X  Intonation

MARY E. BECKMAN
JENNIFER J. VENDITTI

1  The scope of this chapter

As a technical term in phonological descriptions of spoken languages, *intonation* refers to patterned variation in voiced source pitch that serves to contrast and to organize words and larger utterances. In this general statement of its meaning, it is synonymous with the technical term *tone*. In typical usage however, the two terms are differentiated by applying them to different aspects of these linguistic uses of pitch, a differentiation that is reflected in this edition of the *Handbook of Phonological Theory* by the fact of there being a separate chapter on tone (Hyman, this volume). In order to delimit the scope of the present chapter, therefore, we begin by listing the aspects of the linguistic use of pitch that are typically invoked in differentiating intonation from tone.

The differentiation is exemplified by the two parts of the sixth definition for the entry for ‘tone’ in the *Concise Oxford English Dictionary* (11th edition): “(in some languages, such as Chinese) a particular pitch pattern on a syllable used to make semantic distinctions” and “(in some languages, such as English) intonation on a word or phrase used to add functional meaning.” This sixth definition is tagged as the meanings for a technical term in phonetics, and its second part subsumes the term ‘intonation’, which is defined in its own entry as “the rise and fall of the voice in speaking”. In the *COED* entry, then, the primary sense of tone as a technical term in describing sound patterns refers to a localized melodic event (a note or glissando) occurring over the span of a syllable, whereas tone *qua* intonation refers to a pattern of glissandi distributed over a longer span. Also, tone in this primary sense invokes a system of contrastive pitch patterns that act as minimal word-differentiating elements, comparable to the inventory of vowels or consonants of a language, whereas tone *qua* intonation invokes other functions, such as mirroring the syntactic structure of an utterance or indicating its pragmatic role in the larger discourse context. These two sets of contrasting characteristics make for a multidimensional taxonomy of phonetic form in relationship to linguistic function. This much is uncontroversial.

A third aspect of the *COED* definition is more controversial. The two parts of the definition refer to two different sets of languages, reflecting the claim in many broad-stroke surveys such as Hyman (2006) that particular values along the dimensions of form and function tend to coincide in ways that are conducive to a one-dimensional classification of language types, with “some languages, such as Chinese,” at one end and “some languages, such as English,” at the other. Careful descriptions of specific languages at every point along the purported continuum, on the other hand, typically use the terms together in ways that defy the typology. For example, in many descriptions of specific Chinese dialects, such as Chang (1958), “tone” is used to refer to the localized melodic events...
(notes or glissandi) that contrast one-syllable words in citation-form utterances, but “intonation” is also used: to designate notes and glissandi that occur at phrase edges (rather than on designated syllables) with functions other than that of lexical contrast (e.g. marking interrogative speech acts), and to refer to longer-term modulations of the implicit melodic scale that defines the relationship of contrast among different notes and between two different rising or two different falling glissandi. Conversely, in many accounts of the English intonation system, such as Halliday (1967), Goldsmith (1978), and Pierrehumbert (1980), “tone” is used to refer to glissandi or notes that are localized to linguistically significant positions, such as the stressed syllables of some words and the edges of phrases.

The reference to constituents such as “phrases” and “stressed syllables” in these descriptions invokes another technical term – *prosody* – which also occurs in many definitions of intonation, and vice versa. In these definitions, the two words are often treated as being in a hypernym-hyponym relationship whereby ‘prosody’ is the broader cover term that groups various aspects of the pitch pattern of an utterance together with a motley group of other phenomena that defy the assumption of “segmental idealization” – i.e., the assumption that “speech … can appropriately be idealized as a string of ordered discrete sound segments of unspecified duration” (Ladd, this volume). For example, the *COED* defines ‘prosody’ first as “the patterns of rhythm and sound used in poetry” and by extension as “the patterns of stress and intonation in a language”. In order to fully delimit the scope of this chapter, therefore, we need to lay out our assumptions about what prosody is, both in relationship to the pitch patterns of a language and in relationship to the rhythms of its vowel and consonant patterns.

In the rest of this chapter, then, we will elaborate on the various aspects of intonation that are invoked in these definitions of tone and of prosody. We begin by briefly laying out one current understanding of what prosody is, and by illustrating this understanding in terms of how one particular language’s word rhythms can be described using the representational device that is adopted in much current research on intonational phonology (Section 2). We then review the various ways in which phonologists and phoneticians have represented the pitch patterns that they observe in the laboratory and the field (Section 3). We go on to describe how research in phonology and phonetics over the decades since the development of such crucial analytic tools as the source-filter theory of vowel production (Fant 1960) has contributed to the evolution of a taxonomy of the forms (Section 4) and functions (Section 5) of spoken language melody. In these descriptions we will use many examples of how the taxonomy applies to varieties of Chinese and of Germanic (including English) in order to emphasize why we think it is not as useful to delimit intonation from tone in terms of sets of languages, a point to which we will return in our summary (Section 6).

## 2 Defining prosody

### 2.1 Bases for a definition

We define prosody as the set of syntagmatic relationships that hold among an utterance’s tone, vowel, and consonant specifications, as distinct from the paradigmatic relationships of similarity and contrast that hold between these specifications and any other tone, vowel, and consonant specifications that might stand in closely analogous syntagmatic relationships to each other in some other utterance. We suspect that in the course of acquiring a spoken language, these syntagmatic relationships become reified as an abstract structure projected away from the phonetic substance of the paradigmatic relationships, and that this projection then serves as “a rhythmic scaffolding that specifies designated temporal points of convergence and structural alignment among different
components of the grammar” (Arbisi-Kelm & Beckman 2009). The primary device that we adopt to describe this projected syntagmatic structure is the “prosodic tree” — a term first used by Nespor and Vogel (1979) to refer to a particular type of directed acyclic graph that satisfies a set of substantive constraints on the types of node that can occur at different distances from the root and on relationships of (temporal) order among nodes at the same distance (see Selkirk 1980 and chapter 6 of Pierrehumbert and Beckman 1988, inter alia).

The language that we will use to illustrate this device is the Tokyo dialect of Japanese, a language that has played a critical role in the development of computationally explicit compositional theories of the elements of intonation contours and their relationship to the prosodic organization of utterances. For example, Fujisaki and Sudo (1971) and Pierrehumbert and Beckman (1988) provide two different accounts of Japanese intonation patterns in which phrase-level pitch range is specified by continuous phonological control parameters that are independent of the categorical specification of more local tone shapes in the lexicon. This incorporation of continuous (or “paralinguistic”) specifications of phrasal pitch range directly into the phonological description was a critical element in the development of what Ladd (1996/2008) calls the “autosegmental-metrical” (AM) framework.

Ladd coined this term to refer to a class of models of intonational phonology that began to develop rapidly in the late 1970s and early 1980s, especially after Pierrehumbert’s (1980) incorporation of Bruce’s (1977) seminal insights about the composition of Swedish tone patterns into a grammar for English intonation contours. In the decades since, the AM framework has been widely adopted in descriptions of intonation systems and in comparisons of prosodic organization across a variety of languages. (See Jun 2005 for a recent collection of such descriptions and Gussenhoven 2004 for a review of work in this framework since Ladd 1996.) All AM models are deeply compositional, analyzing sound patterns into different types of elements at several different levels. The most fundamental level is the separation of material properties from the structural positions that license them. This is roughly comparable to the separation of lexis from phrase structure in a description of the rest of the language’s grammar.

2.2 Autosegmental content

More specifically, the “A” (or autosegmental) part of an AM description refers to the specification of content features that are autonomously segmented — that is, that project as strings of discrete categories specified on independent tiers rather than being bundled together at different positions in the word-forms that they contrast (see Goldsmith 1976; Hyman, this volume). For example, in Japanese, changing specifications of stricture degree for different articulators can define as many as six different manner autosegments within a disyllabic word-form, as in the word *sanpun* [sampun] ‘3 minutes’, where there is a sequence of stricture gestures for sibilant airflow, for vowel resonances, for nasal airflow, for a complete stopping of air, and then a vowel and a nasal again. However, there are constraints on which oral articulators can be involved in making these different airflow conditions, so that only one place feature can be specified for the middle two stricture specifications in a word of this shape. Therefore, in the grammar of Japanese (as in many other languages), place features can be projected onto a different autosegmental tier from the manner features. This projection allows for a very general description of the allomorphic alternation among labial nasal *[m]* in *sanpun*, dental nasal *[n]* in *sandan* ‘3 steps’, velar nasal *[ŋ]* in *sangai* ‘3rd floor’, and uvular nasal *[ɴ]* in the citation form *san* ‘3’. This kind of place assimilation for coda nasals applies without exception in the derivational and the inflectional
morphology of the language, and it leads to differences between first-language
speakers of Japanese and Korean in their ability to identify place of articulation
of coda nasals in English (Aoyama 2003).

A key insight of Firth (1957), which was developed and elaborated as the
basis of autosegmental phonology by North American interpreters such as
Goldsmith (1976) and Haraguchi (1977) is that the consonant and vowel “units”
that are named by symbols such the [s], [a], [m], [p], [u], and [n] in sanpun are not
necessarily “the basic constituent of a linguistic system” (see Pike 1943, p. 42, as
discussed in Ladd, this volume). Indeed, several long-standing phonological
puzzles, such as “the non-uniqueness of phonemic solutions” (Chao 1934)
disappear when consonant and vowel segments are treated not as an automatic
reflection of a “natural phonetic segmentation” but as an epiphenomenon, a by-
product of the intersection of place and manner specifications for the word-form
when the array of parallel streams of autosegments on different tiers is collapsed
into a one-dimensional string. Thus, the “autosegmental insight” (Hyman, this
volume, p. 82) gradually led to a key insight of CV phonology (Clements and
Keyser 1983) and dependency phonology (Anderson 1986): the idea that “vowel”
and “consonant” are not merely cover terms for two broad sets of autosegmental
substance, but also the names (typically abbreviated as V and C) of two basic
types of syntagmatic relationship, which are analogous to the relationships of
head versus adjunct in the syntactic organization of words. These two
relationships can be represented as two types of leaf node in the prosodic tree
that depicts the “M” part of an AM description of a spoken utterance.

2.3 Metrical frames

This “M” (or metrical) part an AM description, then, specifies the hierarchy of
prosodic constituents that is the meter (rhythmic scaffolding) of an utterance
containing the word-form. For example, in Japanese, there is a low-level prosodic
constituent syllable (typically abbreviated σ) that coordinates the place and
manner features of words to alternate between more sonorant V manners that
are licensed to occur at the head and the less sonorant C manners that licensed to
occur only at the edges. In an utterance of sanpun, the syllables dominate a CV
level meter that is CVC.CVC. In many other words, there is a more regular
rhythmic alternation between the V and C leaf-node types, as illustrated by the
two utterances shown in Figure X.1. All but six of the 26 syllables in these
utterances dominate a perfectly alternating CV sequence.

The exceptional cases are of two types. The first is the bare V at the beginning
of the verbs oyoideru and oborete and of the auxiliary verb -iru in utterance (b) in
the figure. These three V-shape syllables, like the twenty CV-shape syllables, are
short. That is, each contains just one mora (abbreviated μ), a constituent between
the syllable and the CV leaf nodes that licenses the specification of vowel place
and height features at the V nodes within a syllable. The other three exceptional
syllables are long. A long syllable contains at least one extra adjunct mora after
the head V that is the sole obligatory leaf constituent. This following mora can be
either a V or a C. If it is a V, it can be the second part of a geminate vowel, as in
the particle yoo at the end of utterance (b), or it can be a less sonorous vowel than
the head vowel, as in the second syllable of oyoideru in the same utterance, where
the [i] has lesser sonority than the preceding head vowel [o]. If the following
mora is a C, it can be a moraic nasal, as in each of the two syllables of sanpun
discussed earlier, or it can be the first part of a geminate obstruent, as in the
second syllable of the verb hashitteru of utterance (a) in Figure X.1.

As sanpun shows, a long syllable that is closed by the moraic nasal can be
word final. By contrast, a long syllable that is closed by an obstruent must have a
following syllable to provide the second C position for the necessarily geminate
consonant. This constraint is one of the motivations for positing prosodic constituents above the syllable as well as below. That is, the place and manner autosegments for a geminate consonant necessarily associate to a sequence of C nodes that is medial to a prosodic word (abbreviated $\omega$).

In (1), we show a prosodic tree representation of the $\omega$ and $\omega$-dominated constituents for each of the utterances that we have represented phonetically in the four panels of Figure X.1. Below each tree we transcribe the sequence of consonant and vowel segments that results when the array of place and manner autosegments is collapsed into a one-dimensional string. The spectrograms in Figure X.1 provide one kind of (partial) phonetic representation for these place and manner autosegments. Below the transcription of the consonant and vowel segments, we transcribe the ten tonal autosegments that are specified to occur around the C and V constituents of lexically designated syllables in five of the words that occur in the two utterances. The $f_0$ contours in Figure X.1 provide one kind of (partial) phonetic representation for these ten autosegments and several other tonal autosegments that we will discuss at more length in Section 4.1.

Figure X.1 Spectrograms and $f_0$ contours for utterances of (a) Kashino-ga hashitteru. ‘Kashino is running.’ and (b) Yamano-wa oyoideru ga, marude oborete-iru yoo da. ‘Yamano is swimming, but he’s nearly drowning right now.’ [The utterances are from Venditti, Maekawa, and Beckman 2009.]
2.4 Prosodic trees

First, though, we want to bring out more clearly the relationship between the metrical frames that are represented symbolically in (1) and the autosegmental content features that are represented parametrically in Figure X.1. To do this, we return to the analogy that drives the choice of the representational device in (1). The hierarchy of prosodic constituents in the phonological description of an utterance is analogous to the hierarchy of syntactic constituents in its syntactic description. By this analogy, the prosodic constraint that each syllable must dominate a V (vowel) category phone that is positioned to be its head is like the syntactic constraint that each verb phrase must have a V (verb) category word positioned to be its head. It is in this sense that we say that the inventory of segments for a language is analogous to the lexis of the language.

One argument in favor of this analogy is that the prosodic category of a phone such as [i] in Japanese is inherently ambiguous in the same way that the syntactic category of a word such as *dam* in English is ambiguous. In different morphosyntactic contexts, *dam* can be parsed either as the verb meaning ‘to build an obstruction across a stream or river’ or as the noun meaning the resulting obstruction. Similarly, the palatal constriction that gives rise to the low first formant and high second formant on either side of the medial [o] in the word *oyoideru* in spectrogram (b) in Figure X.1 can be parsed either in terms of a C constituent that is directly dominated by σ (where it is traditionally transcribed as [j]) or in terms of a V constituent that is dominated by the adjunct second μ (where it is traditionally transcribed as [i]).

A second argument in favor of this analogy is the way that empty structural positions are interpreted. That is, the implicit but unrealized subject noun phrase in the second clause of utterance (b) in Figure X.1 can be recovered from the syntactic parse of the rest of the clause. Similarly, the implicit head vowel in the second syllable of *hashitteru* in utterance (a) can be recovered from the prosodic parse of the rest of the ω even though there is no interval containing the usual acoustic traces of the vowel manner specification in the signal. This token of the segment is “devoiced” and so cannot show the vowel formant values that cue the explicit [i] in *oyoideru*. Nonetheless, the listener parses the presence of some
implicit head V constituent for the second σ of hashitteru from the fact that [ft:] is not a grammatical sequence of phones for any position in a prosodic word of the language. Moreover, the listener parses an associated sequence of high (H) and low (L) tonal autosegments here, even though there is no periodicity to give rise to a sense of pitch on this lexically “accented” syllable, because the melody over this whole sentence is “underlyingly” (i.e., prosodically) identical to the melody of the Yamano-ga oyoideru clause in utterance (b) in this figure.

A good grasp of how listeners resolve prosodic ambiguities and of how they recover the intended autosegmental features of an implicit prosodic node in phonological ellipsis is critical for understanding the ways in which prosodic structures above the word are manipulated to fill the functions that we associate with intonation as distinct from tone. In introducing the prosodic constituents C, V, µ, σ, and ω in this section, we have defined them primarily in terms of the distribution of place and manner autosegments for consonant and vowel phones. In describing what listeners do to parse the prosodic categories of ambiguous segments or to recover elided segments, we have made passing reference to the autosegmental content of these structures, using the spectrograms in Figure X.1 as a convenient representational device for the place and manner autosegments.

The two other panels in the figure show a convenient representational device for the ten tonal autosegments that differentiate words such as Kashino and oyoideru from words such as marude, a representational device that we discuss next.

3 The representation of tone and intonation

3.1 Phonetic representations

Because of the well-behaved psychophysical relationship between the percept of pitch and the fundamental frequency (f₀) of the periodic voice source, the choice among phonetic measures that one could use to represent tone and intonation patterns is fairly straightforward. Fry (1968) describes several early methods for estimating f₀ values over shorter or longer stretches of speech. These included measuring the durations of successive periods identified in oscillographic records (as in Denes and Milton-Williams, 1962) and tracing the frequency of some higher harmonic visible in narrow-band spectrograms (as in Lehiste and Ivić, 1963). More recently, the ready availability of computer programs for estimating f₀ using autocorrelation-based algorithms in free signal-analysis packages such as Praat (Boersma, 1993) and WaveSurfer (Sjölander and Beskow, 2000) has made the f₀ contour of a recorded utterance an even more obvious choice as a first-pass phonetic representation of its intonation or tone pattern. The obviousness of this choice is reflected in the name by which the f₀ contour often is called. In both Praat and WaveSurfer, the f₀ estimates that are returned by the autocorrelation algorithm are called “pitch” values, and many phonologists and phoneticians use the phrase “pitch track” in referring to a time plot of a sequence of estimated f₀ values over some interval of recorded speech.

It is important to remember, however, that the f₀ contour is only a very rough first-pass phonetic representation of the melodic pattern of an utterance, for at least three reasons. The first is that f₀ is not reliably estimated in stretches of speech where less regular source qualities such as creaky voice are in play. The failure of standard algorithms for estimating f₀ in such regions makes the “pitch track” a poor phonetic representation for melodic events that harness such a “non-modal” voice quality as a cue, as illustrated in Figure X.2. The f₀ contours in the middle panels of the figure were calculated over utterances of two sentences of Putonghua (PRC Standard Chinese) produced by an adult female speaker from Songyuan City in Jilin Province. The creaky voice in the third syllable of
each utterance is a cue to the very low pitch target that characterizes this tone, as shown by Gårding, Kratochvil, Svantesson, and Zhang (1986), among others.

![Spectrograms](image)

**Figure X.2** Spectrograms and $f_0$ contours for utterances of (a) Lào Wáng mǎi ròu. [lau$^{21}$ wan$^{35}$ mai$^{21}$ zo$^{51}$] ‘Old Wang buys meat.’ and (b) Lào Huáng tǎo fàn. [lau$^{21}$ xwa$^{35}$ tao$^{21}$ fan$^{51}$] ‘Old Huang begs for food.’ with a close-copy stylization (solid line) overlaid on each $f_0$ contour in the bottom row of panels.

The second reason is essentially the same as the first, but applies to regions where the $f_0$ is well-defined. When we see the concentrations of energy in a 600-800 Hz band at 0.27 seconds in each of the spectrograms in Figure X.2, we can read this setting of the filter resonances in terms of the combined labial and dorsovelar constriction gestures for the initial [w] of the surname Wáng and for its devoiced allophone after the initial [x] in the surname Huáng. By contrast, when we see the $f_0$ value of 190 Hz at both 0.13 and 0.41 seconds in the right-hand utterance (i.e. at points midway through the [a] nucleus in the two vowels before and after the [xw] of Huáng), we cannot know what combination of pulmonary and laryngeal gestures produced this setting of the glottal source period. There is a perceptibly lower tone target for the vowel at 0.13 seconds as compared to the target for the vowel at 0.41 seconds, and this percept of different targets matches transcriptions using Chao’s (1930) tone-letters. Typically, the tone in Lào in this phrase-medial context is transcribed as a drop or as a dipping down to the bottom of the tonal space (i.e. [lau$^{21}$]~[lau$^{214}$]), whereas the tone in Wáng is transcribed as a rise from a middle region (i.e. [wan$^{35}$]).$^2$ Further support for positing such a difference in the target minima for these two tones comes from electromyographic studies (Sagart, Hallé, de Boysson-Bardies, and Arabia-Guidet 1986; Hallé 1994), which show consistently high activity of the sternohyoid for the very low-pitched target in the low or dipping tone of Lào but
less reliable involvement of this muscle at the beginning of the rising tone of Huáng. The percept of a lower minimum target for [laʊ²] than for [xwa³] in Figure X.2 suggests that we hear the speaker’s intent to produce the creaky voice quality that cues the lower tone target even when the glottal source wave is regular enough that the $f_0$ tracking algorithm does not fail.

The third reason that the “pitch track” can only be a first-pass phonetic representation of the melody of an utterance is the existence of so-called micro-prosodic effects, whereby the aerodynamics of producing contrastive properties of consonants can cause systematic variation in $f_0$ on consonants and neighboring vowels that is related to the percept of the consonants rather than to the percept of the utterance tune. These effects also are illustrated in Figure X.2. The syntactic structure and the sequence of lexical tones are identical between the two sentences. The substantial differences in $f_0$ shape for the second, third, and fourth syllables are due to the different manners of articulation for the syllable-initial consonants. When native speakers listen to utterances such as these, they parse these microprosodic effects for what they are, and perceive the intended tone sequence that is common to both despite the marked differences in $f_0$ shape (see, e.g., Reinholt Peterson 1986, Silverman 1986).

These three limitations of the $f_0$ contour can be overcome to some extent by the use of analysis-by-synthesis techniques such as the ‘close-copy stylization’ method pioneered by researchers at Institute for Perception Research (IPO) in Eindhoven (see, e.g., Cohen and ‘t Hart, 1967; ‘t Hart and Collier, 1975; de Pijper, 1983). A close-copy stylization is defined as a synthetic approximation to the melody of the utterance which meets two criteria: “it should be perceptually indistinguishable from the original, and it should contain the minimum number of straight-line segments with which this perceptual equality can be achieved” (Nootseboom, 1997, p. 646). This kind of downsampling of the $f_0$ contour is very easy to do today, because of the re-synthesis utilities based on LPC analysis or PSOLA (Atal and Hanauer 1971, Moulines and Charpentier 1989, also see Carlson and Granström this volume) that have been implemented in many free signal analysis packages. The bottom row of panels in Figure X.2, for example, shows a close-copy stylization of each of the $f_0$ contours in the figure, created using the implementation of PSOLA re-synthesis in Praat (Boersma and Weenink, 2007).

### 3.2 Analysis by synthesis

Making such a close-copy stylization is a first step in developing and testing a phonological representation of the intonation pattern of an utterance in the models of the British English and Dutch intonation systems described in de Pijper (1983) and ‘t Hart, Collier, and Cohen (1990). Both of these models pick out some of the line segments in a close-copy stylization as corresponding to phonologically significant events. Two types of phonologically significant event are identified: prominence-lending movements that are anchored to stressed syllables, and juncture-marking movements that occur at the edges of phrases. The phonological significance of a line segment is determined by the criterion of ‘perceptual equivalence’ (as opposed to the ‘perceptual equality’ of the close-copy stylization). Two stylized contours are equivalent if listeners perceive them to have the same utterance melody. In this framework for modeling utterance melody in languages such as English and Dutch, cataloguing the recurring patterns of sequences of line segments in ‘perceptually equivalent’ melodies is analogous to cataloguing the inventory of contrasting vowels or consonants in transcriptions of words and phrases elicited in the field. That is, determining the patterns of melodic equivalence and dissimilarity across a sufficiently large and varied corpus of utterances in several iterations of analysis and perceptual
testing of the re-synthesized utterances should yield the set of ‘melodically
distinct pitch movements’ for the language variety. The model parameters that
contrast these line segments then are homologous to other distinctive feature sets
for the language, such as its contrasting vowel heights or frication source places.

In de Pijper’s model of British English, for example, there are eight
melodically distinct glissandi which are parameterized in terms of their direction
(rise versus fall), their slope (steep versus shallow), and the pitch levels between
which they move (e.g., a half rise from lower to middle differs from a full rise
from lower to upper levels). Some of these melodic elements are illustrated in the
two panels of Figure X.3. The top panel shows the original $f_0$ contour and a close-
copy stylization of a two-phrase utterance produced by a young male speaker of
British English. The lower panel repeats the close-copy stylization and overlays
an approximation to the re-synthesized contour that would be generated by de
Pijper’s model. The intonation pattern in each phrase is the variant of the ‘hat
pattern’ depicted in Figure 5.6b in de Pijper (1983). There is a steep prominence-
lending half rise early in each phrase (on the first syllable of royal in the first
phrase, and on came in the second) followed by a steep prominence-lending full
fall near its end (on the first syllable of messenger in the first phrase and on ball in
the second). The only other phonologically significant line segments are the two
‘continuation rises’ over the last part of messenger just before the inter-phrase
boundary and over the second half of ball.

![Figure X.3](image)

**Figure X.3** Extract from a reading of the *Cinders* passage in the IViE corpus
(Grabe, 2004) with a two-acent ‘hat pattern’ on each syntactic phrase. Dotted
lines are a close-copy stylization overlaid on the original $f_0$ (top) and on
approximations to de Pijper’s (1983) model of the melodic elements in this
pattern (bottom).

In addition to these parameters that specify local rises and falls, there are two
other essential components of the model. One is the parameter set specifying the
timing of each rise or fall relative to the segments of the syllable or at the phrase
dge that licenses it. In de Pijper’s model of English, for example, an early steep
fall that is prominence-lending starts early enough to be completed at the onset
of the vowel in the prominent syllable, a neutral fall starts 30 ms after the vowel onset, and a late fall does not start until after the end of the syllable.

The other essential component of the model is the ‘declination line’ that is the implicit lower level for the melodically distinct rises and falls and that describes the \( f_0 \) over sections in between the melodically distinct movements. In complex utterances such as the two-phrase extract in Figure X.3, there will be as many declination lines as there are phrases, with ‘reset’ at the phrase boundary. The local declination line is steeper for shorter phrases and shallower for longer ones, as suggested by the two overlaid grey lines in the figure.

### 3.3 Phonological representations

We have presented de Pijper’s model in detail because this kind of analysis by synthesis has proved to be a invaluable tool for going from a database of phonetic representations of a good sample of utterances to an adequately formalized system of phonological representation for languages such as English, which do not offer the fieldworker the crutch of lexical contrast that supports the indispensable initial methodology for studying utterance melody in languages such as Beijing Mandarin or Hakha Lai (see Hyman, this volume, p. 53). Ladd (2008, p. 13) lauds this “IPO theory of intonational structure” as “in many ways the first ... serious attempt to combine an abstract phonological level of description with a detailed account of the phonetic realisation of the phonological elements.” Other formal frameworks that laboratory phonologists began to develop at about this same time also used an analysis-by-synthesis approach to decompose \( f_0 \) contours into contributions from three model components for (1) the set of localized pitch events, (2) the timing of these events relative to landmarks such as vowel onsets in prosodically relevant syllables, and (3) aspects of backdrop pitch range such as the reset points and declination slopes in the IPO model.

Of course, different frameworks make different claims about the allocation of responsibility among these three components, as well as different claims about the appropriate set of parameters internal to each one. However, all fully formalized frameworks have this kind of compositional phonetics, so that the configuration of parameters of the synthesis model that generates a \( f_0 \) contour that is perceptually equivalent to each original \( f_0 \) contour for a set of utterances that share an intonation pattern can be construed as the phonological representation of that pattern. Formalizing the phonological representation of tune in terms of analysis-by-synthesis model parameters in this way gives phonologists a way to compare models of a language across frameworks, or to compare models of different languages within a framework.

One point of comparison is the size of the smallest sequential element assumed. Where the IPO framework takes melodically equivalent glissandi to be the basic atomic units, many other models of English and Dutch decompose each rise or fall into a finer-grained sequence of endpoint notes. Pierrhumbert (1980) and Gussenhoven (1984), for example, both analyze the steep prominence-lending fall of the English ‘hat pattern’ in Figure X.3 as a transition from a higher pitch target to a lower pitch target. These targets, then, were called “high tone” (abbreviated H) and “low tone” (abbreviated L) in an explicit analogy to the use of the terms in work such as Benedict’s (1948) description of Thai and Cantonese, and Ward’s (1948) description of Efik and Igbo. The analogy made it possible to draw directly on the same kind of compositional phonetics that was beginning to be applied to languages such as Igbo in order to understand better the interplay between the lexically-specified melodic elements that are the “tonemes” of the language and any melodic elements that are produced or parsed “post-lexically” for sequences of words or phrases in their larger discourse contexts. The seminal
example of this kind of model is Bruce’s (1977) description of Stockholm Swedish word tones, which inspired critical elements of Pierrehumbert’s (1980) model of American English intonation patterns as well as of Pierrehumbert and Beckman’s (1988) model of Tokyo Japanese word-accent patterns in sentential context.

The analogy worked in the other direction, too, making it clear, for example, that the segmentation grain for utterance melodies is a theoretically interesting question even for languages with lexically contrastive tone. Thus, where Ward (1948) and others analyze Igbo word and phrase patterns as sequences of tone levels, with high or low specified for each syllable, Clark (1978) proposes a system of ‘dynamic tones’ so that a rise or fall is specified only at linguistically significant syllable junctures, as in the IPO framework models described above. Similarly, where Kindaichi (1957) and Haraguchi (1977) analyze Japanese word- and phrase-level melodies in terms of a succession of low or high tones specified on all moraic segments, Kawakami (1957), Hattori (1961), and Fujisaki and Sudo (1971) analyze them as combinations of underlying rises and falls. We will return to this point in Section 4.1 after describing the ramifications of such differences among models for the symbol sets that are a more typical referent of the term “phonological representation”.

3.4 Symbolic representations

The control parameters for producing and parsing spectral patterns of utterances often are referenced symbolically using transcription systems such as the International Phonetic Alphabet (IPA), in which each basic segmental unit is represented by a letter symbol. The second syllable of the utterance displayed in the upper-left panel of Figure X.2, for example, can be transcribed as a sequence of three segments [w], [a], and [ŋ], whereas the second syllable of the paired utterance to its right might be transcribed as these three segments plus an initial [x]. While phonologists disagree on the ontological status of such symbolic tags (see Ladd, this volume) there is an overwhelming consensus that this grain of syntagmatic discretization is a useful starting point for phonological analysis, and hence, that the IPA provides a useful common vocabulary for annotating recordings made in the laboratory or the field and for comparing models of what talkers and listeners implicitly know about how to differentiate words such as the surnames Wáng and Huáng in these sentences.

There is no comparable standard alphabet for segmenting and tagging pitch patterns. The 1989 Kiel revision of the IPA resolved a rivalry between Africanist and Sinological conventions for tagging pitch patterns in languages such as Igbo versus Cantonese, but only by including both transcription systems. In his summary of discussion by the Working Group on Labeling of Suprasegmentals at the Kiel meeting, Bruce (1989, pp. 36-37) describes the failure to achieve even such a laissez-faire resolution for tagging pitch patterns in languages such as English:

There exists an apparent need for a direct way of symbolizing intonation in a phonetic transcription. However, the opinions diverge regarding the exact way of transcribing intonation. For a phonological transcription of intonation the symbolization is very much dependent on the language and the analysis.

Why might tone and intonation contrasts be so much less amenable to a “phonetic” transcription than are consonants and vowels? We think this is because there is no natural universal segmentation for the pitch pattern shorter than the utterance as a whole. That is, there is nothing akin to the segmentation of filter-resonance patterns afforded by the abrupt transition from a stop-like closure into a more open vocal tract in the CV-like “frame” of canonical babbling
(MacNeilage and Davis, 2000) and in the “vocal motor schemes” (McCune and Vihman, 1987) that become the infant’s first words. Across cultures, mothers may use a common stock of attention-getting tunes to draw very young infants into sessions of imitative turn-taking, as suggested by Papoušeč, Papoušek, and Symmes (1991). And this common stock of pre-verbal melodies may be a basis for trends such as the prevalence of raised pitch and rising terminals in yes-no questions, as” noted by Lieberman (1967), Bolinger (1978), and Ohala (1983), among others. However, these are not universals of syntagmatic alternation within the utterance. They do not provide a compelling “natural phonetic segmentation” of the melody into units that are any smaller than the tunes of the alternating interlocutor turns.

This difference in preverbal rhythmic base gives the symbolic transcription of linguistically significant pitch variation a fundamentally different status from the symbolic transcription of linguistically significant spectral variation. Alphabetic transcription can be a useful pre-theoretical tool for identifying events that are very likely to have phonological significance in the vocal tract filter pattern. The analogous use of symbolic transcription for voice source events makes sense only within the context of a research community in which shared expectations about tunes and their relationship to prosodic constituents above the level of CV units can emerge. For example, the Africanist transcription system that is included in the IPA evolved in a community of Bantuists and of researchers working on non-Bantu languages in the West African Sprachbund, where a long history of language contact has given rise to striking commonalities in syntactic and in prosodic organization above the word. The IPO symbols for English and Dutch prominence-lending shapes and boundary pitch movements, similarly, evolved in a community of researchers who developed the symbols as shorthand for particular sets of parametric specifications in a shared analysis-by-synthesis model of the intonation systems of these two closely related language varieties.

3.5 Parametric representations in prosodic phonology

The lack of a universal preverbal rhythmic base for segmenting speech melodies at any level below the whole utterance may also explain why languages with lexical tone contrasts figured so prominently in the early development of a long lineage of frameworks that led to the emergence of the AM framework. That is, lexical contrast typically provides a more compelling functional basis than pragmatic contrast for segmenting the melodic contour into units smaller than the whole utterance, and there is a long history of applying ideas developed for the description of prosodic structures in languages with lexical tone contrasts to languages such as English. An early example is Kingdon’s (1939) application of the notion of “tonetic stress” in the development of a model of English intonation contours that was the first to recognize that stressed syllables before the syllable bearing the “nucleus tone” also can be marked by tones. More recent examples include Liberman (1975), Leben (1976), and Goldsmith (1978), as well as Pierrehumbert (1980) and Gussenhoven (1984), all adapting notational devices from descriptions of lexical tone patterns in various African languages to the description of English utterance tunes.

In reading earlier work in this lineage, we are often struck by the congruence between concepts that were assumed in describing, for example, “the tonephrasing system of Kongo” (Carter 1974) and the parameterization of intonation patterns in analysis-by-synthesis models such as de Pijper’s (1983) model of British English described earlier. In particular, whether lexically-specified tones are transcribed by diacritics on the vowels (as in the Africanist tradition) or by numerals after each syllable (as in the Sinological tradition), utterance melodies are typically described in terms of a convolution of two parts. One part is the
concatenated sequence of pitch levels for the transcribed tones and the other part is linguistically significant modulations of what Ladd (1992) calls the *tonal space* (see Section 4.3). This partitioning of tunes into tones and tonal space is evident in the Beijinghua utterances in Figure X.4, where the rising glissando on the word *hé-zi* 'box' can be transcribed in terms of a sequence of lower and higher pitch targets ([23] in Chao’s tone numbers or LH by Yip’s, 1980 analysis) in both cases, but the high target at the end of the glissando in the right-hand utterance is higher because it is realized in an expanded tonal space. This partitioning is also congruent with the distinction in the IPO framework between the parameters that specify the melodically significant glissandi and the parameters that specify a sequence of declination lines for successive phrases.

![Figure X.4](image)

**Figure X.4** Extracted \( f_0 \) contours for the sentence *Fàng zài nèi-ge hé-zi lì-biar le*. produced by a female speaker of the Beijing dialect of Putonghua in staged dialogues where the context makes it a statement ‘(I) put (them) in that box.’ (left) or a question ‘Had (she) put (it) in that box?’ (right) [from Lee, 2005].

Carter’s account of “the tone-phrasing system of Kongo” is also characteristic in distinguishing between tones that are anchored to specific syllables in a word and tones that signal other phonologically significant anchoring points such as the edges of larger phrases. This distinction is congruent with the distinction between pitch movements on stressed syllables and boundary pitch movements in the IPO models of English and Dutch, as well as with descriptions of some utterance-level melodic contrasts in many languages with lexically contrastive tone shapes. The declarative and interrogative utterances in Figure X.4 illustrate the Beijinghua case. Whereas the pitch events on earlier syllables reflect the lexically specified tones, the high tone anchored at the end of the last syllable on the right is a pragmatic morpheme that signals its interrogative speech act.

In this chapter, we will call this lineage of frameworks “prosodic phonology” — using lower case to differentiate this common noun from the homophonous name of two particular frameworks within the lineage (Bazell, Catford, Halliday and Robins 1966; Nespore and Vogel 1986), and also to make clear that the key ideas invoked by this term are generic. Each of these ideas was developed more or less independently in at least two frameworks within the lineage and was congruent with approaches being developed at the same time in the allied science of speech synthesis. We will adopt transcription conventions from the relevant prosodic-phonology models of tone and intonation systems when we describe \( f_0 \) contours of example utterances or discuss melodic contrasts in the language varieties from which these examples are drawn. For instance, the ten H and L tones in (1) are a possible way to symbolize a subset of the tonal events in the two utterances depicted in Figure X.1 and the strings of functionally annotated H and L tones in (2) and (3) are possible ways to symbolize the tunes of the three utterances depicted in Figures X.3 and X.4.
In both sets of transcriptions in (2) and (3), there are linking lines. Each such line indicates that a tone or tone sequence below is associated to a designated syllable in the orthographic or phonetic transcription above. Also, in both transcriptions, the + infix conjoins tones that are anchored as a glissando around the designated syllable, and in (2), the * suffix on the L+H indicates that English contrasts two rising glissandi, L+H* versus L*+H, which differ in how they are anchored to the designated syllable. By contrast, each % affix in the transcriptions indicates a tone that is anchored at a phrase boundary rather than internally to a phrase, whereas the - suffix in (2) marks a “floating” L phrase tone that is realized somewhere between the preceding L+H* and following H% targets. While we find it useful to adopt these tagging conventions, however, we must emphasize that the symbol strings in (1), (2), and (3) are not narrow phonetic transcriptions. Moreover, they are not even broad phonemic transcriptions until they are construed as names of meaningful configurations of parameter settings in an analysis-by-synthesis model for the speaker’s dialect of (1) Japanese, or of (2) British English, or of (3) Mandarin Chinese.

(2) (One day) a royal messenger came to announce a ball.

(3) a. Fàng zài nèi-ge hé-zi lǐ-biār le.   b. Fàng zài nèi-ge hé-zi lǐ-biār le?

4 A taxonomy of formal parameters

In this section, we amplify on the relationships between three key developments in prosodic phonology and the components of typical synthesis models that allow symbol strings such as the ones in (1), (2), or (3) to be read as pieces of a broad phonemic transcription in some actual or possible formal model of the tone and intonation systems of each of these three languages. We can identify three key developments.

4.1 Segmenting the melody

The first key development was the notion that the melody of an utterance can be segmented into a string of localized events — single notes or the conjoined notes in glissandi or in more complex sequences such as dipping movements — and that this segmentation is autonomous of the formal properties and functions that allow the native-speaker listener to parse the filter-resonance patterns in terms of the consonant and vowel inventories of a language. This idea gives the name of the autosegmental phonology framework (Goldsmith 1976), but it is not unique to that framework. It is implicit in the cataloguing of significant pitch movements in the IPO model, in the identification of turning points in the Lund model (Gårding 1977, 1993), and in the detection of phrase and accent commands in the Fujisaki model (Fujisaki and Sudo 1971, Möbius, Pätzold, and Hess 1993).

It is useful to begin this discussion of the autonomous status of tone segments by reviewing the phonetic bases for the earlier conceptualization of tone as a suprasegmental feature. A great deal of research over the last five decades highlights how a taxonomy of formal properties of vowel and consonant systems across languages emerges from the interplay between information-theoretic principles and the physiology and physics of speech. Indeed, the very fact that
there are vowel and consonant systems can be related to the ways in which spoken utterances are naturally segmented by the spectral discontinuities that result when constrictions (i.e., consonant gestures) are superimposed on more open vocal tract postures (vowel gestures). As Goldstein (1989), Ohala (1992), and many others point out, even though the consonant and vowel gestures are not themselves sequenced, the acoustic patterns they produce are effectively sequenced because of two facts about the types of constriction that yield the most robust CV segmentation. First, these constrictions block the transmission of spectral information that gives the listener clues to coarticulated vowel postures behind the place where airflow is impeded. Second, some contrastive features of the most effective consonant segments, such as the place of impedance for a stop, are only audible during the acoustic intervals for coarticulated vowel postures, so that spectral properties at the edges of vowel segments must be treated as transitions between consonant states and vowel states in order to recover these “hidden” consonant features.

Although concurrent tone gestures also are “hidden” by these consonant constrictions, source and filter resonances are to a large extent independently controlled during intervals where airflow is not impeded. As Pierrehumbert (2000) points out, this independence of source and filter for vowels means that tone features are carried on a considerably more separable channel of acoustic information when compared to the “hidden” features of consonants (which, as just noted, are carried on exactly the same channel of spectral resonance patterns that carry the vowel features). Vowel segments are thus the more reliable intervals for transmitting information about concurrent tone gestures during a sequence of consonants and coproduced vowels. This is the psychophysical basis for a type of tone system in which each vowel segment in a word or phrase is the nucleus of a syllable that can be counted off in the metrical structure of the utterance by virtue of its having exactly one associated lexically-specified tone. This kind of “tone syllabification” is especially characteristic of utterances in languages such as Cantonese, where most syllables are monosyllabic content words, and tone features typically preserve the syllable count even at very fast speech rates where the vowel features are swallowed up (see Section 5.1).

Pike (1948, pp. 3-5) reserved the term “tonal language” for a language with this kind of tone-syllabification system, whereas many other later researchers such as Voorhoeve (1973), McCawley (1978), Goldsmith (1984, 1987), and Hyman (2001, 2006, this volume) defined “tone language” more broadly and used terms such as “restricted tone system” or “pitch accent” to differentiate the “unrestricted tone system” of Cantonese from their accounts of the underlying tone sequence in languages such as Safwa or Tonga, where words typically are longer and phonotactic constraints strongly restrict what tones can occur where within a word. One of the more fiercely contested questions in prosodic phonology today is the prevalence of tone syllabification as a basis for counting tone targets in the melodic contours of words and phrases. The question arises in part because of the alphabetic bias to model melodic contours of languages with “restricted tone systems” as a succession of “phonetic” tones for all syllables even over stretches where the observed pitch pattern on the vowels depends completely on the pitch targets for nearby ”phonemic” tones.

For example, a common strategy for training models of Mandarin Chinese tone and intonation is to use databases of recorded utterances of sentences such as the two in Figure X.2 (e.g., Lee, Tseng, and Hsieh 1993, Shih and Kochanski 2000). This strategy in effect treats the language as if it had a Cantonese-like prosodic system. By contrast, Kratochvil’s (1998) corpus study suggests that the Beijing dialect at least differs prosodically from Cantonese, and that examples such as the two utterances of the sentence in Figure X.4 are more typical. The pinyin orthographic transliteration of this sentence in (3) on p. 15? shows nine zi
(a grammatical unit that Riha, 2008, terms the “morpheme-syllable”), but of these nine, zài, -ge, -zi, -biar, and le are affixes or grammatical particles with “neutral tone” — i.e., they have no lexical tone specification, as indicated for the four zì other than zài by the lack of a tone diacritic. (The locative particle zài is transliterated with the diacritic for the [51] lexical tone because it is listed in most dictionaries under the entry for the related locative verb, but the particle and verb are not homophones; the particle has neutral tone and there is no fall in pitch even in the declarative utterance shown in the left panel, where the consonant and vowel are not lenited to nothing, as they are in the interrogative utterance on the right.) How can we account for the $f_0$ values in neutral tone zì?

Chen and Xu (2006) follow recent accounts such as Yip (1980/1990) to argue that even when there is no lexically specified tone pattern, each syllable in an utterance has a surface target tone level. The target value for a neutral tone zì is M (i.e., midway between the H and L targets of the four contrastive lexically-specified tones) and the actual pitch value is an average of this M with the pitch value of the immediately preceding tone target. By contrast, Li (2003) follows earlier accounts such as Chao (1932, 1968) in assuming that a neutral tone zì has no tone target even on the surface. The pitch pattern over such a zì can be an extension of the pattern for the last preceding lexically-specified tone (e.g., continuing the fall after the [31] falling tone or realizing the optional rising tail after the [214] dipping tone) or it can be just part of the transition between tone targets on either side (e.g., in (3), the last two syllables are the transition from the L tone on the root morpheme ὶ of ὶ-iyor to the pragmatically-specified L% or H% boundary tone at the end).

The disagreement over which type of account is better is reminiscent of the disagreements that occasionally arise in the literature on transitional elements such as the release phase of obstruents in Berber. Coleman (1998) transcribes the release as a reduced [ə] vowel, a phone which Dell and Elmedlaoui (1998) and Ridouane (2008) insist is not part of the phonological inventory of the language, so that by their analyses many syllables are headed by an obstruent consonant, violating a purported universal minimum sonority constraint on syllabicity, albeit not violating most formulations of the universal as a sonority sequencing constraint. However, those disagreements arise very infrequently relative to the consensus view that consonant and vowel gestures in spoken languages tend to be configured syntagmatically in such a way that native speakers and linguists can identify a string of CV units, as in the IPA transcriptions in (3a) and (3b). By contrast, disagreements like the one that gives rise to different counts for the melodic units in these utterances are endemic across the communities of researchers working on tone and intonation. These disagreements are almost inevitable, because they reflect an inherent ambiguity in the parsing of tonal gestures. This ambiguity stems from the fact that a vowel-by-vowel segmentation of the melody is not intrinsic to the production and perception of tone gestures per se, but instead is parasitic on the CV segmentation of the spectral pattern that is intrinsic to the production and perception of obstruent gestures.

Lieberman (1967) proposes a rather different phonetic basis for segmenting the melodic contour that he describes as emerging from the interplay between syntactic structures governing the flow of information in a discourse and the coordination of respiratory and laryngeal postures to control expiratory airflow for phonation. In particular, he suggests that, absent a “marked” gesture to change laryngeal tension, the posture for sustained phonation results in a rise to high $f_0$ at the beginning of controlled expiration and a rapid fall in $f_0$ toward the end, as in the combination of prominence-lending movements in the IPO “hat pattern” in Figure X.3. This rise and subsequent fall forms a natural unit of segmentation, which Lieberman calls the “unmarked” breath group. He also describes a “marked” breath group, which instead has a final rise produced by a
localized laryngeal-tensing gesture. He claims that a comparably localized gesture to boost subglottal pressure can also make for a more extreme early rise or a rise in other positions to mark focal prominence (“emphasis” or “contrastive stress”) in the discourse context of the utterance.

Although Ohala (1970) and other later work has discredited Lieberman’s characterization of a definitive role for subglottal pressure in the production of local melodic events such as the L+H* rise when used to mark focal prominence on a particular syllable or word in English, Lieberman’s depiction of an early rise and late fall that defines the melodic contour for an “unmarked breath group” captures a fairly common aspect of phrasal melody. Safwa, Basque, Japanese, French, and many other languages use a small set of tone sequences, often involving a rise in pitch anchored near the beginning and a fall in pitch anchored later, to highlight the edges of utterances and to segment them into smaller prosodic phrases. These same prosodic phrases often seem to be the domain for specifying an expanded or compressed tonal space to express the relative prominence of the constituent as a whole (see Section 4.3), and it is less implausible that this expression of phrasal prominence relationships could involve adjustments to the pulmonary expiration rate as a mechanism for overall volume control.

The utterance in Figure X.5 (which repeats the f0 contour from utterance (b) in Figure X.1) illustrates this delimitative aspect of the tone-phrasing system of Japanese. There are four prosodic phrases, each of which is marked by an initial rise in pitch. This rise is analyzed in the X-JToBI tagging conventions as a sequence of a low boundary tone that is anchored strictly at the phrase edge and a high phrasal tone which is timed to follow the low tone at some loosely fixed distance which depends both on the prosodic structure of the first syllable and on distance to the next melodic event. In every phrase but the third, there is also a steep fall at a designated syllable that is marked by an apostrophe in the transliteration of the word in (4a). This lexical specification for anchoring a HL tone sequence at some designated syllable differentiates “accented” words such as oyo’ideru ‘is swimming’ from “unaccented” words such as oborete-iru ‘is drowning’.

Although the verb form in the second phrase is unaccented, there is a steep fall in this VP because the following evidential particle is lexically accented. Even in phrases that contain no accented words, however, there is most typically a fall, albeit often a more gradual one, which the X-JToBI conventions analyze as a transition from the high target at the end of the phrase-initial rise to the low boundary tone at the following phrase edge. Prosodic groups (accentual phrases) can be counted off in an utterance from the distribution of the phrasal rises and subsequent steep or gradual falls. Also, while the tone patterns differ in other dialects, with some having more complicated lexical contrasts and some having no lexical contrasts, the metrical structures that are defined by the distribution of tones relative to accentual phrase (AP) boundaries seems to be shared across dialects.

    [ja.ma.no.oITERALITY q] [o.joi.de.ruITERALITY букв. ḳa maru.de o.bo.re.te.i.ru yo’u-da]  
    %L HL L% HL L% H- L% H- HL L%  

b. Ya’mano-wa oyo*ideru-ga, marude oborete-iru yo*o-da.  
    [H*L L H*L L H L H*L]  

c. [H L L L L H L L L L H H H H H H L]
Accounts such as Kawakami (1957) and Pierrehumbert and Beckman (1988) focus on the way that the tone patterns mark off the salient prosodic groups to make for the pan-Japanese metrical system. In such accounts, melodic contours for utterances in the standard dialect are segmented only into those tones that are anchored relative to the phrase edges and those tones that are anchored at the designated syllables of accented words. All other parts of a contour are described as tonally “underspecified” and modeled as transitions between the nearest tones on either side, making tonal transcriptions of Tokyo Japanese utterances such as the one shown in Figure X.5 look like transcriptions of utterances in the Autosegmental-Metrical model of the American English intonation system that was invoked in (2). The transcription in (4a) illustrates, using the X-JToBI conventions (Maekawa, Kikuchi, Igarashi, and Venditti 2002).

By contrast, in Kindaichi (1957) and Haraguchi (1977), the focus is primarily on making a spare underlying representation for the lexical contrasts between the absence versus presence of the HL sequence and (if present) among different anchoring positions within the word. These contrasts are represented by marking the designated vowel with an * to show where the HL sequence is to be inserted at the initial stage of deriving the surface pitch pattern. The pattern on other parts of the accentual phrase is modeled by derivational rules that conditionally insert L and H tones on the initial and final vowels, as in (4b), and then copy the inserted tones or the lexically-specified tones onto other vowels, to produce a “fully specified” surface pattern, as illustrated in (4c). This tone-spreading account makes the intonation system of Tokyo Japanese look like Voorhoeve’s (1973) picture of the “restricted tone system” of Safwa and also like Goldsmith’s (1984) account of “tone and accent in Tonga” a decade later.

The difference between the 13 tone segments assumed in the transcription in (4a) and the 21 tone segments assumed in the transcription in (4c) is also parallel to the difference between specifying tones for just four of the zi in Li’s (2003)
model that yields the transcription in (3a) as compared to specifying these four plus the five M targets for the neutral tone zi in Chen and Xu’s (2006) model. In both of these cases, one account assumes that the sequence of syllables (or other potential tone-bearing units) is “fully-specified” for tone targets whereas the other account assumes that the nodes at this level of the prosodic hierarchy are “underspecified” for tone. These names characterize the disagreement in terms of their different assumptions about the set of localized pitch events — i.e., the first of the three synthesis model components listed in Section 3.3.

Such disagreements have consequences for the depiction of the “underlying” tone specification. For example, in the fully-specified account of Japanese tone patterns, the starred tone of the H*L word melody is associated to the designated mora (which is marked with a * in the lexicon) at the first stage of the derivation, and then a L is inserted on the first tone-bearing unit of an AP just in case that mora does not already bear a tone specification. This account therefore predicts that there will be no tone difference between a sequence of clauses such as yonde mi’ru ‘call and then see’ and a verb-auxiliary construction such as yonde-mi’ru ‘try calling’ since in both cases the initial vowel of mi’ru will already have an associated H tone at the stage of the derivation when an initial L is conditionally inserted, as in (5c). By contrast, in the underspecified account, that first L is a boundary tone that marks the edge of an AP whether or not the first syllable is accented. Thus, utterances of the two-clause sequence often would be distinct from utterances of the verb-auxiliary construction, because the two-clause sequence often will be produced as two AP, as shown in (5a).

(5)  a. yonde mi’ru ‘call and then see’       yonde-mi’ru ‘try calling’ | | | | % L H - L% H + L L% % L H - H + L L% b. [jondemiru] [jondemiru] c. yonde mi*ru ‘call and then see’      yonde-mi*ru ‘try calling’ | | | | L H H* L L H* L d. [L H H L] [L H H L]

It is important to note that these differences in the analysis of the underlying forms stem from the more fundamental disagreement about the nature of surface phonetic representations. In the X-JToBI account of Tokyo Japanese (as in all ToBI framework accounts), the surface phonetic representation is the actual pitch pattern, as deduced from representations such as Figure X.5, which shows an $f_0$ track calculated from a recording of a specific utterance of (4a), or as shown in (5b), which is a schematic “pitch track” summary of the many $f_0$ contours that we have observed for actual utterances of the phrases in (5a). In Haraguchi’s account, by contrast, the surface phonetic representation is still a symbolic transcription — a sequence of discrete pitch targets associated vowel-by-vowel, as in (5d). On the surface, then, this account makes Japanese look like an unrestricted tone language. How can we decide between these two accounts?

Pierrehumbert and Beckman (1988) made the following predictions. If the fully-specified account is an accurate representation of what the speaker intends to produce, then a sequence of spread L tone targets (as in the last four vowels in oyo’ideru-ga in (4c)) or a sequence of spread H tone targets (as in the second through sixth vowels in oborete-iru-yo’o-da in (4c)) should show the same pattern of actual $f_0$ values over the associated vowels regardless of the length of the sequence. In the underspecified account, by contrast, the $f_0$ contour over such stretches could fall or rise at different rates, depending on the distance between the two tone targets specified at the surface. Pierrehumbert and Beckman tested
these predictions using a set of elicited utterances of three-phrase sentences in which both the accent status and the number of syllables in the words in the middle phrase were systematically varied. For unaccented medial phrases, they measured the slope of the $f_0$ downtrend over the interval between the peak $f_0$ near the beginning of the accentual phrase to the minimum $f_0$ at the next phrase boundary — i.e. over an interval that would be represented as a sequence of H tones in Haraguchi’s account but as a mere transition from a phrasal H- to a L% boundary tone in the underspecified account. For accented phrases, they fit two slopes, differentiating the steep fall of the H+L tones at the designated syllable (which they predicted to have a fixed duration and slope) from the shallower decline over the variable-length region from the L of the accent to the L% at the end of the phrase. In both cases, the slope of the downtrend over the variable-length region up to the phrase edge was steeper for shorter intervals and shallower for longer ones, as predicted by the underspecified account.

In differentiating between the fully-specified and underspecified accounts of Tokyo Japanese tone patterns, Pierrehumbert and Beckman (1988) fit very simple (straight-line) curves to the $f_0$ contour over both types of tonally unspecified intervals. As Pierrehumbert (1980, p. 12), van den Berg, Gussenhoven and Rietveld (1992), Beckman and Pierrehumbert (1992), Myers (1998), Ladd and Schepman (2003), and many others point out, however, the shape of a transition over tonally unspecified regions is a research question in its own right. Moreover, it is a question that is tied up inextricably with questions about alignment or anchoring — i.e. about how the speaker synchronizes tone gestures with vowel and consonant gestures so that the listener correctly parses where the targets are anchored in relation to prosodic positions such as stressed syllables and phrase boundaries.

4.2 Anchoring the tones in time

The second key development in prosodic phonology was the idea that tonal autosegments are not suprasegmental features of the vowel segments on which they realized. Rather vowel (and consonant) segments as well as tone segments are associated to positions in a metrical structure, and this structure and the association patterns are objects of study in their own right. This idea is often associated with what is termed the metrical phonology framework (e.g., Liberman 1975/1979, Liberman and Prince 1977, Selkirk 1981), but again, it is not unique to that framework. It is developed more fully in the treatment of coarticulation of consonant and vowel features in what is called the articulatory phonology framework (see Browman and Goldstein 1986, Byrd 1993, Byrd and Saltzman 2003, other work reviewed by Fletcher this volume). For tonal autosegments, this idea is implicit in the functional separation between prominence-lending pitch movements and boundary pitch movements in even the earliest IPO system models, and it corresponds to the distinction between turning points and pivots in the Lund model and to the distinction between phrase commands and accent commands in the Fujisaki model.

To show how this development was separate from the first key idea, we begin by comparing what “association” means in the two different accounts of Japanese discussed above. The fully-specified “phonetic representations” in (4c) and (5d) can do away with the link lines and just list the string of H and L tones, reflecting the assumption that each tone is aligned simply to coincide with the vowel or moraic nasal segment to which it associates by rule. Beckman, Hertz, and Fujimura (1983) describe a synthesis model couched in this fully specified autosegmental phonology framework, which specifies a target $f_0$ value for the H or L tone midway through each vowel or moraic nasal in this way. By contrast, the underspecified surface transcriptions in (4a) and (5a) must show link lines to
identify the accent tones and their designated syllables in accented words. Other tones must be annotated for their anchoring relationships. The annotation conventions differentiate the %L and L% boundary tones that anchor tightly at the phrase edge from the H- phrase tone that is only loosely aligned relative to the edge of the accentual phrase that begins with an unaccented word. Pierrehumbert and Beckman (1988) describe a synthesis program couched in the AM framework which specifies target \( f_0 \) values at various time points that are chosen to relate the linguistically significant \( f_0 \) peaks, valleys, and inflection points ("elbows") in the phonetic representation in Figure X.5 to the functional differences among the accent tones, the boundary tones, and the phrase tones. Although the input is a sequence of tones, the ways in which tone sequences such as the L% H- are anchored to positions such as the phrase edge makes their model much more like Kawakami's account than like Kindaichi's.

Pierrehumbert and Beckman's (1988) model of Japanese tone structure relied crucially on Bruce's seminal model of Stockholm Swedish tone patterns (Bruce 1977, 1982, 1987, 1990). In Bruce's model, there are three types of tone which are anchored in different ways to designated constituents or positions at several levels of a hierarchy of prosodic units. The first two relevant levels are the grouping of consonant and vowel constituents into short (unstressed) and long (stressed) syllable constituents, and the grouping of unstressed syllables together with neighboring stressed syllables into word constituents. The second level is marked tonally by the word accent, a H+L tone sequence that is anchored to a designated strong syllable in each word. This culminative distribution of the H+L sequence means that in longer Swedish utterances, words can be counted off in the melodic contour for an utterance by recognizing the word-accent tones and their anchoring points. Above the word level, whole utterances and prosodic phrases within utterances are delimited by boundary tones such as the L% for the "terminal juncture fall" (Bruce 1983, p. 223). Also, the melodic contour for each phrase must include a H- tone, called the sentence accent in Bruce (1977), the phrase accent in Pierrehumbert (1980), and the focal tone in Gussenhoven and Bruce (1999). The focal tone is realized just after the word accent of the word with "sentence stress" — i.e., a word that is in narrow focus in the discourse context or the last word in the phrase when there is broad focus over the whole phrase. All the tone types are shown in the sample transcriptions in (6). These schematic "pitch contours" are based on \( f_0 \) tracks given in Bruce (1977) and are intended to give a sense of the typical patterns of truncation and undershoot.

\[
(6) \quad \text{a. } \textit{mellan mälén} 'between meals' \quad \text{b. } \textit{mellanmålen} 'snack' \\
\%L H^+L H+L^* H^-L% \quad \%L H^+L H- L% \\
\text{c. } \textit{MELLAN målen} 'BETWEEN meals [not AT meals]' \\
\%L H^*+LH^-H-L^* L% \\
\]

As in most dialects of Swedish, the Stockholm variety has a lexical contrast between two anchoring patterns for the word accent, transcribed by Bruce (1990) as H+L* ("Accent 1") versus H^+L ("Accent 2"). In Accent 1 forms such as \textit{anden} 'the duck', \textit{anamma} 'accept', and \textit{målen} 'meals' produced in contexts with one or more preceding syllables, the H+L* denotes a fall to a low pitch target within the stressed syllable that starts from a pitch peak or a high inflection point (an "elbow") about 120 ms before the low target. In Accent 2 forms such as \textit{anden} 'the ghost', \textit{låmna} 'leave', and \textit{mellan} 'between' there is a peak or high elbow within the stressed syllable and a fall to a valley or a low elbow 120 ms later. A compound word such as \textit{mellanmålen} 'snack' is marked by a H^+L (Accent 2)
anchored to the designated syllable of the first component and no word accent on any later component. This accenting in compound words mirrors the typically initial stress in the native Germanic stratum of the lexicon.

Other important concepts are truncation and undershoot. When an Accent 1 word with initial stress is initial in its utterance, the leading H of the accent will be effectively “hidden” by the preceding silence, so that the underlying H+L* is truncated to be just the L* target on the designated syllable. Also, when an Accent 1 word with final stress is final in its phrase, the close succession of fall for the H+L* followed by rise to H- and fall to L% leaves very little room for the word accent to be realized. There is undershoot so that the L* is effectively a mid tone. By contrast, the trailing L of the Accent 2 fall is typically fully realized, because the designated syllable in an Accent 2 word cannot be final. Moreover, the duration of the transition from the H* target to the elbow for the trailing L is very stable. At the other extreme, the H focal accent has no very fixed constraints on its alignment other than that it is after the accent tones of the focalized word.

In compound words, it is especially late, because the trailing L of the word accent has a secondary anchoring point at the stressed syllable of the second (or last) word in the compound. This account of the focal H- as a “floating tone” is invoked in AM-framework transcriptions by showing no line linking it to a designated syllable.

Two aspects of Bruce’s work are especially noteworthy. The first was his rigorously controlled laboratory phonology methods. He designed his materials to allow a systematic comparison of melodic contours for Accent 1 and Accent 2 words of different lengths in both final and non-final position and in both non-focal and narrow focus contexts. This was necessary for him to be able to disentangle the tones that are specified by the lexicon from the tones that mark other levels of prosodic organization. He used analysis by synthesis to verify the segmentation of the melodic contour into these disparate elements and to examine their timing relative to the consonants and vowels at phrase boundaries and at the designated syllables within each phrase.

An equally important aspect of his work was his rigorously imaginative adaptation of key ideas from prosodic phonology. He did not let broad-stroke typologies dictate what analogies could be drawn between the tone patterns of Swedish and the intonational accents of English, and was among the first to grasp the implications for prosodic phonology of Bolinger’s (1958) theory of pitch accent in English as interpreted by Vanderslice and Ladefoged (1972). He saw that the syllable bearing the word accent is not the only potential site for anchoring a tone target in a citation-form utterance of a Swedish word, and that tones realized at variable distances from the accented syllable in many dialects (including the Stockholm one) might reflect rhythmic organization above the word. This let him re-conceptualize the originally simpler theory of “association” only between autonomously segmented tones and some unspecified temporal location within the set of “tone-bearing-units” at just one relatively low level of the prosodic tree (that is, either to just the V nodes, just the μ nodes, or just the σ nodes) as a more complex synchronization at “critical timing points” (Bruce 1983, p. 234) that speakers and listeners control to resolve potentially conflicting demands in different phonological domains.

For the speaker, these conflicts involve “the interaction between the timing of phonatory and articulatory gestures” (Bruce 1983, p. 222), which cannot follow an invariant rhythm because the words in a sentence can be one syllable or longer, initially-accented or accented on a later syllable, in focus or subordinated to a neighboring constituent, and so on. Consonant and vowel gestures in a particular utterance of a string of words must be synchronized with each other so that the listener can parse the syllable count, hear whether each syllable is stressed or unstressed, and, if stressed, whether it is an extended vowel gesture.
or a coda consonant gesture that contributes the second mora in the syllable. Tone gestures also must be synchronized with the consonant and vowel gestures so that the listener can hear which stressed syllables are accented, whether an accent is H+L* or H*+L, and what word is highlighted by the focal H-. These different prosodic functions impose different demands. Realizing the word-level contrasts between short and long syllables and between H+L* and H*+L accents places stringent demands on the timing of the targets. In realizing the utterance-level contrast between focus and background, on the other hand, the exact timing of the focal H- is less relevant than achieving a particular target peak value, since the latter signals prominence relationships among words and phrases as well as among syllables within each word. Conflicts among these demands can be reconciled by adapting the tone targets (e.g. through truncation or undershoot) or by adapting the vowel and consonant targets (e.g., lengthening a final accented syllable to realize a complex sequence of word-accent tones, focal H-, and boundary tones, as suggested by Lyberg 1981). To model the relevant interactions, the segments and tones must be observed in more contexts than citation form utterances.

In Bruce’s original formulation of this “synchronization hypothesis” he differentiated between two orientations for evaluating the synchronization. From the “phonological point of view” of a “production oriented model” it is useful to specify the critical timing points for the underlying tone targets, but these may not map neatly onto the “perceptually critical” f0 events such as rising or falling glissandi. For example, in his own perception experiments on the timing of the H+L* targets of Accent 1 versus the H*+L targets of Accent 2, Bruce found that the times of the starting and ending points traded off with the steepness of the fall in a way that suggested that subjects listened for the point of maximum velocity in the middle of the fall. However, reference to this midpoint time “is possible only in a sonorant environment” so that “from a perceptual point of view, it is probably an estimate of the timing of the entire f0 change ... that is decisive” (Bruce 1983, p. 231).

Bruce’s hypothesis was developed to account for the variable realizations of Swedish word accents across different sentence contexts and different dialects, but it was an important precursor to the AM model of Japanese tone structure presented in Pierrehumbert and Beckman (1988), as well as to the development of the AM framework generally. Initially, development of the framework was addressed more to the production-oriented aims of finding “invariant” or “underlying” tone targets and their modes of association to phonologically defined positions in the hierarchy. For example, Pierrehumbert and Beckman (1988) proposed that the L% and H- tones in their model of Japanese are associated initially to the two accentual phrases on either side of the boundary that the pitch rise marks, but then that each tone is also associated secondarily at a later derivational stage to the first unaffiliated mora in the accentual phrase that begins at the boundary. They observed differences in the shape of the rise and in measured f0 minima for what they called the “strong L%” versus the “weak L%” and they attributed these differences to a contrast between having and not having a secondary association to the first mora in the following accentual phrase. Gussenhoven and Bruce (1999) similarly propose to account for the shape of the trough in citation form utterances of compound words in Stockholm Swedish in terms of a secondary association of the trailing L of the H*+L accent. Grice, Ladd, and Arvaniti (2000) catalogue other examples of languages where a phrase accent can be analyzed as having a dual affiliation to both the edge of a larger prosodic domain and to some designated syllable within the domain. This focus on tone targets and their anchoring relative to the prosodic structures that the speaker controls is congruent with the production-oriented approach of the articulatory phonology framework (e.g. Browman and
Goldstein 1990, Byrd 1996). For example, Xu and Liu (2007) apply a model of Putonghua lexical tone alignment to examine syllables of both Putonghua and English in order to probe for universal patterns in how an onset consonant gesture is anchored to its syllable to be coarticulated with the relevant vowel. This application suggests some of the questions that can be addressed fruitfully using production-oriented models that assume invariant underlying tone, vowel, and consonant targets for the speaker that are aligned with each other to reflect the “temporal signature of prosody” (see Fletcher 2009).

Other recent work, however, suggests that the time is ripe to begin to reorient our models to incorporate constraints on the listener, too. For example, Arvaniti, Ladd, and Mennen (1998, 2000) show that the timing of prenuclear rising accents in Greek does not fall out from a simplistic model that designates either the L or the H as the target that is associated to the designated syllable. Rather, the L is anchored just before the syllable-initial consonant and the H is anchored to coincide with the CV boundary in the following syllable. Unlike Swedish, Greek has only five vowels, with no prosodic contrast between short and long vowels or short and long consonants. Many syllables are CV and vowels tend to be quite short. Also, whereas many Swedish words follow the common Germanic pattern of root-initial stress, the position of the stressed syllable in a Greek word is constrained only to occur on one of the last three syllables. Within this three-syllable window stress placement is “phonologically unpredictable” (Arvaniti 1999, p. 171). Given these characteristics of the language, the observed anchoring pattern for Greek prenuclear rising accents may have emerged as a way to provide the listener with a robust “estimate of the timing of the entire \( f_0 \) change” in order to reliably parse the location of each accented syllable in an utterance.

These demands on the Greek listener are different from the demands on the listener from a language such as Dutch, where there are many more than five vowels in the inventory, vowels are typically longer, and there also is a much larger variety of typical syllable structures, including a contrast between syllables with short vowels and syllables with long vowels. Ladd, Mennen, and Schepman (2000) show that in Dutch, the timing of the end of a rising accent is not fixed in the same way as in Greek. Rather, it is later relative to the end of a syllable with a short vowel and earlier relative to the end of a syllable with a long vowel, and this difference in anchoring of the endpoints supports the vowel length contrast even for speech rates and discourse contexts where the vowel durations themselves are not robustly different.

Arvaniti, Ladd, and Mennen (2000) end their paper with a call for more research both to refine what “association” means for our models of the prosodic structures that the speaker intends to produce and to devise better methods for understanding how targets and their timing properties are realized in the speech signal that the listener parses. One promising line of research in this vein is comparative work such as Smiljanić (2006). Smiljanić looked at accent-related rises in standard Serbian and Croatian, language varieties which are mutually intelligible but which differ in whether there is a lexical contrast between word accents with an early versus a late peak. Smiljanić found that speakers of both varieties signal focal prominence on a word by manipulating the timing as well as the maximum \( f_0 \) value of the pitch rise to the accent peak. However, the timing effect is much smaller in Serbian, where the anchoring of the rise also signals the contrast between the two word accent types. We need more such comparative work in other language groups to develop our understanding of the potential role of functional load in the interaction between demands on production and demands on perception. We also need more work that does what both Bruce (1977) and Smiljanić (2006) did — namely, to observe tones in words across a good variety of sentential and discourse contexts, to see how variation in the demand for precise “horizontal” anchoring of tone targets relative to critical
positions within a word interacts with variation in the demand for precise “vertical” positioning of the tone targets relative to the tonal space.

4.3 Tone scaling and the tonal space

The third key development in prosodic phonology was the idea that speakers can raise or lower and expand or compress the local tonal space as a whole and also independently scale tone targets up and down within the tonal space, to reflect both autosegmental contrast and relative metrical strength, as well as other sorts of linguistic (or “paralinguistic”) relationships. While there is a broad consensus that this separation of “vertical” position into two parts is necessary, the separation is realized differently in different AM-framework models, and the linguistic nature and formal status of the independence remain controversial. We illustrate one way in which the independence of tone-scaling and tonal-space specification parameters has been modeled, using an AM-framework description of Tokyo Japanese that was first developed and tested in an analysis-by-synthesis system by Pierrehumbert and Beckman (1988) and then modified by Maekawa, Kikuchi, Igarashi and Venditti (2002) in developing the X-JToBI conventions that were used in tagging the Corpus of Spontaneous Japanese (Maekawa 2003). The separation of parameters in the model corresponds roughly to the specification of variable accentuation levels for turning points independent of the parameters of the tonal grid in the Lund model of various dialects of Swedish and to the independent specification of amplitude values for accent commands and phrase commands in the Fujisaki model of various dialects of Japanese.

As noted earlier, we have adopted Ladd’s (1992) term “tonal space” to talk about the effects that the IPO-framework models generate by specifying variable starting values and slopes for declination lines over different parts of an utterance. Ladd chose this term to have a framework-neutral way of referring to what Chao (1930) called the pitch “range” when he proposed his “system of tone letters” and the corresponding numerical notation that we used to indicate the lexically contrastive pitch pattern on each of the syllables in the transcriptions of the Putonghua utterances in the caption to Figure X.2. Chao (1932, p. 124) identifies “several abstractions” that must be made to record the pitch patterns that differentiate the tone classes in any dialect of Chinese. Specifically, each pitch level must be calculated “relative to the speaker’s range of voice, so that what would be a low tone for a soprano is actually higher in pitch than the high tone of a tenor.” Moreover, “the range of pitch between different tones and within the limits of moving tones is also a variable quantity depending on force of articulation and force of vocalization.”

The abstraction over different speakers’ voices is analogous to the abstraction over different vocal tracts when computing targets in some speaker-normalized representation of the vowel formant space. The abstraction over variable “force of articulation” is analogous to the constancy of vowel-class identity across the hyperarticulation-hypoarticulation continuum (Lindblom 1990). An important difference between these two spaces is that the “force of articulation” and the “force of vocalization” effects on vowel formant values are necessarily small compared to speaker effects, because maneuvers such as contracting the strap muscles to lower the larynx can change a soprano’s vocal tract length by only a small amount relative to the typical length difference between her vocal tract and a tenor’s. By contrast, the “force” effects on pitch values can be extremely large relative to the differences across speakers, so that the soprano’s H tone in a very subdued speaking style can be much lower than the tenor’s H tone in a very forceful speaking style. A phonological consequence of this difference between the phonetic spaces is that when force of articulation and force of vocalization effects on vowel formant values are phonologized as linguistically significant
markers of strong versus weak positions in the prosodic hierarchy of a language, the markers typically can be described in terms of a small number of discrete prosodic constraints on what vowel targets can be specified for moras or syllables in different positions of the hierarchy. Analogous prosodic constraints on what tone targets can be associated in different positions are fairly common across spoken languages (cf. Section 5.3), but an even stronger universal is the phonologization of the control parameters for positioning tones within the tonal space and for varying the dimensions of the tonal space itself so that these can act not just as discrete markers of the set of categorical contrasts in prosodic organization, but also as gradient markers of more subtle differences in relative metrical strength as well as of other linguistic scales.

The bottom panel of Figure X.5 illustrates the parameterization of the tonal space and of tone scaling that Pierrehumbert and Beckman (1988) built into their synthesis model for Japanese, as these parameters are understood in the version of this model that was incorporated into the X-JToBI labeling conventions on the basis of later research that is reviewed in Venditti, Maekawa and Beckman (2008). In this model, there are tonal-space or tone-scaling effects that refer to three different types of prosodic constituent — the intonation phrase (IP), the accentual phrase (AP), and the prosodic word (\( \omega \)).

At the beginning of the first IP, the reference line that defines the bottom of the tonal space is initialized to reflect overall engagement or volume within the speaker’s voice range. The reference line for the utterance in Figure X.5, for example, is initialized at 70 Hz. This value is maintained until late in the last IP of the utterance, where the effect of “final lowering” kicks in, to signal discourse-level functions such as topic shifts or yielding of the floor to the other speaker. Final lowering is a change in the reference line time function, from having a fixed value to showing a decline over some span at the end of an IP. In the turn-final utterance shown here, for example, the effect reaches in to lower the reference line by 44 Hz per second starting at 0.45 seconds from the end of the last IP.

The IP is also the level of prosodic structure where the value for the top of the tonal space is (re)initialized. The initial topline values for the three IPs in the utterance in Figure X.5, for example, are set at 130, 66, and 110 Hz above the reference line. The IP is also the domain of downstep, a compression of the tonal space triggered at each lexical accent. This effect is implemented in the model by reducing the distance of the topline from the reference line by a fixed ratio. The downstep ratio in the first IP that is triggered by the accent on the first syllable of the prosodic word Yamano, for example, is 0.62 — compressing the tonal space to 62% of its original span.

Tone targets at the level of the IP, the AP, and the \( \omega \) are then positioned within the local tonal space that is defined by the additive effects of the initial IP topline specification, the compression at each previous downstep, and edge effects such as final lowering. Position within the tonal space first of all defines the discrete contrast between H tones (the targets that are closer to the topline) and L tones (the targets that are closer to the reference line). The level of the AP, for example, is defined by the rise from the %L or L% boundary tone to the H-phrase tone. At the level of the \( \omega \), the lexical contrasts among accented and unaccented words are expressed by the presence and (if present) the location of the H+L accentual fall.

The top and bottom of the tonal space also act as a reference for continuous within-category contrasts in metrical strength. Stronger L tones are scaled lower, to be closer to the reference line, and stronger H tones are scaled higher, to be at (or even above) the topline. Some of these strength contrasts are intrinsic to the tone target type. Within an AP containing an accented \( \omega \), for example, the H tone of the H+L word accent is intrinsically stronger than the H- of the phrase-initial
rise; it will be higher relative to the topline, other things being equal. Other strength contrasts are extrinsic and reflect other types of linguistic structure, such as the discourse-level differentiation between given and new information. The imagined context for the performance of the utterance in Figure X.5, for example, is a conversation between two spectators at a triathlon relay race. The other speaker has just asked whether the athlete who is swimming could be Yamano. The H of the word accent in the first AP goes above the local topline to reflect the discourse-level prominence of Yamano as a contrastive topic. The H of the word accent in the second AP is much lower, reflecting both the compression of the tonal space at the downstep and also the given status of the verb oyoideru ‘is swimming’ in this dialogue context.

The effects that are illustrated in Figure X.5 are parameterized in somewhat different ways in the Fujisaki model that Hirai, Higuchi and Sagisaka (1997) used to analyze several large multi-speaker corpora in order to develop the intonation component of CHATR, a concatenative speech synthesis system with prosody-based unit selection (Campbell and Black 1997). For example, in the Fujisaki model, downstep is not modeled explicitly, but instead falls out from the choice of amplitude values for successive accent commands. At the same time, there are important commonalities between these two models. In particular, both models encode relative prominence relationships among tone targets using two different sets of parameters. In the Fujisaki model, there is a step function (the phrase command) to (re)initialize the backdrop tonal space at the beginning of each new IP and a matched pair of step functions (the accent command) that generates the rise to the first H target in each AP (as well as the fall at the accent or at the end of the AP if there is no accented W in the phrase), and the amplitude of each of these two commands is a continuously variable parameter. That is, the distinction between these two amplitudes corresponds roughly to the distinction between the tonal space parameters that are initialized at the level of the IP and tone scaling parameters that are specified for the tone targets that are obligatory at the level of the AP in the AM-framework model depicted in Figure X.5.

One critically important difference between the two models is the treatment of L-tone scaling. As noted already, prominent L tones are scaled downward toward the reference line in this AM-framework model. In the utterance in Figure X.5, for example, the L% tone at the IP boundary after oyoideru-ga is lower in the local tone space than the L% tone at the mere AP boundary after Yamano-wa, reflecting the difference in metrical strength between those two positions in the prosodic hierarchy. In Osaka Japanese, where there is a contrast between %L-beginning and %H-beginning words, there is a similar downward scaling of this initial L tone as well as a delay in the beginning of the following rise in L-beginning unaccented words under focal prominence (Kori 1987) and in pragmatically loaded questions (Miura and Hara 1995). These effects would be difficult to model in the Fujisaki framework without introducing another type of (downward pulsing) accent command that can be positioned at places other than the beginning of the AP.

Another critical difference is in the treatment of effects such as final lowering. Beckman and Pierrehumbert (1986) suggest that extreme final lowering defines one end of a continuum which has (at the other end) an effect that they call “final raising” which they observed in syntactically unmarked questions. As already stated, in the AM model in Figure X.5, this kind of edge-in effect is modeled directly as a change in the shape of the tonal space at the end of some phrasal grouping, analogous to the way that phrase-final lengthening and initial strengthening are treated in the π-gesture model of Byrd and Saltzman (2003) and other articulatory-phonology framework models (see review in Fletcher 2009). In the Fujisaki framework, by contrast, such edge-in upward or downward slope differences cannot be modeled directly. There is a necessary downtrend
across the tonal space for the whole IP, because the phrase command impulse is
smoothed by a filter function and the resulting curve is convolved with the
concurrent accent commands, each of which is also smoothed by a different filter
function. However, since these filter shapes are intended to reflect “hard”
physiological constraints (cf. Öhman 1967, Fujisaki 1983), they are not under the
speaker’s direct control. In order to vary the slope as a way of marking structural
properties such as the discourse property of being turn final, the modeler must
insert a phrase command with just the right amplitude at some place near the
end to counter the downtrend from the damping function. The inability to model
systematic slope variation directly makes the Fujisaki model fundamentally
different not just from the AM-framework model of Japanese, but also from
Grønnum’s very different model of functionally similar effects in Standard

Grønnum’s model, on the other hand, is fundamentally different from both
the Fujisaki model and the AM model in that all aspects of the tone pattern are
treated in terms of a hierarchy of trend lines, with slopes that are specified for the
nested spans of the individual stress groups within individual clauses within a
semantically coherent text. Factors affecting these slopes are the length of the
span (e.g., the clause-level slope is very steep for clauses containing fewer stress
groups and very shallow for clauses containing many stress groups) as well as
the same discourse-level factors that Beckman and Pierrehumbert (1986) identify
as the function of edge-in effects such as final lowering.

As Grønnum (1990, p. 199) points out, the Danish effects are formally distinct
from the Japanese effects in that they are global and not localized to the phrase
end. The downtrend that signals finality “does not just reach one half-second in
from the end, it reaches in all the way back, across several ... stress groups to the
onset of the utterance.” In Liberman and Pierrehumbert’s (1984) AM-framework
re-interpretations of Grønnum’s results, this longer-range clause-level slope
function is modeled in terms of downstep triggered locally at each successive
accent. The speaker would have to be able to specify a different downstep ratio at
the beginning of each IP in order to simulate the difference in slope between a
final and non-final clause. The even longer-range slope of the text, on the other
hand, is modeled in terms of the speaker’s specific choices for reference line or
topline initialization values for the successive phrases. Grønnum (see Thorsen
1984, p. 307) criticizes this “local” approach as arbitrarily allocating responsibility
to disparate sets of formal parameters to account for the functionally uniform
hierarchy of syntactic and semantic coherence.

Grønnum (1995, p. 348) voices a related criticism in her review of the equally
“local” treatment of downtrends in Möbius’s (1993) Fujisaki-framework model of
German. Specifically, she points to his results that the amplitude of the phrase
accent command depends on the number of accents and also (in short sentences)
on whether the accent is early or late. After quoting from Möbius’s comparison
between his more “local” approach and her “hierarchical” one, she agrees:
That is exactly ... the problem with FUJISAKI’S model as adapted by the
author: it permits phrase command amplitudes to depend on accent
location, and it does not supply criteria for a principled choice between
several sets of phrase and accent parameters which each render an
acceptable f0 copy of an original, if such are conceivable. And that, I think,
is incompatible with a model which purports to be physiologically and
linguistically motivated.”

Ladd (1992, 1993) and others point to a comparable “degrees of freedom”
problem for the tone scaling and tonal space parameters of the AM-framework
models of English and Japanese associated with Pierrehumbert and her
colleagues, but it surely is a problem for Grønnum’s model, too, once one goes
beyond carefully scripted lab speech. Indeed, this indeterminacy will be a
problem for any analysis by synthesis model that is sophisticated enough to simulate the ways in which tonal space and tone pattern interact in speech but relies exclusively on goodness of \( f_0 \) fit as a criterion for choosing among parameter settings. In short, there are very pressing research questions that need to be addressed before we have a good model of tone scaling and its relationship to tonal space control, including the overarching one that Grønnum identifies in her review of Möbius’s model: What kind of criteria can be applied to distinguish among models or among different parameter settings within a model?

As noted in Grønnum’s review, Möbius defends his choice of framework on the grounds that the tonal space parameters in Fujisaki model are physiologically motivated. The basis for this claim is in Öhman’s seminal model of Scandinavian “word and sentence intonation” in which he posited two distinct laryngeal gestures for word accents and sentence-level patterns, and suggested that these could be identified with independent activation of two different parts of the cricothyroid muscle (Öhman 1967, pp. 29-30). Fujisaki (1983, pp. 53-54) follows Öhman to posit the same physiological correlates for the different damping functions that he proposed for the phrase command and the accent command. Work on the control of \( f_0 \) in speech has not supported this idea. Neither has it identified evidence of separate “gestures” for tonal space versus tone scaling parameters, because there is no compellingly obvious way to conceptualize the task space. In this respect, the articulation of \( f_0 \) is fundamentally different from the articulation of spectral correlates of consonant constrictions. There are some suggestive ideas in work on physiological correlates of tone and pitch accent contrasts, such as Gårding, Fujimura and Hirose (1970), Erickson (1978, 1993), Erickson, Honda, Hirai and Beckman (1995), Beckman, Erickson, Honda, Hirai and Niimi (1995), Hallé (1994), and Sugito (2003). There is also research such as Herman (2000) and Epstein (2002), documenting perceptible differences in vowel amplitude and voice quality associated with the final lowering effect. These non-\( f_0 \) correlates perhaps could help in conceptualizing the task space for tonal space gestures if examined at the articulatory level, as suggested in Herman, Beckman and Honda (1996). However, the interactions among laryngeal tension, vocal fold thickness, and pulmonary effort are complex and not completely understood. There looks to be a great deal of basic research yet to be done before physiological evidence can be brought to bear directly on the degrees of freedom question.

Another avenue of attack that may yield more immediately applicable criteria is to develop experimental paradigms for assessing whether native listeners treat tone scaling and tonal space separately, as in Herman (2000) and Gussenhoven and Rietveld (2000). Such experiments might be especially useful if paired with studies designed to pin down the meaning differences associated with minimally contrasting melodic contours where tone scaling or tonal space differences seem to act as a primary cue or as an enhancing secondary cue, as in Hirschberg and Ward (1992), Grice and Savino (1995), Venditti, Maeda and van Santen (1998), Caspers (2000), and Lee (2000, 2005). As Gussenhoven (1999) points out, however, this avenue of research requires that we look more closely at the types of linguistic functions that are linked to different formal parameters in different languages, and think carefully about how particular experimental tasks might preclude discovery of the use of some pattern of tones, tonal anchoring, tone scaling, or tonal-space settings for a particular function.

This highlights the fact that we need a better understanding of the range of linguistic functions that can be encoded in spoken language melody and of how these functions are realized in related language varieties as well as in different language families. In the next section, therefore, we will briefly describe some of the functions that have been identified, beginning with the “tonemic” function of constituting a small finite set of meaningless contrasting patterns that can be
combined with elements from other sets of meaningless contrasting patterns (consonant constrictions and vowel postures) to build an indefinitely large lexicon.

5 A taxonomy of linguistic functions

5.1 Tonemes and tonal morphemes

The basic “tonemic” function is most easily illustrated with utterances and words from a language such as the standard Hong Kong dialect of Cantonese. In this variety of Chinese, most words are monosyllabic (that is, any given zi probably is a word), and every syllable is specified for one of the tone patterns exemplified by the contrasting wordforms in (7).

(7) a. [wɐi⁵⁵] ‘power’ [wɐi³³] ‘fear; pleasant’ [wɐi²²] ‘guard’
    b. [wɐʔt⁵] ‘dense’ [wɐʔt³] ‘revolve’ [wɐʔt¹] ‘kingfisher’

Figure X.6 shows example utterances of the six wordforms with sonorant rhymes in (7a) produced as citation form sentences. The extremely low onset of the toneme that is transcribed with [⁵⁵] reflects a sound change in progress in the Hong Kong standard dialect (see So 1996, who transcribes it as [²⁵], and reviews the literature on this and other recent tone changes and merges). The black and gray f₀ tracks in Figure X.7 illustrate how the mid-level tone of the homophones meaning ‘fear’ and ‘pleasant’ is realized in two other intonational contexts. The morpheme just before [wɐi³³] ‘pleasant’ in the utterance plotted with gray in that figure also has this same mid-level tone. The pitch perturbation at the syllable boundary is a juncture-marking creaky voice quality that sets off and emphasizes the final word, which is as long as the total duration of the first four morphemes of the utterance. All of the earlier morphemes in this utterance, as well as in the utterance plotted with the black line, are shortened by a process that Wong (2006) calls “syllable fusion”. When morphemes are conjoined into compound words or frequently uttered phrases, speakers can signal the particularly close juncture by weakening or deleting medial consonants and merging the two syllables’ vowels. Except in the most extreme cases, however, the percept of each syllable’s tone specification is preserved to maintain the syllable count. Thus in this variety of Cantonese the tone specifications are contrastive properties of syllables fully on par with such properties as the palatalized offglide in the rhymes in (7a) as opposed to the glottalized plosive coda in the rhymes in (7b). The typical shapes of words in combination with the extremely “isolating” or “analytic” nature of the grammar drives a robust segmentation of the melody into syllable-anchored tone units.

The pitch patterns on the final syllable in Figure X.7 illustrate another way in which tones can function in lexical contrast. The [wɐi³³] syllable in each of these utterances is prolonged to be three or five times the average length of syllables earlier in the utterance. This prolongation leaves room for the realization of one or two more tone targets that are transcribed using the notational conventions described earlier for the transcriptions in (1)-(6). This “code-switching” between transcription systems follows the C-ToBI conventions proposed by Wong, Chan and Beckman (2005) to clearly distinguish the morphemic function of the tones transcribed as H% and HL% in these utterances from the tonemic function of the tones transcribed as [²³], [²¹], [²¹], and so on in (7). The meanings of these two morphemes H% and HL% are reflected in the glosses. The H% at the end of the
utterance in gray makes it an incredulous echo question as indicated by the ‘?!’ at
the end of the gloss, whereas the HL% at the end of the utterance in black
imparts the sense of discovery or sudden realization glossed by the ‘Oh I get it!’

Figure X.6 Example \( f_0 \) contours of citation intonation utterances of the level tone
wordforms (black) and contour tone wordforms (gray) listed in (7a) produced by
an adult female native speaker of Hong Kong Cantonese. The utterances are from
Wong, Chan and Beckman (2005).

Cantonese has an extremely rich set of pragmatic morphemes like these final
boundary tones. Many of these sentence particles are composed of vowel and
consonant phonemes as well as the tonemes, but several of them are just the
toneme affixed to the final content word, as illustrated here. The H% boundary
tone in the Beijinghua utterance in Figure X.4b, similarly, is one of two tonal
morphemes among the twenty-eight sentence particles that Chao (1968) counts.
The count for any of the Mandarin dialects is somewhat easier, since the non-
tonal components of the sentence-final particles of Mandarin are analyzed as
being neutral tone and combinations of particles are never counted separately from the particles that are simple syllables or simply tonal. By contrast, counts for Cantonese range widely (as many as 206 by Yau’s 1980 count), depending on whether polysyllabic sequences, monosyllabic particles that are potentially fused forms of polysyllabic sequences, and other complex forms are counted separately. For example, Law’s (1999) count of between 35 and 40 includes sets that are traditionally described as being minimally differentiated by tone, such as the minimal pair [tse\(^5\)] and [tseʔk\(^5\)] studied by Chan (1998). Fung (2000) suggests that Cantonese sentence particles such as these can be grouped into a much smaller number of “families” of phonologically related particles that have a common core meaning. That is, she proposes that the meanings of [tse\(^5\)] and [tseʔk\(^5\)], for example, can be analyzed in terms of the core meaning of the [ts] family in combination with the meanings of the tones (which correspond to the tonal morphemes transcribed in C-ToBI as H\% versus -\%). Sybesma and Li (2007) analyze Fung’s families in more detail, and propose that each of the forty most common sentence particles is composed of three parts: (1) an onset morpheme that is either the default null (glottal stop) initial or one of the fully specified consonants [h, g, l, m, l\~n, ts], (2) a nucleus morpheme that is either the default vowel [e:] or one of [\~o, \~\~], and (3) a tonal morpheme that is either the default [\%] (tagged as a protracted neutral target :\% in C-ToBI ), [\%\%] (tagged as -\% in C-ToBI), [\%] (H\%), or [\%] (L\%). By this analysis, then, the HL\% transcribed for the utterance plotted in black in Figure X.7 might be a compound of Sybesma and Li’s tonal morphemes [\%] and [\%], or it might be the tonal morpheme corresponding to [\%\%], which as a toneme has merged with [\%] in the Hong Kong dialect (see So 1996, among others).

The difficulty of counting the number of Cantonese sentence particles as compared to the ease of counting the nine Cantonese lexical tonemes in (7) is noteworthy. It may reflect the elusiveness of pragmatic “meaning”, which is difficult to paraphrase outside of the specific contexts where a pragmatic morpheme is appropriate, as compared to the stark difference in referential meaning that lets us identify the polysemous nature of the wordform [w\~ni\(^3\)]. It also may speak to a more basic difference between tonemes and tonal morphemes that stems from the design principle of duality of patterning (Hockett 1960) — i.e., the principle that the lexicon of any human language is a self-diversifying system in which a small number of discretely different elements can be combined to make a large number of potentially extremely complex morphemes without losing their discrete distinctiveness (Goldstein and Fowler 2003). Consider the analogous difference for vowel segments. As Hyman (this volume, p. 37) points out, it is relatively easy to count the number of tonemes in a language such as Dadibi, Nupe, Chatino, Kam, Putonghua, or Cantonese in the same way that it is relatively easy to count the number of vowel phonemes in these languages. And it is relatively easy to recognize that the tones in the second and third syllables of the first utterance in Figure X.7 are the same toneme and that both are different from the toneme on the first syllable, just as it is easy to recognize that the vowels in the two syllables of the English compound A-frame are the same vowel phoneme but that the vowels in the two syllables of A-team or AWOL are different vowel phonemes. It is harder to say whether the vowel [e:] in the first syllable of each of these words is the same morpheme, or even whether [e:] constitutes a morpheme in AWOL in the way that the [e:] in the first syllable of A-frame obviously constitutes the first morpheme of a compound word.

These two sources of difficulty have long confounded the analysis of the tonal morphemes of English. Is the tune in the second phrase of Figure X.3 a sequence of four tonal morphemes, as suggested in the transcription in (2), which follows the analysis in Pierrehumbert and Hirschberg (1990)? Or is it two tone
morphemes $H^*L H^*LH$ to which a linking rule has applied to anchor the L of the first morpheme to the second stressed syllable, as proposed by Gussenhoven (1984)? What kinds of experiments can we use to differentiate between these two morphological analyses? Ladd (2008, Chapter 4) gives an insightful description of the difficulties for English and a few other related languages, as well as a review of arguments advanced by proponents of different analyses and of the relevant experimental studies.

5.2 Prosodic grouping

In describing the “tonemic” function using Cantonese examples, we emphasized the monosyllabic word shapes and isolating morphology of the language, because the more general function of lexical contrast will be realized using very different segmentation and anchoring parameters in a language where words are polysyllabic or the grammar is of a more “agglutinative” or “synthetic” nature. For example, every modern Chinese dialect has a system of lexical tone contrasts that is a reflex of the same original tone categories that give rise to the Cantonese tonemes in (7), but in a Wu dialect, the tone pattern that corresponds to a toneme of a Cantonese word typically will not be realized in the same way on the cognate Wu morpheme. Words are typically at least two syllables, and very productive morphological processes (typically called “tone sandhi” — see Chen 2000) insure that just one toneme is specified for each compound word or phrasal construction in an utterance, as illustrated in (8).

The Shanghai examples in (8) are from Zee and Maddieson’s (1980) study, and the schematics are based on the $f_0$ tracks they show. The compound nouns in (8c) and (8d) are derived from the sets of four zi in (8a) and (8b), respectively. These examples illustrate the tone sandhi processes that relate the patterns of derived words to the citation form tone patterns of the zi from which they are derived. The most general description is what Chan and Ren (1989) call “Pattern Extension”; the underlying toneme of the first zi is the only one realized, and its component tones are extended to cover the whole word or phrase, as in (8c). All of the Wu dialects use some variant of this Pattern Extension process, although details such as the typical phonological anchoring pattern may differ across tone types and across dialects. For example, Zee and Maddieson analyze the abrupt fall in (8d) in terms of a constituent-final tone that they posit for all such compound words, whatever the initial toneme, but they do not discuss the early anchoring point for the tone in some cases, such as (8d).

   HL   LH   MH   HL
   H    HL   MH   HL
c. [ɕin.vɔŋ.tɕi.tse] ‘reporter’
   H    L    MH   L
d. [tɛInstrumental.ʃi.ɔŋ.ɹ] ‘marriage license’
   H    H    L

Kennedy’s (1953) description of a very similar abrupt fall in Tangxi compounds that have initial syllables with checked tone rhymes suggests an alternative analysis in which the abrupt fall is the realization over longer material of the creaky voice register that characterizes the checked tone. More recent work by Chen (2008) suggests that both analyses may be supported for variant realizations of longer compounds for at least some younger speakers. The cross-dialect differences in anchoring point can be appreciated by comparing the Shanghai falling-tone example in (8c) to the three Wuxi falling-tone examples in
The four examples in (10) are an alternative pattern for Wuxi compounds that begin with a falling-tone zi.

(9) a. [sɛ̃] ‘three’  b. [sɛ̃ŋiɛ] ‘3 years’  c. [sɛ̃.dʌɯ.mɔ/].dõ] ‘3 big wooden tubs’
   H   L   H     HL   H     HL

   HL   L   LH   L   LH   L   LH

The transcriptions and schematics in (9) and (10) are based on the descriptions and \( f_0 \) tracks in Chan and Ren’s (1989) account of the history of two different morphological processes that they identify in this dialect. They describe the Pattern Extension process in Wuxi as typically applying to number+classifier expressions, as in (9), and also to reduplicated verbs, verbs with resultative or directional complements, and reduplicated nouns in child-directed speech. The Wuxi “Pattern Substitution” process in (10), by contrast, is typically applied to verb phrases with direct objects, to reduplicated nouns in the adult lexicon, as well as to the very productive compound word formation process illustrated in (10), where [fi.tɕi] is ‘fly machine’, [fi.tɕi.pʰɪ.lɤ] is ‘fly machine ticket’, and [fi.tɕi.pʰɪ.lɤ.tɕia] is ‘fly machine ticket price’. Chan and Ren relate these two Wuxi processes to a contrast that Kennedy (1953) describes for the Tangxi dialect, where the morphosyntactic difference is clearer. When the two Tangxi processes apply in combining the morphemes [tso] ‘to fry’ and [ve] ‘rice’ the Pattern Extension process yields the compound noun [tso.ve] ‘fried rice’ whereas the second type of process yields the verb phrase [tso.ve] ‘to fry rice’.

Despite the differences across the examples in (8)-(10), however, the function is essentially the same. The toneme specification is a property of the constituent as a whole, and the boundaries between successive constituents are marked by a transition to the next lexically contrastive tone pattern. The contrasting melodic contours, then, effectively group strings of syllables into coherent prosodic constituents (tone sandhi groups) that align to constituents or domains specified by other parts of the grammar. When utterances are short and decontextualized, as in the utterances examined in Zee and Maddieson (1980) and Chan and Ren (1989), the domains are described in terms of morphosyntactic relationships. When utterances are longer or produced in more elaborated discourse contexts, other types of relationship, such as the articulation of an utterance into topic and focus or given and new, come into play, as discussed by Selkirk and Shen (1990) among many others.

This same function of prosodic grouping is invoked by Carter’s (1974) description of the “tone-phrasing system of Kongo” and by our description of the distribution of L\( \% \) and H- tones in Japanese in Section 4.1 above. It also is a critical element in Halliday’s (1967, p. 9) description of English utterances as “an unbroken succession of tone groups each of which selects one or another of the five tones” as well as of Pierrehumbert’s (1980, p. 19) definition of the “tune” in English as “the melody for the intonation phrase”. As should be obvious from this list of languages, as well as from the differences between Cantonese and Shanghai, the ways in which melody is harnessed for the function of prosodic grouping are orthogonal to the ways in which melody is harnessed for the function of lexical contrast. Cantonese and Shanghai have inherited the same set of tone categories from their common ancestor language, but Cantonese does not have any morphological process like these “tone sandhi” processes in Shanghai and the other Wu dialects and instead uses the consonant- and vowel-focused
Mary E. Beckman and Jennifer J. Venditti

process of syllable fusion. Thus, the surface melodies of cognate compound words and phrases make for very different tone groups in the two languages. The modern Japanese dialects offer the complementary evidence, for a double dissociation. Although the accentual phrase melodies of Japanese mark off analogous tonally delimited prosodic phrases in very similar ways in the Tokyo and Osaka dialects, the tone at the AP boundary in Tokyo is invariantly L, whereas Osaka preserves an older tonemic contrast between %L-beginning and %H-beginning words.

5.3 Metrical prominence

In accounting for the melodic differences between the disyllabic compound noun [tsɔ.^vɛ^1] ‘fried rice’ and the verb phrase [tsɔ.^vɛ^2] ‘to fry rice’ in Tangxi, Kennedy (1953) talked about the prosodic grouping function that the two patterns have in common, but he also described differences in the “stress pattern”, with the compound noun pattern having “louder stress” on the first syllable and the verb phrase having it on the second. A related segmental difference is specific to the checked tone; the glottal coda in morphemes such as [ba^3] ‘white’ or [tɕ^hʊk^5] ‘to drink’ can trigger gemination of a following syllable onset in the compound-noun pattern but not in the verb phrase pattern, as in [bas^3.sɛ^5] ‘white water’ versus [tɕ^hʊ.tʰsɛ^3] ‘to drink soup’, an interaction that is reminiscent of the stress conditions on Raddoppiamento Sintattico in many varieties of Italian (see, e.g., D’Imperio and Gili Fivela 2003).

In other Wu dialects, also, the syllable in the tone sandhi group that is associated with the toneme bears segmental hallmarks that are associated with phrasal or lexical stress in other languages. For example, Zee (1990) documents a process of vowel lenition in Shanghai whereby the high vowels of the language [i, ɨ, u] can be devoiced or deleted in certain environments. This is essentially the same process as the “syncope” that Cedergren and Simoneau (1985) describe for [i, y, u] in Quebec French, and the vowel “reduction” that Dauer (1980) describes for [i, u], which Arvaniti (1994) uses as a metric of stress on pretonic syllables. As in these other two languages, devoicing in Shanghai is a variable process that depends on speech rate as well as on the identity of the neighboring consonants. It is also tonally conditioned. In Quebec French, syncope never affects vowels in the final syllable of the constituent that Cedergren, Perreault, Poiré and Rousseau (1990) define as the domain of final pitch accent. This is the constituent that Jun and Fougeron (2002) call the accentual phrase, highlighting the demarcative function of the obligatory final pitch accent and the optional initial rise. In Greek, similarly, devoicing does not occur on the stressed syllable in a word — i.e., the syllable which is associated with one of the tonal morphemes of the utterance melody. In Shanghai, too, devoicing never affects the first syllable in the tone sandhi group — i.e., the syllable to which the toneme is associated phonologically. Chao (1968, pp. 31 & 141) also notes high-vowel devoicing in neutral tone syllables in the Beijing dialect of Putonghua, a relationship that he describes in Chao (1932, p. 129) in terms of the notion “stress-accent” or “tonic stress”:

Stress-accent does not play any important role in most Chinese dialects. But in a few dialects, including that of Peiping, tonic stress plays such an important part that unstressed syllables not only tend to have their vowels obscured, but also lose their proper tones, and acquire a level, usually short tone, the pitch being determined by the preceding syllable. Thus, in all four of these languages, the property of being eligible to bear an associated toneme or tonal morpheme prohibits application of a process that
weakens or deletes vowels. This is true both of Beijinghua and Greek, where the location of this “tonic stress” is not predictable from the prosodic grouping into words and phrases, and of Shanghai and French, where the fixed position of the “tonic stress” serves to demarcate the tone sandhi group or accentual phrase.

By contrast, high-vowel devoicing is not constrained by the tone pattern of the accentual phrase in either Japanese or Korean. Maekawa and Kikuchi (2005) observe devoiced vowels in the Corpus of Spontaneous Japanese in many syllables that are aligned to the phrasal H- in unaccented phrases or associated to the H of the H+L lexical pitch accent. Jun and Beckman (2002) likewise document pervasive high-vowel devoicing in a corpus of enacted lab speech dialogues of Korean, both in syllables that are AP-medianal and in syllables that are associated to the LH or HH sequence that marks the beginning of the AP.

This reduction of vowels in the first syllable of the AP in these two languages contrasts with other segmental effects in this position. In Japanese, for example, older speakers who produce the nasal allophone of [g] in AP-medianal positions (such as in the –ga particle in the Yamano- wa oyoideru-ga clause in Figure X.1b) do not produce [ŋ] in AP-initial position. Keating, Cho, Fougeron and Hsu (2003) and others show that the beginning of the AP in Korean also is a position marked by “initial strengthening” of the consonant. The consonantal effects are more in line with the segmental effects of metrically strong position in languages with tonic stress. In Shanghai, for example, voiced stops are voiceless with breathy voice releases when they are onsets of syllables at the beginning of a sandhi tone group, but are voiced with short closures in tone-group medial position where the syllable does not bear a tone specification (Cao and Maddieson 1992). That is, they show the same allophonic patterns with respect to the tone sandhi group that Jun (1993/1996) and others document for the Korean lax stops in AP-initial versus AP-medianal position. There are similar effects in syllables with tonic stress versus syllables with neutral tone in the Beijing dialect. For example, the voicing of the [ts] in the second syllable of fàng zài in Figure X.4a is a cue to the neutral tone status of the syllable. One possible way to characterize the different treatments of the vowel in Japanese and Korean as compared to these other four languages, then, is to say that the vowels are less important than the consonants in defining the syllables and the rhythms of anchoring points for tones in these two languages. Another way to characterize the difference is to say that Japanese and Korean emphasize edges at all levels of metrical structure, from the consonant-focused definition of the syllable to the primarily demarcative use of tonemes and tonal morphemes, whereas Beijing Mandarin, like English, emphasizes heads.

This difference has ramifications for the realization of focal prominence. In Korean and Japanese, focal prominence is realized primarily by an expansion of the tonal space to enhance the demarcative rise at the beginning of the first AP coupled with a post-focal erasure of AP boundaries (see, e.g., Venditti, Jun and Beckman 1996). Other prominence-enhancing mechanisms include the choice of IP-final boundary tones such as the H% tone of Tokyo Japanese (see, e.g., Venditti et al. 2008). In Beijing Mandarin, English, and Swedish, by contrast, focal prominence instead singles out a syllable with tonic stress and then either reduces or deletes the tones associated to following stressed syllables (see Jin 1996, Xu and Wang 2001, among many others, for Mandarin, Chapter 6 of Ladd 2008, for a review of the literature on English, and Bruce 1977, 1982, among many others, for Swedish). Chapter 7 of Ladd (2008) gives a particularly insightful discussion of this difference between edge-focus and head-focus strategies. He also suggests a common underlying unity. The syllable with tonic stress in languages such as English and Swedish plays a culminating role in marking words and larger morphosyntactic constituents, as illustrated by the tone pattern that marks the compound word in (5). The word that is the domain of the focal
H- in Stockholm Swedish, similarly, plays a culminating role in marking intonation phrases and their alignment with the domains of focus in the information structure of the sentence. Pierrehumbert (1980) posits a similar “phrase accent” for English, as in the transcription in (2). Gussenhoven (1984), by contrast, treats the rise-fall-rise over messenger and ball as a single H*LH tonal morpheme. By either analysis, however, the word that contains the stressed syllable to which the H* tone is associated plays a similarly culminating role viz-a-viz the focus domain in English. Ladd (2008, p. 278) suggests that both the culminating function and the demarcative function can be viewed as ways of identifying levels of grouping in the metrical hierarchy of a language:

If ... we see prosodic phrasing as the ultimate basis of sentence stress, we may see that the correct way to pose questions about universals of the prosody-focus link is not ‘Why is the main accent in this sentence on word X rather than on word Y?’ but rather ‘Why is this sentence divided up into phrases the way it is?’

A further advantage of thinking of “sentence stress” in this way is that focus and other aspects of information structure at the level of the sentence then become the local expression of the same types of discourse structure relationships that are encoded in such effects as final lowering, as discussed in Nakatani (1997) and Venditti (2000).

6 Defining tone

As the preceding section should make clear, it does not seem very useful to talk about “stress” as if it were an autosegmental content feature, on par with tone features, manner features, and place features. Rather, stress is better treated as a syntagmatic property of nodes in the prosodic tree, like the property “syllabic” (which is another way of calling the autosegment that stands as the head of some o). In many languages, stressed nodes in the prosodic tree are defined first and foremost by the licensing of tonal autosegments. In some language varieties, such as Hong Kong Cantonese, syllabic nodes also are so defined (cf. Wong et al. 2003, and the discussion of “syllable fusion” above). In many languages, tones also are stereotypically used to mark the edges of prosodic constituents above the prosodic word, a function that is less commonly associated with vowel or consonant features. Tone is remarkably versatile in the roles that it plays in realizing and interpreting the prosodic trees of spoken languages. As Hyman (this volume, p. 35) [emphasis in the original] puts it, “Tone can do everything that segmental and metrical phonology can do, but the reverse is not true.”

This versatility raises again the question with which we started this chapter: How can we define intonation in a way that delimits the scope of this chapter from the scope of the chapter on tone? Or focusing the question the other way around: What is tone, and how does it differ from intonation?

In his chapter on tone in this volume, Hyman (this volume, p. 22) poses the first part of this question, but then replaces it with a question about language types: “How do we know if a language has tone?” In answer, he repeats his earlier “working definition of tone” (Hyman 2006, p. 229), which is adapted from Welmers’s (1959) definition: “A language with tone is one in which an indication of pitch enters into the lexical realisation of at least some morphemes.” He rejects a distinction between “pitch accent” and “tone” (corresponding to Voorhoeve’s distinction between “restricted” and “unrestricted” tone systems described above). In his earlier paper, Hyman points out that systems cited as examples of the “pitch accent” type are a varied lot, including languages as different as Tokyo Japanese (where “accent” does not imply metrical prominence and the majority of native Yamato-stratum words are unaccented) and Stockholm Swedish (where every word has at least one syllable with tonic stress and a compound word has
exactly two). While he rejects the idea of this third type, however, he maintains that a tenable distinction can still be made between a “tone language” prototype and a “stress-accent language” prototype. His criteria for setting up this distinction require that he treat stress as a “suprasegmental” property on par with H tone, rather than as a structural property on par with syllabicity. That is, he proposes that the prototype stress-accent system is one in which “every word has at least one stress accent” and “the stress-bearing unit is necessarily the syllable.”

We cite Hyman (2006) here because this paper is very representative of a widely-held assumption: that there are fundamental prosodic differences among spoken languages which naturally fall out from the difference between using tone “to make semantic distinctions” and using it “to add functional meaning.” This assumption is at the heart of nearly every typology of tone and intonation systems. The difference that is deemed critical in these typologies is a distinction between the tonemic function of lexical contrast and everything else — between languages such as Cantonese, where many of the tones in the melody of a typical utterance are tonemes that combine productively with the consonant and vowel phonemes of the language to make a large and expandable set of morphemes, and languages such as English, where the tones are pragmatic morphemes chosen from a small and relatively closed set. This is a useful distinction, because it predicts that there would be sharp differences in native speakers’ and linguists’ metalinguistic awareness of the tone count, as suggested in Section 5.1. However, contra Hyman (2006) we do not see that it correlates neatly with all of the other distinctions that could be made on the basis of the functions outlined in Sections 5.2 and 5.3. That is, we can appreciate the difference in ease of counting tones in Putonghua versus English that falls out from the fact that a L+H that is anchored to a stressed syllable in Putonghua is a toneme whereas a L+H* that is anchored to a stressed syllable in English is a pragmatic morpheme. But this difference does not change the fact that these two languages are far more like each other in many other respects than either is to a language such as Japanese. There is no useful classification of prosodic types that falls out from the classification of languages in terms of the tonemic function alone.

Hyman begins his chapter on tone by saying that, “Except for a brief period in the late 1970s and early 1980s, tone has generally fallen outside the central concerns of theoretical phonology” (p. 17). He ends it by calling for renewed attention to questions about “the interdependency of tone with other features” — questions such as “What is or can be a TBU?” and “What are the correct set of tonal features?” (p. 362). He concludes by saying that “the above set of questions may even contain a misunderstanding that we still have either about tone, or more likely, about phonology in general.” (p. 377)

As we have tried to make clear in this chapter on intonation, we think that, contrary to Hyman’s assessment of the last three decades, tone qua intonation has been very central to major developments in theoretical phonology, and that work on intonational phonology has addressed the questions about interdependencies and proposed answers that suggest that the questions do indeed “contain a misunderstanding” about the nature of tone. That is, the work that we reviewed in Section 4 suggests to us that the presuppositions of these questions constitute the tonal counterparts to the assumptions of “segmental idealization and “universal categorization” in segmental phonology, as discussed by Ladd (this volume). In short, we suspect that defining tone (and delimiting intonation from it) in terms of the tonemic function alone may have delayed progress in our understanding of tone. It may have obscured the true diversity of ways in which speakers of different languages use pitch variation to structure their words, utterances, and larger spoken discourses. We close by reminding the reader that the languages for which we have a good solid description of the system of lexical
tone contrasts vastly outnumber the languages for which we have even a cursory description of the tone patterns that are associated with other levels of the grammar. Until that gap in coverage is filled in a bit more, any delimitation of tone from intonation based on a classification of language types seems premature.

References


Carlson, Rolf, and Björn Granström (this volume). Speech synthesis.


Hyman, Larry M. (this volume). Tone: Is it different?


MacNeilage, Peter F., and Barbara L. Davis (2000). Deriving speech from nonspeech: A View from ontogeny. Phonetica, 57, 284-296.


Notes

1Sections 3-5 of this chapter are very nearly identical to sections 2-4 of our chapter on “Tone and intonation” in the second edition of the Handbook of Phonetic Sciences (Hardcastle, Laver, and Gibbon, eds., 2009). The material covered in these sections is reused here with the concurrence of the editors of this volume, since it supports the different mandate of this chapter (which was to review the contributions that intonational phonology has made to phonological theory) just as well as it supported the mandate of the other chapter (which was to lay out the bases of a taxonomy of forms and functions of linguistically significant pitch variation).

2Chao’s tone letters locate notes and glissando turning points in the local tonal space in terms of five points numbered from 1 for the bottom to 5 for the top of the tone space.

3The traditional use of the term “accent” both for pragmatic tonal morphemes in English and for lexically specified tone patterns in Japanese is the source of frequent confusion among scholars of both languages. Further confusion is caused by the fact that the Japanese word akusento which ‘accent’ translates here refers to the entire configuration of tones for the level of prosodic grouping that is called the accentual phrase, including both the lexically specified pitch fall at the designated syllable and the “post-lexically” specified pitch rise at the AP beginning. See Venditti, Maekawa, and Beckman (2008) for an explication of the differences between the two phenomena.

4See also Table (66) in Hyman (this volume), where the fully-specified account is assumed and the surface tone string transcribed with Africanist tonal diacritics.

5While we focus on the tonal aspects of the definition here, each of these levels of prosodic grouping is also marked by segmental effects such as differing degrees of “initial strengthening” and “phrase-final lengthening (see review of this approach to these phenomena in Fletcher 2009).

6Hyman (this volume, p. 287) analyzes these patterns in term of tone-spreading to result in a fully specified surface tone pattern, but this is not the only possible analysis (see Yip, 1989).

7Following Zee and Maddieson (1980), we transcribe the apical vowel here and in the examples with [ɹ] rather than with the non-IPA symbol used in the Sinological literature.