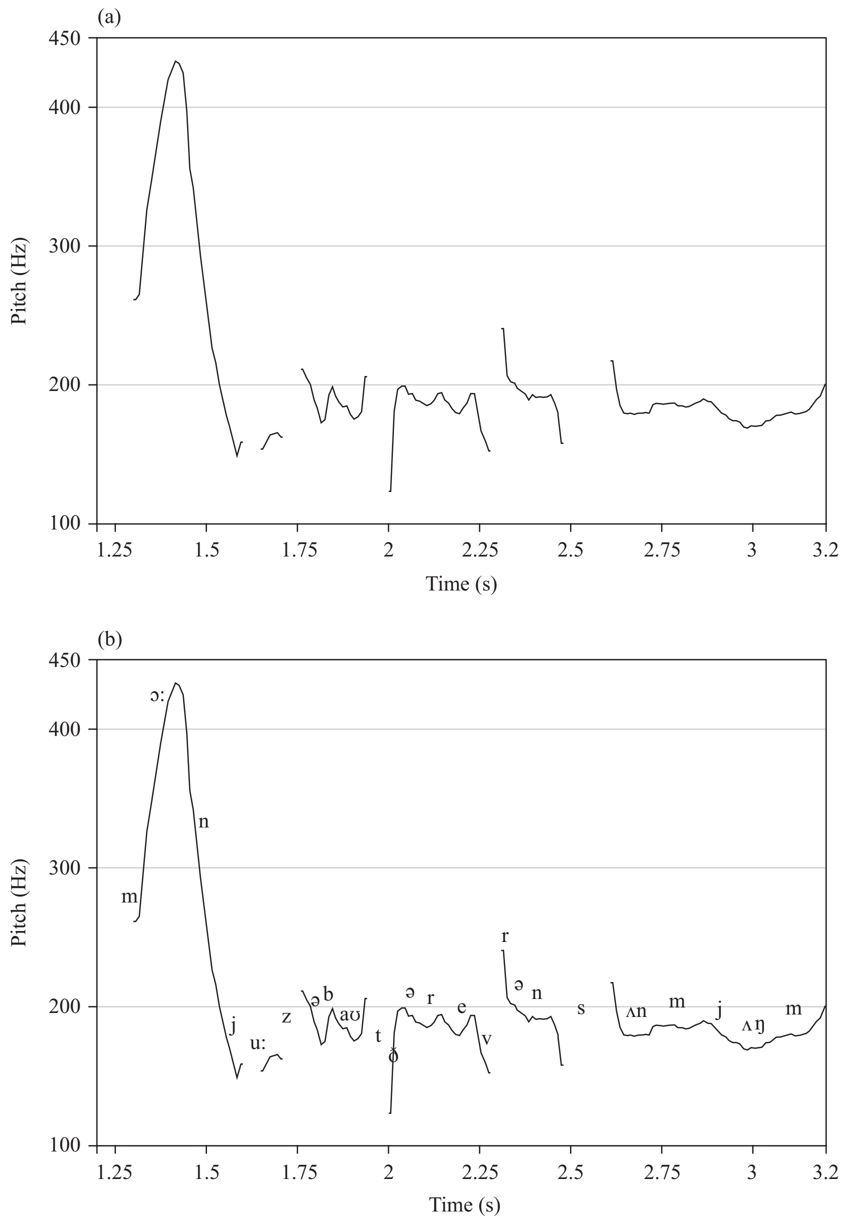


Figure 3.1
Two-second extract of the F0 curve corresponding to "More news about the Reverend Sun Myung M(oon)." (a) Raw F0. (b) F0 and phonemes.

human perception unconsciously bridges the silent gap by filling in the missing part of the pitch contour.

Linguists have been aware for a long time that fundamental frequency curves obtained from utterances containing only sonorants and vowels are much better behaved than raw F0 curves obtained from unrestricted speech. It is for this reason that linguists have often constructed sentences consisting of mainly sonorants and vowels, such as Eva Gårding's (1998) "Madame Marianne Mallarmé har en mandolin

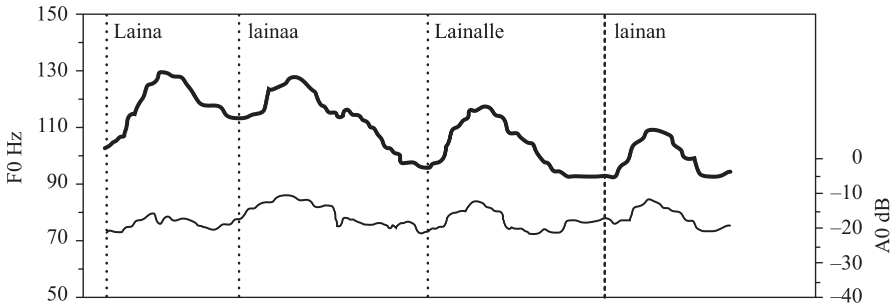


Figure 3.2
An example of a Finnish intonation pattern on a sentence with only sonorant phones. The sentence is "Laina lainaa Lainalla lainen" (Laina lends Laina a loan) (Iivonen 1998). The top curve is the fundamental frequency, and the lower curve is the intensity.

från Madrid" (Madam Marianne Mallarmé has a mandolin from Madrid) for Swedish; Hiroya Fujisaki's (2004) "Aoi aoinoewa yamanouenoieni aru" (The picture of the blue hollyhock is in a house on top of the hill) for Japanese; or Annti Iivonen's (1998) "Laina lainaa Lainalla lainen" (Laina lends Laina a loan) for Finnish (pictured in figure 3.2).

3.5.3 Macromelody and Micromelody

The idea, then, is that a raw intonation pattern is the interaction between two independent components: a macromelodic component determined by the accentuation and intonation of the utterance and a micromelodic component determined by the segmental phonemes.

If we compare two simple French phrases such as *À ton papa* (To your daddy) and *À ma maman.* (To my mummy), pronounced as statements, we can see that there is the same underlying macromelodic pattern for the two utterances and that the surface differences are simply due to the different phonemes of the utterances, voiceless stops in figure 3.3a and sonorant nasals in figure 3.3b.

What is particularly worth noting is that the F0 curve shown in figure 3.3a is practically superposable on that of figure 3.3b. It seems as if the F0 curve continues to change during the voiceless segments of the utterance, even though it is not visible. This is actually not surprising if we think in terms of a continuous change of tension of the vocal folds, which can continue to change even during voiceless segments.

If we now compare these patterns with those observed on the same phrases pronounced as questions, as in figure 3.4, we see once more that the two contours are practically superposable. And once again it seems clear that the F0 curve in figure 3.4a continues to change during the voiceless segments of the utterance.

In particular, the final rise on *papa* does not begin at the onset of the final vowel: the F0 at this point is already considerably higher than that at the end of the preceding vowel.

Notice that this idea of a continuously varying underlying pitch contour is not the model that is generally assumed in phonological descriptions of tonal and intonation contours. In the majority of these studies, it is assumed that tones are directly associated with vowels, which are the tone-bearing elements (see Goldsmith 1990, 44, for example) and that the fundamental frequency observed on the consonants is simply an interpolation between the tones on the vowels. If that were so, then it might be

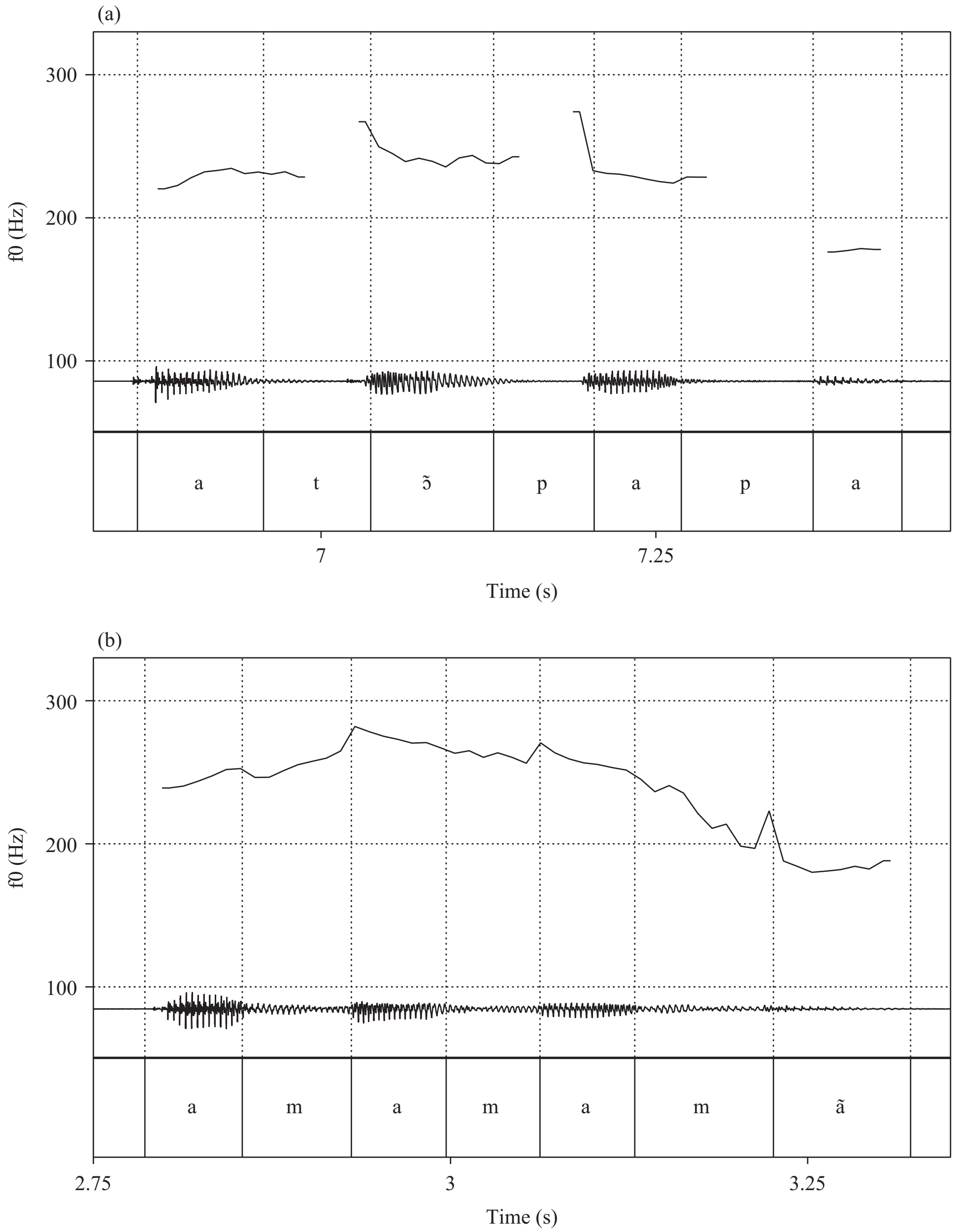


Figure 3.3
Two French phrases pronounced with a declarative intonation pattern. (a) *À ton papa*. (To your daddy.) (b) *À ma maman*. (To my mummy.)

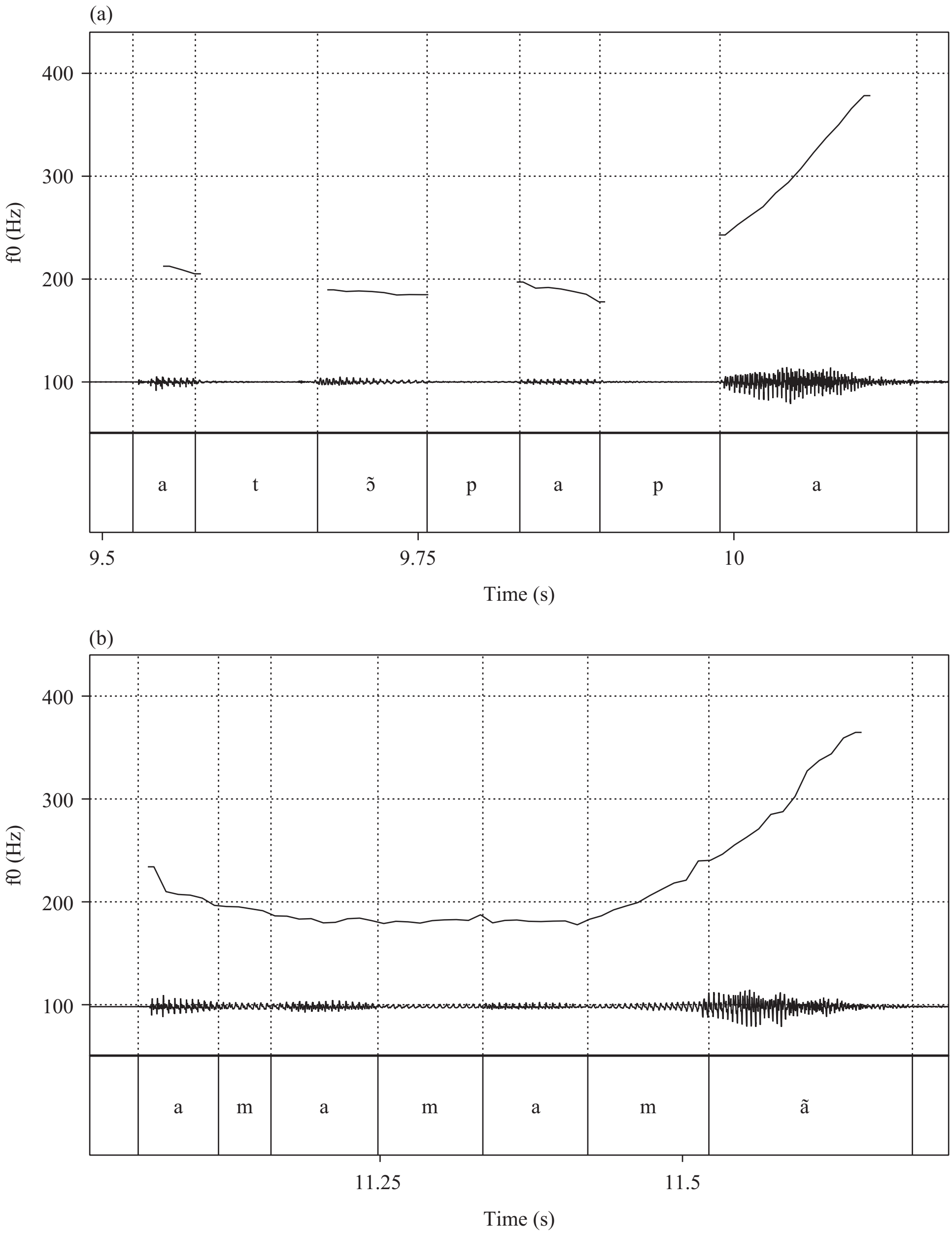


Figure 3.4
The same two French phrases as in figure 3.3 pronounced with an interrogative intonation pattern. (a) *À ton papa?* (To your daddy?). (b) *À ta maman?* (To my mummy?).

thought that the pitch curve visible in figures 3.3a and 3.4a is actually closer to the underlying form than that in figures 3.3b and 3.4b, which are simply the result of an interpolation on the sonorant consonants. The fact that the F0 curve follows the same trajectory in utterances with voiceless consonants as the smooth and continuous curve observed on the utterances with sonorants, however, and, in particular, the fact that the curve continues to evolve during the nonvoiced portions of the utterance, seems to me convincing evidence that the planning of these curves is the result of an underlying macromelodic pattern on which the micromelodic variations are subsequently superimposed.

3.5.4 A Model for F0 Curves

3.5.4.1 Macromelodic and micromelodic profiles The underlying macromelodic component of an intonation pattern, then, has, I shall assume, the two characteristics of being smooth and continuous. This is fortunate because modeling a discontinuous or irregular function is much more difficult than one that is continuous and smooth. To return to the example in figure 3.1, the underlying macromelodic profile of the curve might be something like the continuous curve in figure 3.5a, which could correspond to what we would produce if we were to hum the sentence instead of pronouncing it. I will return (section 3.5.5) to how the continuous curve was actually obtained; for the moment let us just assume that we have such a macromelodic profile.

Once we have a macromelodic profile, we can derive the micromelodic profile by dividing each value of the raw F0 curve by the corresponding value of the modeling function. This gives a result as displayed in figure 3.5b, where the smooth, continuous curve is the macromelodic profile and the discontinuous curve is the micromelodic profile derived as just described.

Notice that such a modeling technique is not simply a stylization of the raw F0 curve, because the raw curve has actually been factored into two orthogonal components without any loss of information. Multiplying each value of the continuous curve from figure 3.5b by that of the discontinuous curve in the same figure gives the original raw F0 as in the curve in figure 3.1a. For speech synthesis, it would be possible to model the micromelodic profile itself and to use this to improve the segmental quality of the utterance (for an application to Arabic, see Chentir, Guerti, and Hirst 2009). For the study of intonation, the resynthesis of the utterance with the macromelodic profile is generally of sufficiently high quality to evaluate the appropriateness of the modeled curve.

3.5.4.2 F0 transitions One of the simplest ways to model a smooth, continuous function such as that in figure 3.5a is as a piecewise sequence of transitions between successive points on the curve. In previous work, until recently, I referred to these points, following a fairly long tradition, as *target* points, but it should be noted that this name was not intended to imply that the “target points” necessarily have any actual psychological reality for the speaker and listener. To make it clear that these points are not intended to represent cognitive targets, like those, for example, in the model proposed in Xu and Wang (2001), I now prefer to use the term *anchor points*, which more clearly corresponds to their role in describing a pitch curve.

The advantage of a piecewise function over a global function is that each segment of the curve is defined locally by its own set of parameters, which means that a modification of one portion of the curve does not entail modifications throughout the rest of the curve. The simplest model would be a linear transition between two anchor points as in figure 3.6a, where the transition is defined by the function:

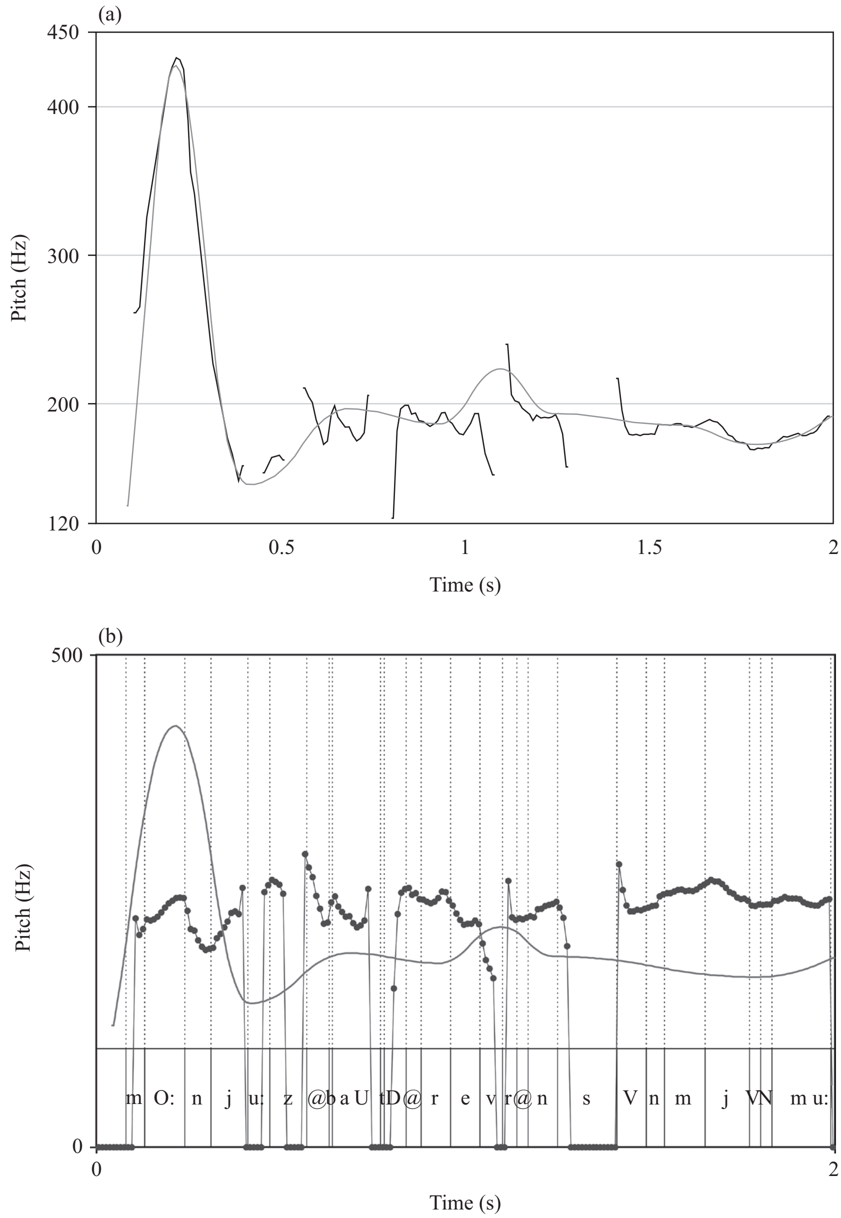


Figure 3.5
(a) Raw F0 (discontinuous) together with macromelodic profile (continuous) for the first two seconds of recording A01. (b) Same as panel (a), with the superimposed micromelodic profile.

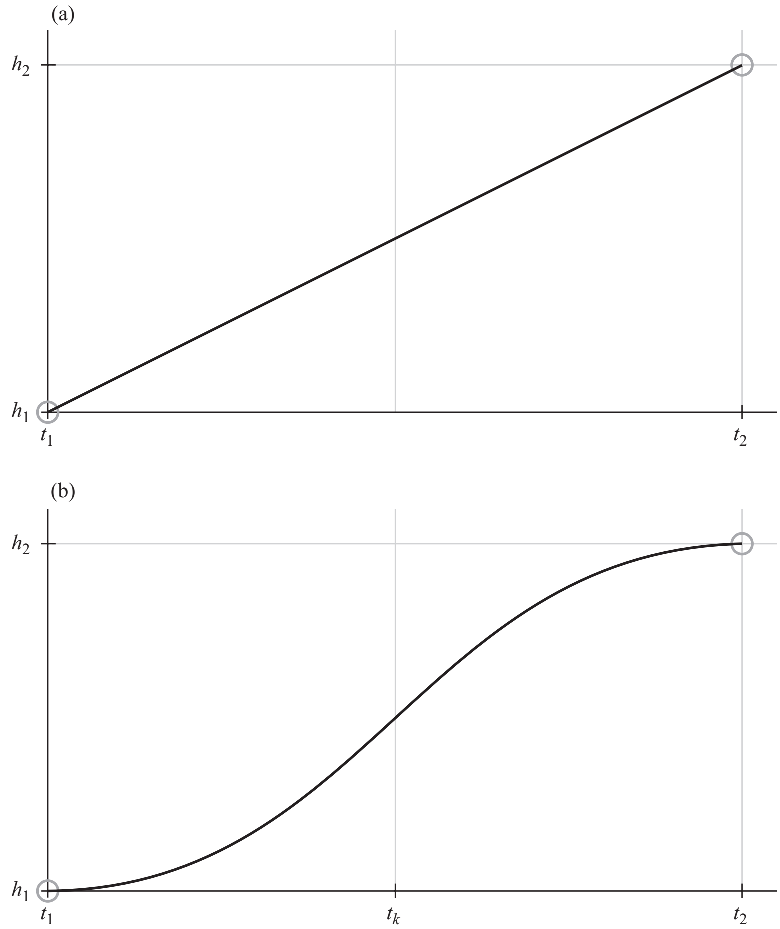


Figure 3.6
Linear and quadratic transitions from a first anchor point <t_1, h_1> to a second point <t_2, h_2>. (a) A linear F0 transition. (b) A quadratic F0 transition.

$$h_i = h_1 + \frac{(t_2 - t_i)}{(t_2 - t_1)} \cdot (h_2 - h_1) \tag{3.1}$$

where h_1 and h_2 are the F0 values of two adjacent anchor points and t_1 and t_2 are the corresponding time values of these anchor points. Here the F0 value (h_2) of the second anchor point is higher than that (h_1) of the first anchor point, but the same reasoning would apply if it had been lower.

Naturally occurring F0 curves are curvilinear, not linear. A number of mathematical functions have been used in the past to model such functions. One of the simplest of these is a quadratic transition, corresponding to a constant acceleration followed by a constant deceleration of the pitch change as shown in figure 3.6b and as defined by the functions:

$$t_i \in [t_1 \ldots t_k] : h_i = h_1 + \frac{(h_2 - h_1) \cdot (t_i - t_1)^2}{(t_k - t_1)(t_2 - t_1)} \tag{3.2}$$

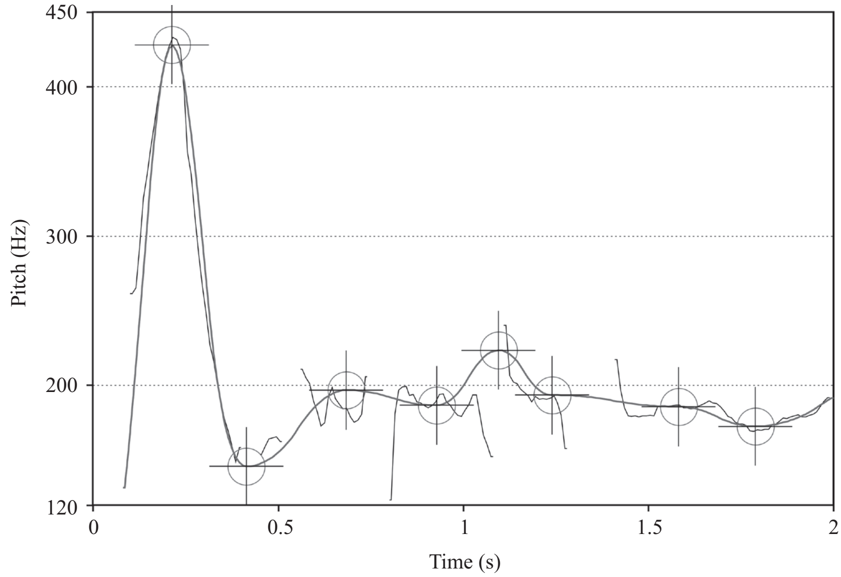


Figure 3.7
Macromelodic profile for a two-second extract from recording A01, defined as quadratic transitions between anchor points.

$$t_i \in [t_k \ . \ . \ . \ t_2] : h_i = h_2 + \frac{(h_1 - h_2) \cdot (t_i - t_2)^2}{(t_k - t_2)(t_1 - t_2)} \tag{3.3}$$

As can be seen from figure 3.6b, in the case of a rise, the transition consists of a concave curve from time t_1 to time t_k, the point of maximum slope, followed by a convex curve from t_k to t_2.

Figure 3.7 shows the same extract we have seen several times now, with the anchor points that define the curve represented as circles.

3.5.5 Momel

A piecewise quadratic function such as that illustrated here is known as a quadratic *spline* function and has been in use in our laboratory since the early 1980s to model intonation patterns using an algorithm called *Momel* (for "modeling melody").

The Momel representation is formally equivalent to a subset of the contours that can be produced by the rise/fall/connection (RFC) model of intonation later developed by Paul Taylor (1994) as a tool for speech synthesis. The only difference is that the RFC model allows linear interpolations between two successive anchor points, as well as quadratic interpolations. If two successive anchor points have the same value of F0, then the transition will be linear (i.e., flat) with Momel too. I have personally never observed a case where a nonflat linear transition gives a better approximation to an F0 curve than a quadratic one.

The original implementation of Momel allowed the user to define anchor points manually by clicking on a representation of the F0 curve on the computer screen. These anchor points were converted into a quadratic spline function, and the user could then resynthesize the utterance using PSOLA resynthesis. Today, this can be done with Praat (Boersma and Weenink [1992] 2015) by creating a manipulation object and then removing and adding pitch points manually. Praat displays the pitch curve with linear interpolation between the pitch points, but an approximation of the quadratic spline function can be obtained by the command interpolate quadratically . . . , which was my own modest contribution to the Praat software.

Manual modeling of F0 is highly subjective, and it was for this reason that my colleague Robert Espesser and I developed an automatic version of the algorithm (Hirst and Espesser 1993) based on our experience of using the manual implementation of the model over a period of several years. The algorithm, which is described in detail in Hirst, Di Cristo, and Espesser (2000), uses a form of robust regression to optimize the modeling of raw fundamental frequency curves with a quadratic spline function.

The algorithm was later evaluated on a corpus of read speech in five languages (Eurom1 corpus) during the course of the Multext European project (Véronis, Hirst, and Ide 1994). Evaluators were instructed to add or delete anchor points of the modeled speech only when such corrections made an audible improvement to the resynthesis of the speech. Table 3.1 shows the statistics[4] for these corrections for the corpus of read speech, together with those for a corpus of spontaneous speech in French.

The results of the evaluation were very encouraging. The F-measures for the different languages showed a global efficiency of around 95 percent, and even the corpus of spontaneous French showed an F-measure of 93.4 percent even though the algorithm had not at all been optimized for spontaneous speech.

Examination of the errors in the anchor points showed that one type of error in particular occurred systematically. This concerned a pitch rise before a silent pause where, frequently, the algorithm missed the final rise entirely. An example of this type of error can be seen in figure 3.8. This error is understandable because the algorithm uses a local modeling technique, fitting a parabola to portions of the curve. As can be seen in the example, the raw F0 exhibits a concave portion corresponding to the acceleration of the pitch rise but hardly any convex portion corresponding to the decelerating portion of the rise. This is the reason why the original algorithm fails to produce a final anchor point.

The Momel algorithm has since been implemented as a Praat plug-in (Hirst 2007), which allows users to use its functions directly from the Praat menus without needing

Table 3.1
Results for the evaluation of the automatic Momel algorithm on read speech for English, French, German, Italian, and Spanish and for spontaneous speech in French (Fref corpus)

		Number of points			Evaluation		
Corpus	Language	Detected	Added	Deleted	Recall	Precision	F-measure
Eurom	English	8,380	623	125	93.0	98.5	95.7
	French	6,547	423	130	93.8	98.0	95.9
	German	13,595	1,145	506	92.0	96.3	94.1
	Italian	9,475	337	330	96.4	96.5	96.5
	Spanish	8,985	651	16	93.2	99.8	96.4
Fref	French	9,835	532	744	94.5	92.4	93.4

Note: The columns show the total number of anchor points detected, number of anchor points added manually, number of anchor points deleted, and the statistical measures of recall, precision and F-measure. *Source*: Data from Campione (2001).

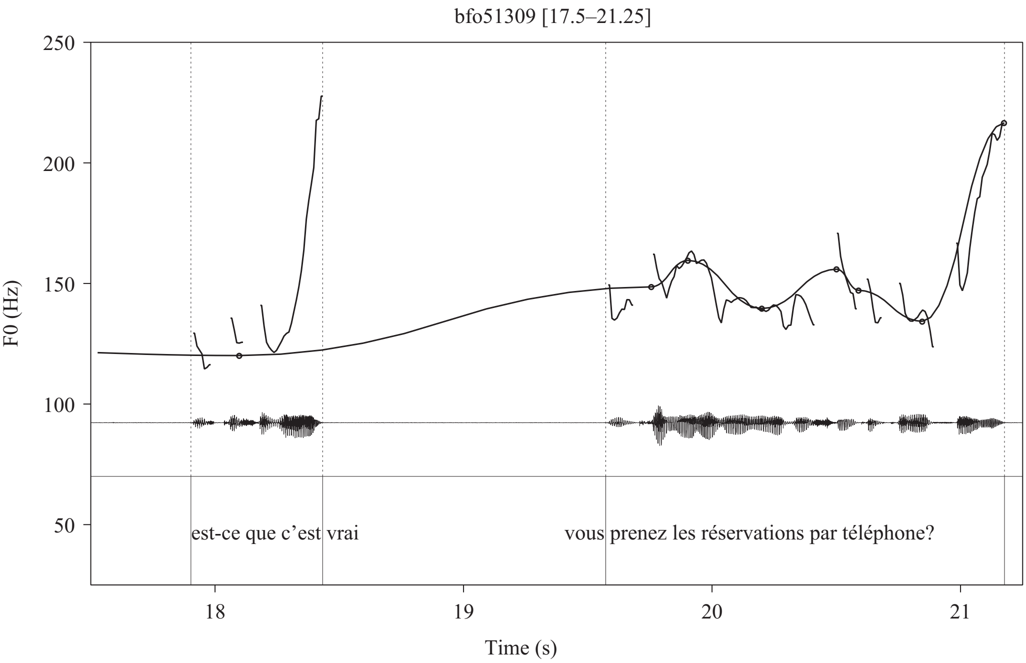


Figure 3.8
Old version of the automatic Momel algorithm for the French utterance, *Est-ce que c’est vrai? Vous prenez les réservations par téléphone?* (Is it true? You take bookings by phone?). Raw (discontinuous) and modeled F0 (continuous).

to handle scripts. The systematic error we have just seen was corrected by a special treatment before silent pauses. The concave part of the pitch rise is now extended to include a similar-shaped convex portion. In other words, to obtain the best fit for this pitch rise, a high anchor point is calculated and situated in the silent pause as near as possible to the previous low anchor point. The anchor point is calculated so that the concave portion of the pitch rise follows the raw data points as closely as possible. The result of this improved algorithm can be seen in figure 3.9.

An evaluation of the improved algorithm was carried out on a corpus of read speech in Korean (Hirst et al. 2007). It showed a significant and systematic improvement as compared to the older version of the algorithm. It is, naturally, desirable that the modeling tools we use should be as theory-neutral as possible. Complete neutrality, though, is obviously not entirely feasible, because any model necessarily makes some assumptions about the nature of underlying representations, as we saw in the discussion (section 3.5.4.1) of whether the underlying contour should be based only on the contours observed on the vowels or whether it should be modeled as a continuous underlying contour.

Rather than suggest that Momel is theory-neutral, then, I like to think that it is what we could call *theory-friendly*. I believe that the algorithm can be compatible with a number of different theoretical approaches to the description of speech melody. It has been used in the past as a first step for modeling with the Fujisaki model (Mixdorff 1999). It has also been used as first step for ToBI for both English (Maghbouleh 1998;

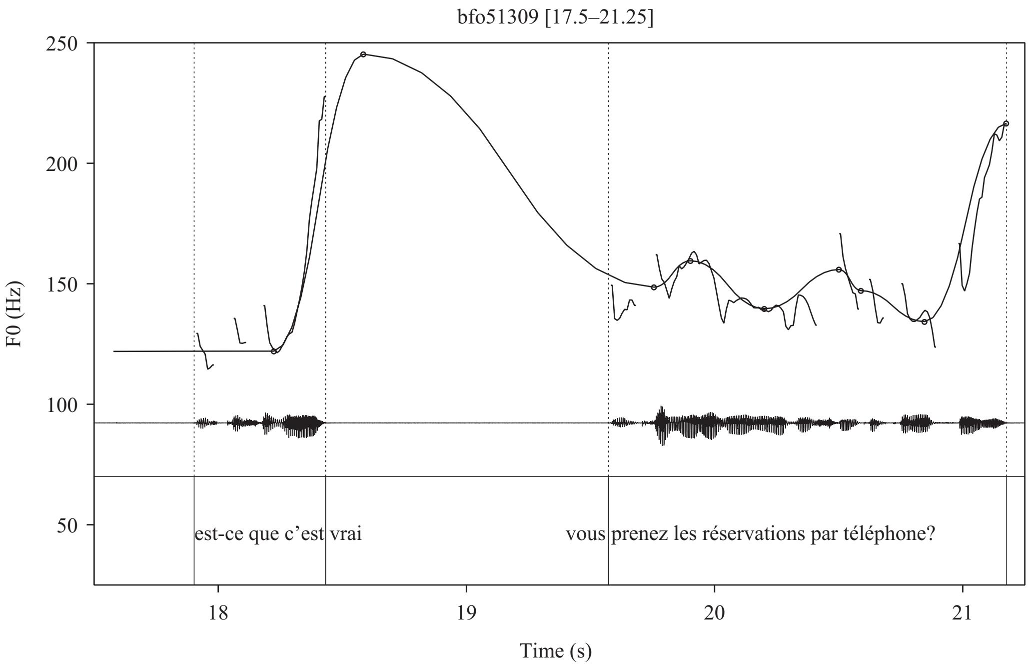


Figure 3.9
Improved version of the automatic Momel algorithm for the same utterance as in figure 3.8. Raw (discontinuous) and modeled F0 (continuous).

Wightman and Campbell 1995) and Korean (K-ToBI) (Cho and Rauzy 2008). It is also compatible with our own surface phonological representation alphabet, INTSINT.

3.6 Surface Phonological Representation

3.6.1 INTSINT: An International Transcription System for Intonation

Momel, as we have seen, provides a reversible modeling of the raw F0 with no loss of information because the raw F0 is factored into two components, macromelody and micromelody. Multiplying the two together recovers the original F0.

INTSINT was originally developed as a tool for linguists to provide a surface phonological representation of an intonation pattern. The original version of the system (Hirst 1987) was based on an inventory of minimal pitch contrasts found in published descriptions of the intonation patterns of numerous languages. The aim was to provide a tool for the systematic description of these intonation patterns, something along the lines of a narrow transcription using the IPA. Like the IPA, it was intended that INTSINT could be used for preliminary descriptions of intonation patterns, even for languages that had not previously been described. Notice that this aim is very different from that of the ToBI system (Silverman et al. 1992), which presupposes that the inventory of intonation patterns for the language being described has already been established. The official website for ToBI (http://www.ling.ohio-state.edu/tobi/) makes this particularly explicit:

> Note: ToBI is not an International Phonetic Alphabet for prosody. Because intonation and prosodic organization differ from language to language, and often from dialect to

dialect within a language, there are many different ToBI systems, each one specific to a language variety and the community of researchers working on that language variety.

The INTSINT system (whose name was suggested to us by Hans 't Hart in a personal communication) was presented in Hirst and Di Cristo (1998) and was used there for the annotation of the intonation of nine different languages. Basically, it describes an intonation contour as a sequence of tonal segments labeled using an alphabet of eight symbols. The tonal segments are assumed to be of three types:

Absolute tones: *t* (top), *m* (mid), *b* (bottom). These are assumed to refer to the corresponding position of the speaker's current pitch range.
Relative tones: *h* (higher), *s* (same), *l* (lower). Unlike absolute tones, relative tones are assumed to be refined with respect to the preceding tonal segment.
Iterative relative tones: *u* (upstepped), *d* (downstepped). These are also defined relative to the preceding tonal segment but generally involve smaller pitch changes and often occur in a sequence of steps either upward or downward.

In the chapters in Hirst and Di Cristo, the INTSINT tones were represented by graphic symbols between two horizontal lines and aligned with the text. These symbols were top [⇑], bottom [⇓], higher [↑], same [→], lower [↓], upstepped [<], and downstepped [>]. The mid tone was reserved for the unmarked onset of an intonation unit and was not marked.

In most later publications, the capital letters T, M, B, H, S, L, U, and D were used instead of the graphic symbols. Since Hirst (2011), I prefer to represent the INTSINT tones with lowercase rather than uppercase letters. This may help to avoid confusion with other more abstract coding schemes such as ToBI (Silverman et al. 1992), or the even more abstract underlying representation used in Hirst (1998) and section 3.8, both of which use some of the same letters as INTSINT.

3.6.2 Mapping from INTSINT to Momel

Although INTSINT was introduced as a descriptive tool for linguists, since its introduction was later than the creation of the F0 modeling tool, I already had in mind the possibility that this surface phonological annotation could be linked to the analysis of the F0 curve as a sequence of phonetic anchor points. I anticipated, then, that it might be possible to map the output of the Momel algorithm onto a sequence of symbols from the INTSINT alphabet. To do this, following the idea of an analysis by synthesis paradigm, it was first necessary to define a mapping in the other direction, that of synthesis. Some of the history of the way in which this mapping was defined is described in Hirst (2005).

In its current implementation, the mapping depends on two speaker-/utterance-specific parameters called *key* and *span*, which together define the speaker's pitch range.[5] The *key* (like a musical key) defines a central reference point for the speaker's pitch range, and the *span* defines the maximum and minimum pitch values of the range, which are taken to be symmetrical (on a logarithmic scale) above and below the speaker's key.

The two parameters together define three absolute tones—top, mid, and bottom—with respect to the speaker's pitch range, as in the following formulas, which assume that the value of key is given in hertz and the value of span in octaves:

$$t = key \times \sqrt{2^{span}} \quad m = key \quad b = \frac{key}{\sqrt{2^{span}}} \tag{3.4}$$

The pitch anchor points corresponding to the relative tones are then defined with respect to both the preceding anchor point (here called *p*) and the top (*t*) or bottom (*b*) of the range.

An anchor point coded *h* is simply defined as the geometric mean (i.e., the mean on a log scale) of the preceding anchor point and the top of the range. It thus corresponds to a pitch movement that moves up halfway toward the value of *t*. As can be expected, an anchor point coded *s* is defined as the same as the preceding anchor point. Symmetrically to *h*, an anchor point coded *l* is defined as the (geometric) mean of the preceding anchor point and the bottom of the range.

$$h = \sqrt{p * t}; \quad s = p; \quad l = \sqrt{p * b} \tag{3.5}$$

For the iterative tonal segments *u* and *d*, the implementation defines the values as the (geometric) mean of the value of the preceding anchor point and that which would be obtained if the anchor point were coded *h* or *l*, respectively. In other words, these anchor points correspond to a pitch excursion one-quarter of the way to the top or bottom of the pitch range.

$$u = \sqrt{p * \sqrt{p * t}}; \quad l = \sqrt{p * \sqrt{p * b}} \tag{3.6}$$

Assuming, once again, a logarithmic scale for the pitch range, these values are illustrated graphically in figure 3.10.

This implementation obviously makes a number of assumptions, most of which would be open to empirical investigation. This is one of the major advantages of an explicit model such as this.

One consequence of the model, which was not specifically intended but which turns out to be fortunate, is that a sequence of alternating *h* and *l* tones will automatically introduce an iterative lowering of the tones, much like the probably universal effect of downdrift that has been described in the literature on tone and intonation as occurring in languages throughout the world. This downdrift effect, as illustrated in figure 3.11, is thus an automatic by-product of the way in which the relative tones are defined.

Since it was first developed, the Momel algorithm has been applied relatively successfully to a number of different languages, including English, French, Italian, Catalan,

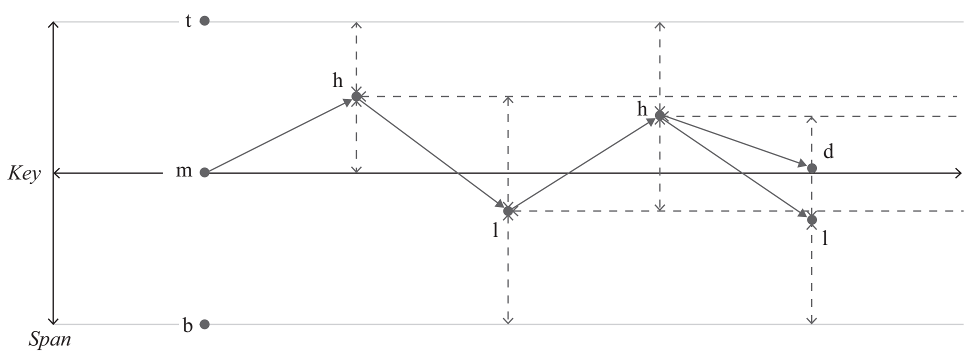


Figure 3.10
Graphic illustration of the mapping from INTSINT to Momel with pitch range defined by the two parameters key and span.

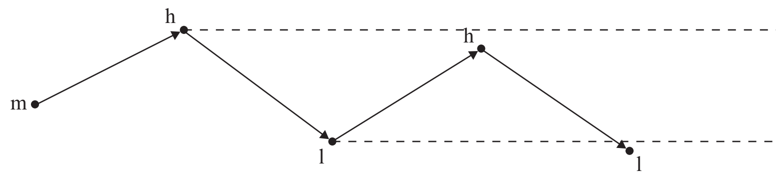


Figure 3.11
A graphic illustration of the fact that in this implementation, downdrift is an automatic by-product of the way in which the relative tones are defined.

Brazilian Portuguese, Venezuelan Spanish, Russian, Arabic, isiZulu, and Korean (for references, see Hirst 2007). More recently (Zhi, Hirst, and Bertinetto 2010), the algorithm was applied to a corpus of speech in Standard (Beijing) Chinese. This was particularly challenging, because the corpus used was spontaneous speech and involved a language with a rich lexical tone system. An attempt to optimize window size for the algorithm showed no overall improvement with respect to the manually corrected data. This was taken to confirm the fact (as suggested by Xu and Sun 2002) that pitch change in a lexical tone language like Chinese is not notably faster than in languages with no lexical tone. The annotated data obtained during this application will constitute a useful yardstick for evaluating improvements to the automatic algorithm, which is expected to be far more robust than data annotated for languages with no lexical tone.

3.6.3 Mapping from Momel to INTSINT

Having defined a mapping from tonal segments (INTSINT) to pitch anchor points (Momel), the same model can be used to establish a reverse mapping from the anchor points to the tonal segments. As is generally the case, such a reverse mapping from continuous variables to discrete categories is much less straightforward than the mapping from categories to continuous variables. The approach we have adopted is an exhaustive search of the target space for the optimal values of the two parameters key and span, together with the optimal coding of the sequence of anchor points, given those two parameters.

The procedure, as described in Hirst (2005), has been implemented as a Perl script. It assumes that the relevant target space is defined as follows:

$$\begin{aligned} \text{key} &= \text{mean} \pm 50 \text{ Hz (step: 1)} \\ \text{span} &= 0.5 \ldots 2.5 \text{ octaves (step: 0.1)} \end{aligned} \tag{3.7}$$

The script thus tries each of the possible values of the two parameters within this target space. For each of the two thousand possible couples ⟨*key;span*⟩ the script evaluates every possible coding of the anchor points using the formulas in equations (3.4), (3.5), and (3.6) and calculating the sum of the square of differences between the predicted value and the observed value. The output of the script is thus the optimized value (within the target range) of the parameters *key* and *span*, together with the optimal INTSINT coding using these parameters. The output of the script is a text file such as in (20), corresponding to the application of the script to the same extract from recording A01 that was shown in figure 3.5 above and elsewhere.

(20) ; A01_01.intsint created on Tue Aug 24 08:12:47 2010
; by intsint.pl 2.11 ; from A01_01.momel
; 32 values mean = 191 <parameter span = 1.4> <parameter key = 235>

0.113	M	221	235
0.219	D	205	208
0.434	D	182	190
0.746	B	120	145
1.177	S	120	145
1.423	T	428	382
1.623	B	146	145
1.894	U	197	184

. . .

In this output, the optimized values of key and span (here, 235 Hz and 1.4 octaves, respectively) are given, together with the sequence of anchor points. After the values for key and span, each line gives the time value (in seconds), the optimized INTSINT code, the original anchor point used as input, and the predicted anchor point derived from the coding with the current values of key and span.

Unlike with Momel, there is some loss of information with the INTSINT coding (as can be seen from the differences between the third and fourth columns of the output in (20); the model is still, however, a reversible one that can be used for synthesis. Figure 3.12 shows the result of recalculating the pitch anchor points from the results of the automatic INTSINT analysis, applied to a five-sentence passage from the Eurom1 corpus, and compared to the original anchor points obtained from the application of the Momel algorithm.

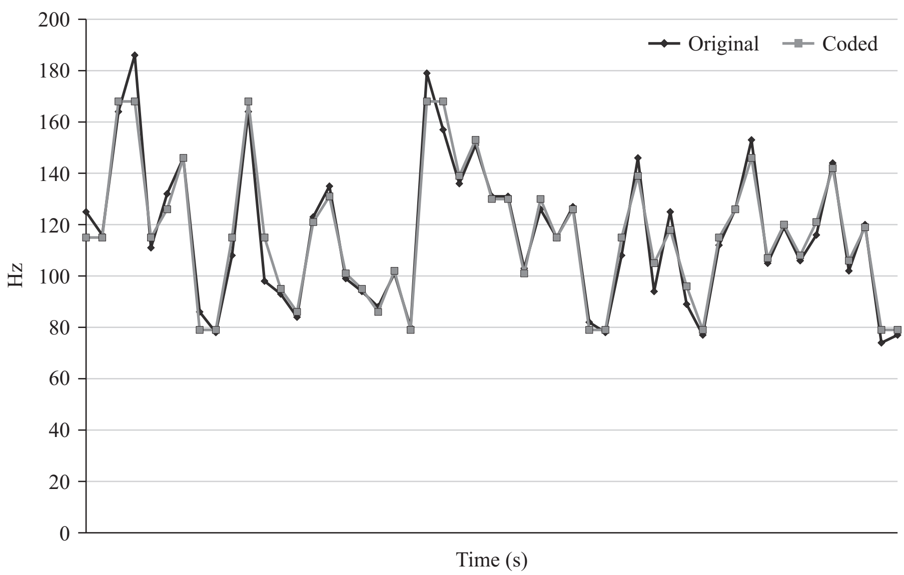


Figure 3.12
Coded (grey squares) versus original values (black diamonds) of pitch anchor points for a five-sentence passage after automatic coding with the INTSINT alphabet.

3.6.4 Longer-Term Characteristics of Pitch Range

The implementation of INTSINT as I have described presupposes that there are no variations in key and span within the segment of speech that is analyzed. In authentic speech, such changes naturally occur quite frequently and are obviously very significant and important. One solution is to implement the INTSINT coding on smaller segments of speech such as breath groups, making what seems a fairly reasonable assumption that changes of key and/or span are more likely to occur between breath groups rather than within them. For an investigation of the possibility of automatizing such a process, see De Looze (2010), which implements an algorithm (AdoReVa) applying a cluster analysis for this task.

3.7 ProZed

The availability of explicit models of speech melody such as those I have described in this chapter makes it possible to use such models to implement more abstract representations of prosody. With this aim in mind, my current work in progress concerns the implementation of a prosody editor for linguists called *ProZed* (Hirst 2015), which aims to provide linguists with a tool allowing them to experiment with different abstract phonological models, providing them with an acoustic output with which they can, at least informally, evaluate the relative value of different models.

A version of this editor applied to speech rhythm is described in Hirst and Auran (2005), which implemented an empirical linear model for rhythm where each rhythm unit is characterized by three parameters: q, a (possibly speaker-dependent) parameter defining a unit of quantal lengthening; t, a long-term parameter of tempo; and k, a local scalar effect of lengthening specific to each rhythm unit. With these parameters the duration of the rhythm unit is defined as:

$$\hat{d}_p = t \cdot \left\{ \sum_{i=1}^{m} \bar{d}_{i/p} + k \cdot q \right\} \tag{3.8}$$

where di/p is the mean duration of all the phones labeled as the same phoneme p as that occurring in position i in the rhythm unit.

ProZed has been implemented as a plug-in to the Praat software. It allows the manipulation of the rhythmic and tonal aspects of speech as defined on three specific tiers, in addition to the phoneme tier. The first two of these are the RU (rhythm unit) tier and the TU (tonal unit) tier. These two tiers control the short-term variability of segmental timing and tonal variation, respectively. Longer-term variations of rhythm and melody can be controlled via a third tier, the IU (intonation unit) tier.

The speech input to the program may be natural recorded speech, the prosodic characteristics of which will then be modified by the software, or, alternatively, it may be the output of a speech synthesis system with, for example, fixed durations for each speech segment.

The current version of the program is designed as the resynthesis step of what is planned to be a complete analysis by synthesis cycle. This will be directly integrated with the output of the Momel-INTSINT and ProZed rhythm analysis models and interfaced with the automatic alignment system SPPAS (Speech Phonetization Alignment and Syllabification; Bigi and Hirst 2012, 2013).

3.8 Underlying Phonological Representation

We have seen that it is possible to derive automatically from the acoustic signal a phonetic representation of an intonation contour (Momel) and that this can then be converted automatically into a time-aligned sequence of discrete symbols (INTSINT), which, together with the output of the automatic alignment software, can constitute at least a first approximation to a surface phonological representation of the prosody of an utterance. It now remains to be seen how we can bridge the gap between the abstract functional representation of prosody outlined in section 3.3 and this surface phonological representation.

For British English, the classical description of the intonation of nonemphatic declarative utterances and *wh*-questions, as in Armstrong and Ward (1926), Kingdon (1958), and O'Connor and Arnold (1961), is of a pitch pattern that begins on a mid-level with a rise to a high pitch on the first accented syllable, followed by a step-like lowering of pitch on each accented syllable until the last accented syllable, followed by a fall to the bottom of the speaker's pitch range.

Among the examples of this "tune," Armstong and Ward (1926) give:

(21) It was the 'last 'thing I ex- 'pected to 'find there.

In terms of prosodic structure, as described in section 3.4, this utterance can be parsed into one intonation unit containing six tonal units:

(22) [[^ it was the] [last] [thing I ex-] [-pected to] [find there]]

The INTSINT annotation of the tonal representation of this utterance (ignoring here the question of the relative alignment of the tones and the segmental material) could then be:

(23) [**m** It was the [**h** last] [**d** thing I ex-] [**d** -pected to] [**d** find there **b**] **b**]

where the initial [m] and final [b] tones are attached directly to the intonation unit and the other tones are attached to the tonal units.

Pike (1945) describes a neutral intonation pattern for declarative utterances in American English as a sequence of falling contours associated with each accented syllable. This could be annotated as:

(24) [**m** it was the [**h** last **l**] [**h** thing I ex- **l**] [**h** -pected to **l**] [**h** find there **b**] **b**]

Pike characterizes a "descending stress series" such as in (23) as expressing "EXTREME precision, or certainty" (70). Because of this difference in interpretation, to American ears the British intonation may sound overprecise or pompous, while to British ears the American pattern may sound overenthusiastic or exuberant.

This dialectal difference between American and British English intonation patterns is interesting since a downstepped tone in many African tone languages can often be traced to a historical low tone, which is sometimes described as a *floating* tone (Clements and Ford 1979), that is, a tone that is not phonetically realized as a pitch target but has the effect of lowering the following high tone.

The two dialects could consequently be described as having the same underlying representation, with the difference that in British English, the low tone is somehow "delinked" from the prosodic structure.

We can compare this intonation to that of a French translation of (22):

(25) C'était la 'dernière 'chose que je m'atten-'dais à 'trouver 'là.

We are likely to find that instead of a sequence of falling pitch movements as in American English (as described by Pike 1945) or a sequence of downsteps as in British English, the French utterance is produced as a sequence of rising pitch movements, culminating with the accented syllable and with a final falling pitch on the last syllable (Hirst and Di Cristo 1984; Di Cristo 1998).

We have suggested (Hirst and Di Cristo 1984) that one of the major prosodic differences between English and French is that in English, tonal units consist of an accented syllable *followed* by a number of unaccented syllables, and in French, they consist of an accented syllable *preceded* by a number of unaccented syllables. Example (25) could consequently be annotated in the following way:[6]

(26) [[c'était la der-] [-nière chose] [que je m'attendais] [à trou-] [-ver là]].

with which we can associate an INTSINT representation:

(27) [**m** [**m** c'était la der- **h**] [**l** -nière chose **h**] [**l** que je m'attendais **h**] [**l** à trou- **h**] [**d** -ver là] **b**]

For an underlying phonological representation of these tunes, I will assume, following Pierrehumbert (1980), that there are only two values for underlying tones, L and H, and that the specific INTSINT interpretations of these underlying tones are allophonic variants determined by the context. Under this interpretation, it seems clear, then, that for all three varieties we have looked at—American English, British English, and French—the underlying tones associated directly with the intonation unit (= boundary tones) are [L L].

We can assume, then, a template like the following to account for this association:

(28)
```
  IU
  / \
 L   L
```

For the tonal units, we can assume that both British English patterns and American English patterns are derived from a single template:

(29)
```
  τ
  / \
 H   L
```

and that in British English, but not in American English, there is a downstepping rule that will have the effect of interpreting an underlying sequence [H L] [H . . .] as [h] [d . . .]

French patterns can be derived from:

(30)
```
   τ
  / \
 L   H
```

assuming simply that in French, a downstepping rule applies only to the last tonal unit, interpreting [. . . H] [L H] as [. . . h] [d].

For the intonation of yes/no questions in English and French, the only change that needs to be made to the underlying representation for statements is that the final L is replaced by H, triggered by the presence of the question particle [Q] so that we have:

(31)
```
   IU
  / \
 L   H
```

For French, the presence of a question particle triggers not only the application of the template in (31) but also the application of a downstepping rule on each accent so that a sequence [. . . H] [L H] is interpreted as [. . . h] [d] for each tonal unit in the intonation unit.

For English, the presence of an emphatic particle [E] has two effects. The pitch accents on the accented syllables are either deleted entirely or considerably reduced, as if there were an intermediate prosodic constituent between the tonal unit and the intonation unit:

(32) a. [[John has] [two] [wives]]. (*John has two wives*)
b. [[[John has] [two] [wives]]] (*JOHN has two wives*)
c. [[John has] [[two] [wives]]] (*John has TWO wives*)
d. [[John has] [two] [[wives]]] (*John has two WIVES*)

There might, then, be a condition that only the first tonal unit within this emphatic constituent is assigned tones. This cannot be the whole story, though, because the last example of (32) would then be assigned the same tonal representation as the first, unemphatic, example. In fact, the pitch accent on an emphatic constituent is systematically higher than other accents. One possibility, then, would be to assume that the emphatic constituent is assigned [H L] tones just like the tonal units, so the underlying representation of (32) would then be something like:

(33) a. [**L** [**H** John has **L**] [**H** two **L**] [**H** wives **L**] **L**] (=John has two wives)
b. [**L** [**H** [**H** John has **L**] [two] [wives] **L**] **L**] (=JOHN has two wives)
c. [**L** [**H** John has **L**] [**H** [**H** two **L**] [wives] **L**] **L**] (=John has TWO wives)
d. [**L** [**H** John has **L**] [**H** two **L**] [**H** [**H** wives **L**] **L**] **L**] (=John has two WIVES)

which could then be interpreted as the INTSINT annotations (for British English), where the **h** or **d** followed immediately by **h** would ensure a higher pitch on the following syllable than when there is just a single **h**:

(34) a. [**m** [**h** John has] [**d** two] [**d** wives **b**] **b**] (=John has two wives)
b. [**m** [**h** [**h** John has **b**] [two] [wives] **b**] **b**] (=JOHN has two wives)
c. [**m** [**h** John has] [**d** [**h** two **b**] [wives] **b**] **b**] (=John has TWO wives)
d. [**m** [**h** John has] [**d** two] [**d** [**h** wives **b**] **b**] **b**] (=John has two WIVES)

3.9 Conclusion

I have presented in this paper a sketch of a fairly complete model of prosody, although a great number of details concerning the implementation of the model need to be specified. The model extends from a functional annotation of intonation at the most abstract level, via an underlying phonological representation and a surface phonological representation, to a phonetic representation, which is directly convertible into an acoustic signal. The model is described in more detail in my book (Hirst 2020), together with more detailed justification of the approach presented here. It is obvious that at each level, I have made a number of arbitrary choices, which may not stand up to further investigation, but I am nonetheless convinced that a multilevel, multilingual model of this type is what is needed to further our knowledge of the complex way in which prosody contributes to our interpretation of utterances.

Notes

1. Faure (1962) makes a similar distinction between "les caractères" (forms) and "le rôle" (functions).

2. I would today prefer to call this [±ACCENT] rather than [±STRESS].

3. From the first recording (A01) of the Aix-Marsec corpus (Auran, Bouzon, and Hirst 2004).

4. The statistics used here were *recall* and *precision*, which are commonly used in information retrieval and binary classification tasks. *Recall* is the proportion of positive values that are correctly identified, and *precision* is the proportion of values identified as positive that are correct. In our context, precision is the percentage of all automatically detected targets that were considered correct, while recall is the percentage of all "correct" targets that were automatically detected. The *F-measure*, considered a measure of global efficiency, is a combination of the two values and is calculated as the harmonic mean of recall and precision (van Rijsbergen 1979).

5. In some earlier publications, I have used the word *range* for what I here call *span*. I now prefer to use *span* to refer to the interval, independent from the value of the *key*. The two values *key* and *span* together define the speaker's *range*. So we might say that a given speaker has a pitch range from 100 Hz to 200 Hz, corresponding to a span of one octave and a key of 141 Hz.

6. French words are regularly accented on the final syllable, but nonfinal words may also be accented on the first syllable, especially when immediately followed by an accented syllable as in *dernière chose* and *trouver là* in this example. See Hirst and Di Cristo (1984) and Di Cristo (1998) for details.

References

Abercrombie, D. 1964. "Syllable Quantity and Enclitics in English." In *In Honour of Daniel Jones*, edited by D. Abercrombie, D. Fry, P. MacCarthy, N. Scott, and J. Trim, 216–222. London: Longman.

Armstrong, L., and I. Ward. 1926. *A Handbook of English Intonation*. Cambridge: Heffer.

Auran, C., C. Bouzon, and D. J. Hirst. 2004. "The Aix-MARSEC Project: An Evolutive Database of Spoken British English." In *Speech Prosody 2004*, International Conference; Nara, Japan, March 23–26, 2004, edited by Bernard Bel and Isabelle Marlien; ISCA Archive. http://www.isca-speech.org/archive/sp2004.

Beckman, M., and J. Pierrehumbert. 1986. "Intonational Structure in Japanese and English." *Phonology Yearbook* 3:255–309.

Bigi, B., and D. Hirst. 2013. “What’s New in SPPAS 1.5?” In *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, edited by Brigitte Bigi and Daniel Hirst, 62–65. Aix-en-Provence, France.

Bigi, B., and D. J. Hirst. 2012. “Speech Phonetization Alignment and Syllabification (SPPAS): A Tool for the Automatic Analysis of Speech Prosody.” In *Speech Prosody 2012*, Sixth International Conference, Shanghai, China, May 22–25, 2012, edited by Qiuwu Ma, Hongwei Ding, and Daniel Hirst; ISCA Archive. http://www.isca-speech.org/archive/sp2012.

Bird, C. S., J. P. Hutchinson, and M. Kante. 1977. *Beginning Bambara (An Ka Ba-manankan Kalan)*. Bloomington: Indiana University Press.

Boersma, P., and D. Weenink. 1992 (2015). *Praat, a System for Doing Phonetics by Computer*. Version 5.4.06. Computer program. http://www.praat.org.

Bolinger, D. 1958. “A Theory of Pitch Accent in English.” *Word* 14:119–149.

Campione, E. 2001. “Etiquetage Semi-Automatique de la Prosodie dans les Corpus Oraux—algorithmes et Méthodologies.” PhD diss., Université de Provence.

Chentir, A., M. Guerti, and D. Hirst. 2009. “Extraction of Standard Arabic Micromelody.” *Journal of Computer Science* 5:86–89.

Cho, H., and S. Rauzy. 2008. “Phonetic Pitch Movements of Accentual Phrases in Korean Read Speech.” In *Speech Prosody 2008*, Fourth International Conference, Campinas, Brazil, May 6–9, 2008, edited by Plinio A. Barbosa, Sandra Madureira, and Cesar Reis; ISCA Archive. http://www.isca-speech.org/archive/sp2008.

Chomsky, N., and M. Halle. 1968. *The Sound Pattern of English*. New York: Harper.

Clements, G., and K. Ford. 1979. “Kikuyu Tone Shift and Its Synchronic Consequences.” *Linguistic Inquiry* 10:179–210.

Couper-Kuhlen, E. 1986. *An Introduction to English Prosody*. London: Edward Arnold.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

De Looze, C. 2010. “Analyse et Interprétation de L’empan Temporel des variations Prosodiques en Français et en Anglais.” PhD diss., Université de Provence Aix-en-Provence.

Di Cristo, A. 1998. “Intonation in French.” In *Intonation Systems*: *A Survey of Twenty Languages*, edited by D. J. Hirst, and A. Di Cristo, 195–218. Cambridge: Cambridge University Press.

Di Cristo, A., and D. J. Hirst. 1986. “Modelling French Micromelody: Analysis and Synthesis.” *Phonetica* 43 (1–3): 11–30.

Faure, G. 1962. *Recherches sur les Caractères et le Rôle des Éléments Musicaux dans la Prononciation Anglaise*. Paris: Didier.

Fujisaki, H. 1991. “Modeling the Generation Process of F0 Contours as Manifestation of Linguistic and Paralinguistic Information.” In *Proceedings of the Twelfth International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs/icphs1991.

Fujisaki, H. 2004. “Information, Prosody and Modeling—with Emphasis on Tonal Features of Speech.” In *Speech Prosody 2004*, International Conference; Nara, Japan, March 23–26, 2004, edited by Bernard Bel and Isabelle Marlien, ISCA Archive. http://www.isca-speech.org/archive/sp2004.

Gårding, E. 1998. “Intonation in Swedish.” In *Intonation Systems*: *A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 117–136. Cambridge: Cambridge University Press.

Goldsmith, J. A. 1990. *Autosegmental and Metrical Phonology*. Cambridge, MA: Blackwell.

Halliday, M. 1967. *Intonation and Grammar in British English*. Berlin: Mouton.

Hirst, D. J. 1974. "La Levée de L'ambiguïté Syntaxique par les Traits Intonatifs: Essai de Formalisation." PhD diss., Université de Provence.

Hirst, D. J. 1977. *Intonative Features*: *A Syntactic Approach to English Intonation*. Berlin: Mouton.

Hirst, D. J. 1983. "Interpreting Intonation: A Modular Approach." *Journal of Semantics* 2:171–181.

Hirst, D. J. 1987. "La Représentation Linguistique des Systèmes Prosodiques: Une Approche Cognitive." Thèse de Doctorat d'Etat (Habilitation Thesis), Université de Provence.

Hirst, D. J. 1998. "Intonation in British English." In *Intonation Systems*: *A Survey of Twenty Languages*. Edited by D. J. Hirst and A. Di Cristo, 56–77. Cambridge: Cambridge University Press.

Hirst, D. J. 2005. "Form and Function in the Representation of Speech Prosody." *Speech Communication* 46:334–347.

Hirst, D. J. 2007. "A Praat Plugin for Momel and INTSINT with Improved Algorithms for Modelling and Coding Intonation." In *Proceedings of the Sixteenth International Conference of Phonetic Sciences*, edited by Jürgen Trouvain and William J. Barry, 1233–1236. Saarbrücken, Germany.

Hirst, D. J. 2011. "The Analysis by Synthesis of Speech Melody: From Data to Models." *Journal of Speech Sciences* 1:55–83.

Hirst, D. J. 2012. "Empirical Models of Tone, Rhythm and Intonation for the Analysis of Speech Prosody." In *Rhythm*, *Melody and Harmony in Speech*: *Studies in Honour of Wiktor Jassem*, edited by D. Gibbon, D. J. Hirst, and N. Campbell, 23–33. Poznan, Poland: Polish Phonetic Association.

Hirst, D. J. 2015. "ProZed: A Speech Prosody Editor for Linguists, Using Analysis-by-Synthesis." In *Speech Prosody in Speech Synthesis*: *Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis*, edited by K. Hirose and J. Tao, 3–17. Berlin: Springer-Verlag.

Hirst, D. J. 2020. *Speech Prosody*: *From Acoustics to Interpretation*. Berlin: Springer-Verlag.

Hirst, D. J., and C. Auran. 2005. "Analysis by Synthesis of Speech Prosody: The Prozed Environment." In *Interspeech'2005—Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4–8; 200 ISCA Archive. 3225–3228. Lisbon. http://www.isca-speech.org/archive/interspeech_2005.

Hirst, D. J., and C. Bouzon. 2005. "The Effect of Stress and Boundaries on Segmental Duration in a Corpus of Authentic Speech (British English)." In *Interspeech'2005—Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4–8, 2005; ISCA Archive, 29–32. Lisbon. http://www.isca-speech.org/archive/interspeech_2005.

Hirst, D. J., H. Cho, S. Kim, and H. Yu. 2007. "Evaluating Two Versions of the Momel Pitch Modeling Algorithm on a Corpus of Read Speech in Korean." In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association,* Antwerp, Belgium, August 27–31, 2007; ISCA Archive, 1649–1652. Antwerp, Belgium. http://www.isca-speech.org/archive/interspeech_2007.

Hirst, D. J., and A. Di Cristo. 1984. "French Intonation: A Parametric Approach." *Die Neueren Sprachen* 83:554–569.

Hirst, D. J., and A. Di Cristo. 1998. *Intonation Systems*: *A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Hirst, D. J., A. Di Cristo, and R. Espesser. 2000. "Levels of Representation and Levels of Analysis for the Description of Intonation Systems." In *Prosody*: *Theory and Experiment*: *Studies Presented to Gösta Bruce*, edited by M. Horne, 51–58. Amsterdam: Kluwer.

Hirst, D. J., and R. Espesser, R. 1993. "Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function." *Travaux de l'Institut de Phonétique d'Aix* 15:75–85.

Iivonen, A. 1998. "Intonation in Finnish." In *Intonation Systems*: *A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 331–347. Cambridge: Cambridge University Press.

Jassem, W. 1952. *Intonation of Conversational English (Educated Southern British)*. Wroclaw, Poland: Wroclawskie Towarzystwo Naukow.

Jun, S.-A. 2005. *Prosodic Typology*: *The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Kingdon, R. 1958. *The Groundwork of English Intonation*. London: Longmans.

Ladd, D. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Liberman, M., and A. Prince. 1977. "On Stress and Linguistic Rhythm." *Linguistic Inquiry* 8:249–336.

Maghbouleh, A. 1998. "Tobi Accent Type Recognition." In *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference*, Sydney Convention Centre, Sydney, Australia, 30th November–4th December 1998; ISCA Archive, Paper 0632. http://www.isca-speech.org/archive/icslp_1998.

Mixdorff, H.-J. 1999. "A Novel Approach to the Fully Automated Extraction of Fujisaki Model Parameters." In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99*, Phoenix, Arizona, USA, March 15–19, 1999. IEEE Computer Society.

Nespor, M., and I. Vogel. 1986. *Prosodic Phonology*. Dordrecht, the Netherlands: Foris Publications.

Nooteboom, S. 1997. "The Prosody of Speech: Melody and Rhythm." In *The Handbook of Phonetic Sciences*, edited by W. Hardcastle and J. Laver, 640–673. Oxford: Blackwell.

O'Connor, J., and G. Arnold. 1961. *Intonation of Colloquial English: A Practical Handbook*. London: Longmans.

Palmer, H. E. 1924. *English Intonation with Systematic Exercises*. Cambridge: Heffer.

Pierrehumbert, J. B. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pike, K. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Rijsbergen, C. J. van. 1979. *Information Retrieval*. 2nd ed. London: Butterworth.

Selkirk, E. 1984. *Phonology and Syntax*: *The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. "TOBI: A Standard for Labeling English Prosody." In *Proceedings of the Second International Conference on Spoken Language Processing*, 867–870. Banff, Canada: ISCA.

Taylor, P. 1994. "The Rise/Fall/Connection Model of Intonation." *Speech Communication* 15:169–186.

Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

't Hart, J., R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation*: *An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Trubetzkoy, N. S. 1949. *Grundzüge der Phonologie*. Translated by J. Cantineau (1957) as *Principes de phonologie*. Paris: Klincksieck.

Véronis, J., D. Hirst, and N. Ide. 1994. "NL and Speech in the MULTEXT Project." In *Proceedings of AAAI Workshop on Integration of Natural Language and Speech*, 72–78. Seattle.

Wakefield, J. 2009. "The English Equivalents of Cantonese Sentence-Final Particles: A Contrastive Analysis." PhD diss., Hong Kong Polytechnic University.

Wakefield, J. 2012. "A Floating Tone Discourse Morpheme: The English Equivalent of Cantonese lo1." *Lingua* 122:1739–1762.

Wells, J. C. 2006. *English Intonation*: *An Introduction*. Cambridge: Cambridge University Press.

Wightman, C. 2002. "ToBI or Not ToBI"? In *Speech Prosody 2002, International Conference*; Aix-en-Provence, France, April 11–13, 2002; ISCA Archive, 21–29. Aix-en-Provence. http://www.isca-speech.org/archive/sp2002.

Wightman, C., and N. Campbell. 1995. "Improved Labeling of Prosodic Structure. In *Proceedings of IEEE Transactions on Speech and Audio Processing*. Piscataway, NJ: IEEE.

Xu, Y., and X. Sun. 2002. "Maximum Speed of Pitch Change and How It May Relate to Speech." *Journal of the Acoustical Society of America* 111:1399–1413.

Xu, Y., and Q. E. Wang. 2001. "Pitch Targets and Their Realization: Evidence from Mandarin Chinese." *Speech Communication*, 33:165–181.

Zhi, N., D. Hirst, and P. M. Bertinetto. 2010. "Automatic Analysis of the Intonation of a Tone Language. Applying the Momel Algorithm to Spontaneous Standard Chinese (Beijing)." In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Japan, September 26–30, 2010, edited by Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura; ISCA Archive. http://www.isca-speech.org/archive/interspeech_2010.