This is a section of doi:10.7551/mitpress/10413.001.0001

Prosodic Theory and Practice

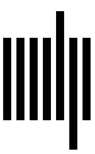
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel

Citation:

Prosodic Theory and Practice
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel
DOI: 10.7551/mitpress/10413.001.0001
ISBN (electronic): 9780262543194
Publisher: The MIT Press
Published: 2022

OA Funding Provided By:

OA Funding from MIT Press Direct to Open



The MIT Press

11 The PENTA Model: Concepts, Use, and Implications

Yi Xu, Santitham Prom-on, and Fang Liu

11.1 Introduction

Speech is a communication system for transmitting information from one human being to another.[1] The information transmitted is rich and multifaceted, but it is coded by an articulatory system in such a way that the listener can readily decode it. These facts, which may seem too obvious to be worth restating, are the premise of the articulatory-functional view of speech that forms the basis of the parallel encoding and target approximation (PENTA) model (Xu 2005). PENTA is therefore a theory of how multiple layers of information are effectively conveyed through prosody with a neurally controlled biomechanical system. In other words, PENTA is about how prosody works as a communication system, how it can be learned, and how it goes through changes over time—in short, how it operates. The mission of PENTA therefore differs from those of many other theories that focus on directly accounting for observed prosodic forms. By focusing on operation as its primary goal, PENTA accounts for prosodic forms only as a by-product, rather than as an end in itself.

11.2 The Conceptual Framework

Beyond the basic facts just stated, PENTA makes a number of assumptions that are highly hypothetical. The first is that syllable-synchronized sequential target approximation (the TA part of the model) is the rudimentary mechanism of speech prosody, based on which all of the information coding is done. The second is that prosody conveys multiple layers of information simultaneously, through encoding schemes that are in parallel to each other, that is, without a hierarchical structure (the PE part of the model). Third, the phonetics of the encoding schemes are specified parametrically rather than based on symbolic representations. Due to their hypothetical nature, each of these assumptions needs independent justifications, which we will provide after a brief sketch of the model.

Figure 11.1 is a schematic of PENTA in its most general form, representing not only prosody but also other aspects of speech (Xu and Liu 2012). The first block from the left represents communicative functions that are conveyed by speech. The functions are parallel to each other, as illustrated by the nonhierarchical stacking in the schematic. The second block represents encoding schemes that are associated with the communicative functions. The schematization here makes it clear that communicative functions do not control surface acoustics directly but through a set of specific encoding schemes. The encoding schemes can be highly stylized and language-specific or more gradient and universal. The third block represents the articulatory parameters that are

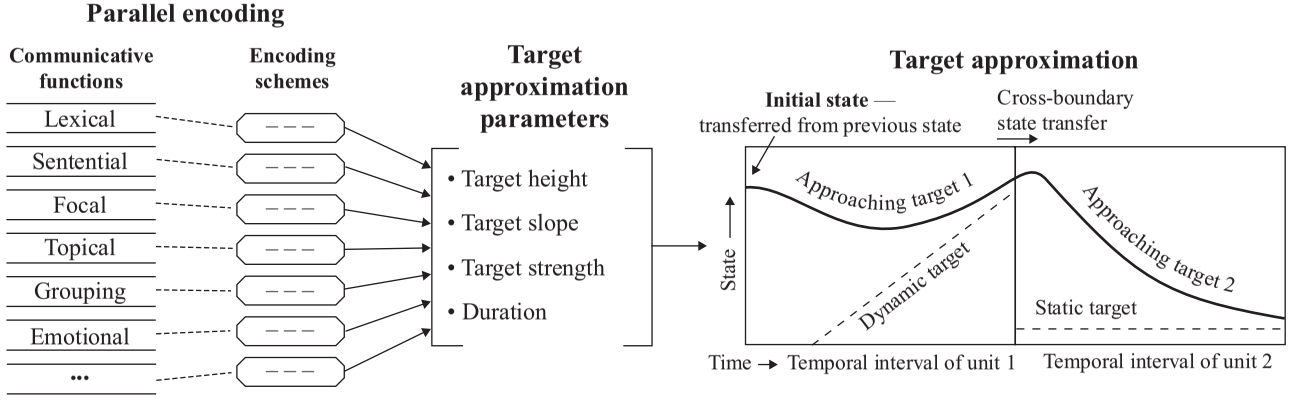


Figure 11.1
A schematic sketch of the PENTA model (Xu 2005; Xu and Liu 2012; Xu and Wang 2001).

specified by the encoding schemes. These parameters control the articulatory process of target approximation represented by the fourth block. This biomechanical process ultimately generates surface acoustics.

The TA model, as depicted in the rightmost block in figure 11.1, assumes that each syllable is assigned an underlying pitch target specified in terms of not only height but also slope. The surface contour is then the result of sequential approximations of successive targets, each articulated in synchrony with a syllable. At the boundary between adjacent syllables, the final articulatory state of the earlier syllable is transferred to the next syllable. Such transfer often results in a delay of the surface alignment of a turning point. The model therefore has no specifications for the temporal alignment of surface turning points, unlike in some other models (e.g., Hirst 2005; Pierrehumbert 1980; Taylor 2000).

11.2.1 Articulatory Mechanisms

Like the theory as a whole, the articulatory aspect of PENTA also starts from self-evident facts. One of the most basic is that, given the need to use different articulatory states to represent different prosodic components (because information has to be coded by differential representation), transitional movements between the states are inevitable, and each movement takes time. Two empirical questions need to be answered about these movements:

Question 1: How much time does each take? If the amount of time is largely negligible, then there is little need to take it seriously in theoretical considerations.
Question 2: What is the manner of the transitional movements? Knowledge about this may help to explain many details in the observed surface acoustics.

The second basic fact is that the laryngeal movements that generate F_0 contours have to co-occur with supralaryngeal movements that generate segments, and the two movements have to be coordinated in time. For this, there are two empirical questions:

Question 3: What is the basic mechanism of the temporal coordination?
Question 4: How much freedom does the speaker have in this kind of coordination?

With regard to question 1 (how much time the F_0 movements have to take), the empirical findings (Sundberg 1979; Xu and Sun 2002) have shown that the time needed for F_0 movements due to change of laryngeal state is not negligible. The following quasi-linear relations are found between the size of F_0 movement and the mean minimum time it takes to complete the movement:

$$\textit{Rise: } t_r = 89.6 + 8.7\ d \tag{11.1}$$

$$\textit{Fall: } t_f = 100.04 + 5.8\ d \tag{11.2}$$

where d is the F_0 excursion size in semitones, and t is the duration in milliseconds. These equations show that it takes about one hundred milliseconds to make even the smallest pitch movement, rising or falling. Given that the average speech rate is about five to seven syllables per second, meaning that each syllable takes about 143 to 200 milliseconds, a significant portion of each syllable has to be used for pitch transitions.

Regarding the manner of the transitional movements (question 2), systematic examination of lexical tones in Mandarin produced in connected speech has provided

relevant clues. As can be seen in figure 11.2, when a tone is preceded by different tones, the corresponding F_0 contours all asymptotically approach a linear trajectory that is characteristic of its underlying properties: high level for the high tone, rising for the rising tone, and falling for the falling tone. Cross-tonal transitional movements can therefore be characterized as asymptotic approximation of the underlying target (Xu and Wang 2001).

As for the basic mechanism of the temporal coordination of laryngeal and supralaryngeal movements (question 3), figure 11.2 also provides relevant clues. We can see that the divergence between different tones is the largest at the beginning of the syllable and smallest at the end of the syllable. This indicates that the movement toward a tonal target starts from the onset of a syllable and ends at its offset. Thus, the target approximation movement of a tone seems to be synchronized with respect to the syllable. In other words, each target approximation movement occurs strictly within the syllable that carries the tone. These observations have led to PENTA's core assumption about the basic articulatory mechanism of F_0 production: syllable-synchronized sequential target approximation (Xu 2005).

There are actually two aspects to this basic articulatory mechanism: pitch target in every syllable and syllable-synchronized sequential target approximation. "Target in every syllable" means that even if there appears to be a global contour over more than one syllable, each syllable still needs to be assigned a pitch target. This is based on the articulatory consideration that it is physically impossible, as illustrated in figure 11.3, to first generate a pitchless syllable (lower left panel) and a carrierless F_0 contour (upper left panel), and then combine them to form the full surface acoustic signal (right panels). It is imaginable that the F_0 contours in the upper left panel and the formant trajectories in the lower left panel could be first formed in the brain, and then the articulation process faithfully reproduces them during articulation. In fact, as already mentioned, much of the F_0 movements result from articulatory transitions between the ideal pitch targets, and their slow speed (relative to the syllable) is due to the limit of maximum speed of pitch change. The same is shown to be true of formant movements as well (Cheng and Xu 2013). Thus, the F_0 and formant contours in figure 11.3 are mostly a by-product of physical inertia. Therefore, if the transitions were already represented in the brain-generated F_0 and formant commands sent to the muscles, the effect of inertia would be applied twice!

Thus, it is unlikely that continuous surface F_0 contours are generated independent of the segmental events and then added to the segmental string during articulation. Instead, it is more likely that, at the control level, each syllable is specified with all the underlying articulatory goals associated with it, including segmental targets, pitch targets, and even phonation (voice quality) targets. This is illustrated in the left block of figure 11.4 for pitch and formants. Here the formant patterns are representations of the corresponding vocal tract shapes that are presumably the actual targets. The articulation process then implements all the targets in tandem, all through target approximation (top right). This biomechanical process ultimately generates continuous surface F_0 and formant trajectories (bottom), which consist of mostly transitions toward the respective targets. Thus, every syllable, before its articulation, would have to be assigned both segmental and suprasegmental targets as control signals for the articulatory system. And, importantly, the effects of inertia are applied only once: during articulatory execution, the final stage in the production chain.

Further support for pitch target for every syllable comes from the finding that not only stressed syllables, but also unstressed syllables in English and the neutral tone

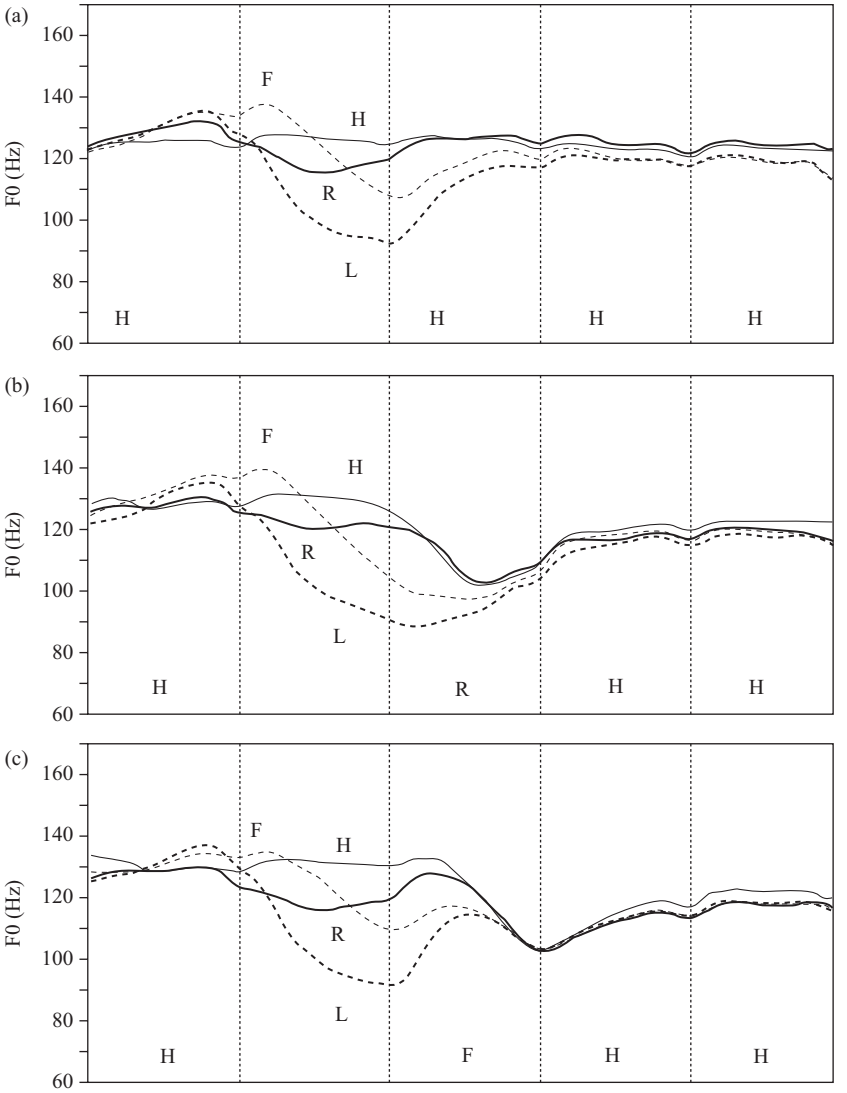


Figure 11.2
Mandarin tones produced in various tonal contexts. These are mean time-normalized F_0 contours of 猫咪/迷/米/蜜 摸/拿/卖 猫咪 (Kitty/Cat-fan/Cat rice/Cat honey strokes/picks up/sells kitty). Each contour is an average of forty tokens said by four male speakers of Beijing Mandarin (five repetitions by each). In all three plots, vertical lines indicate syllable boundaries. The tone names H, R, L, and F stand for high, rising, low, and falling tones, respectively. *Source*: Adapted from Xu (1999).

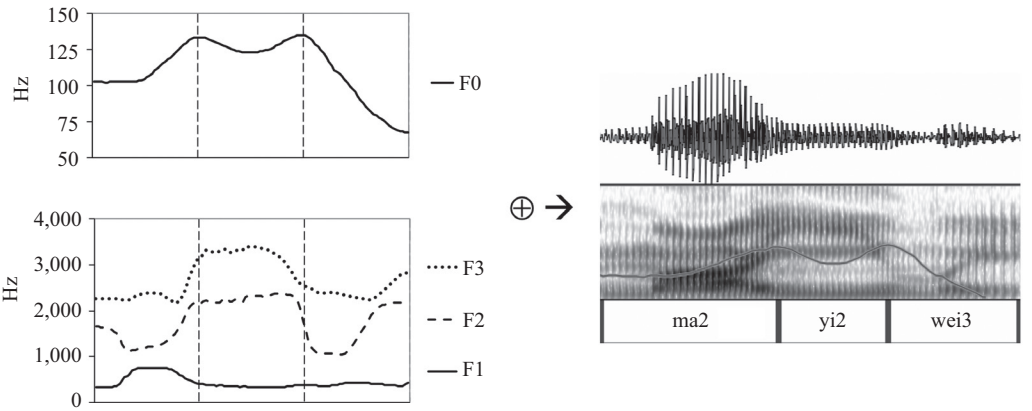


Figure 11.3
(Left) Continuous F_0 (top) and formant (bottom) tracks of the Mandarin utterance *(bi3) ma2 yi2 wei3 (shan4)* (More hypocritical than Aunt Ma). (Right) Waveform, spectrogram, and F_0 track of the same utterance. *Source*: Raw data courtesy of Xu (2007).

in Mandarin, show signs of functionally contrastive pitch targets. For example, when a word was focused in English, the unstressed syllables following a stressed syllable actively lowered pitch, as if they were part of postfocus domain. But the lowering is not as fast as in a postfocus stressed syllable (Xu and Xu 2005). This indicates that unstressed syllables are assigned postfocus targets, but with a weak target approximation strength. As found in both acoustic analysis (Chen and Xu 2006; Xu and Xu 2005) and computational modeling (Liu et al. 2013; Xu and Prom-on 2014), such weak strength can account for the high variability (and hence a seeming lack of target) of the pitch of the unstressed syllables in English and the neutral tone in Mandarin.

The other hypothesized aspect of the pitch production mechanism is the syllabic synchronization of target approximation movements. That is, laryngeal and supralaryngeal movements are synchronized with respect to the syllable (Xu 2020; Xu and Liu 2006, 2012). This assumption is motivated not only by observations like those shown in figure 11.2, but also by findings that the motor system is able to coordinate multiple movements at a fast speed only through full synchrony (Kelso 1984; Kelso, Southard, and Goodman 1979; Mechsner et al. 2001). This synchrony constraint could be due to a general problem in motor control, that is, the high dimensionality of the motor system makes the control of any motor action extremely challenging (Bernstein 1967; Latash 2012). Bernstein (1967) proposes that this problem can be alleviated by functionally freezing degrees of freedom (DOF) during motor learning. The freezing of DOF is analogous to allowing the wheels of a car to be controlled by only a single steering wheel. This makes the movements of the wheels fully synchronized and their DOFs merged. In the case of speech, it is possible that the syllable is an evolved motor synchronization mechanism to solve the problem of multigestural encoding of information. As conceptualized in the synchronization model of the syllable (Xu 2020; Xu and Liu 2006, 2012), all the articulatory target approximation movements are synchronized with respect to the syllable, except that syllable-initial consonants complete their movements earlier than both the vocalic and laryngeal movements.

There are a few issues that the current version of the TA model, especially its quantitative implementation via qTA (Prom-on, Xu, and Thipakorn 2009), has not yet fully

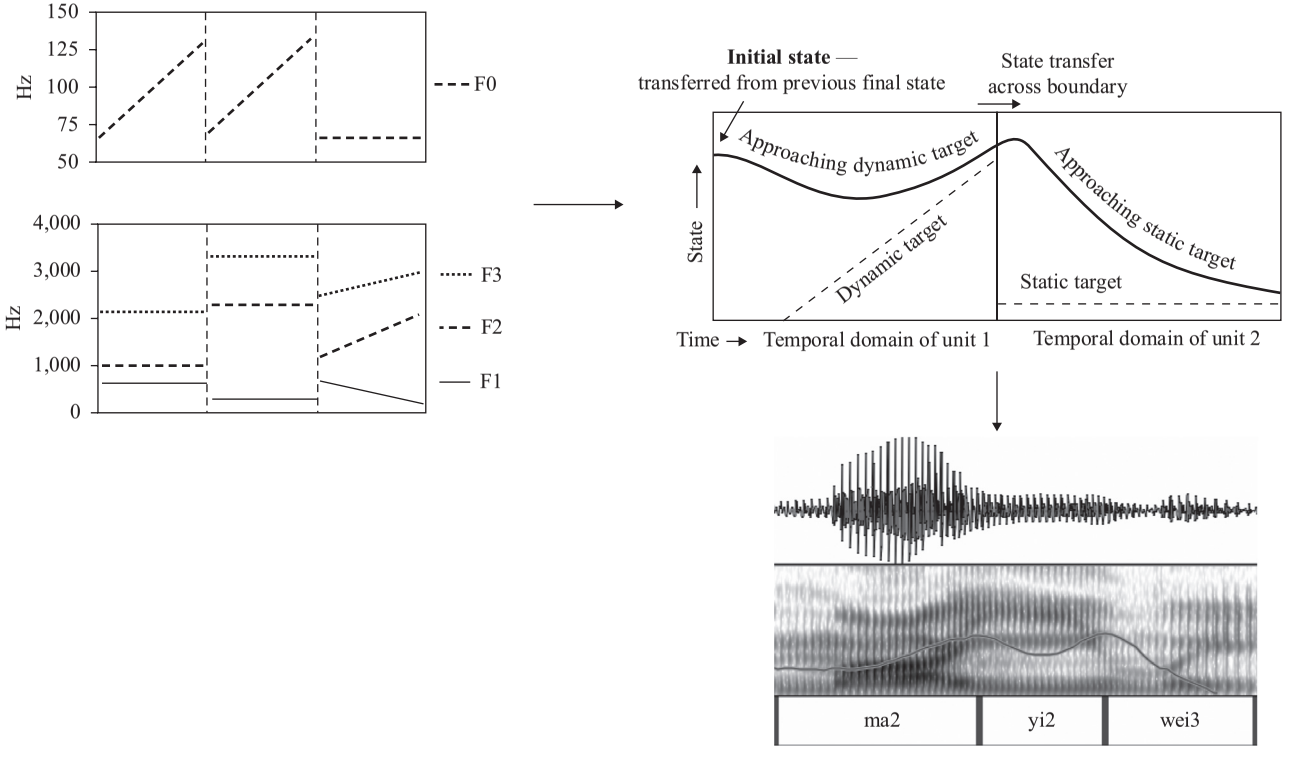


Figure 11.4
(Left) Hypothetical underlying pitch (top) and formant (bottom) targets for the Mandarin utterance shown in figure 11.3. (Right) The TA model (Xu and Wang 2001). (Bottom) Waveform, spectrogram, and F_0 track of the same utterance that are presumably generated with the underlying targets through an articulatory process similar to the TA model. *Source*: Raw data courtesy of Xu (2007).

addressed. The first is the delay in target approximation when a target is dynamic. It has been found that in Mandarin, the final slope of a rising or falling tone remains consistent when syllable duration is excessively long (Xu 1998, 2001). A similar finding has been reported for the final slope of formant trajectories in English diphthongs (Gay 1968). A possible articulatory mechanism is that speakers have the ability to adjust their articulatory effort over time to achieve consistent movement velocity by the end of a syllable. But this mechanism has not yet been implemented in the current version of qTA.

The second issue is the possibility of having two consecutive pitch targets within one syllable. An example is that the citation form of the low tone (tone 3) in Mandarin has a rising movement following the initial low-approaching movement. Although the entire shape looks like that of the rising tone, which also has an initial dip followed by a rise, the rising tail in the low tone is entirely missing when the tone is followed by any other tone. Thus, the final tail of the low tone is optional, and the underlying target is unlikely to be similar to that of the rising tone. Rather, it is possible that the citation form of the low tone consists of a low target followed by either a mid/high or rising target. Such consecutive targets in a syllable would still obey the synchronization constraint because the onset of the first target and the offset of the second one would coincide with the syllable onset and offset, respectively. Like delayed target approximation, the mechanism of consecutive targets also has yet to be implemented in qTA.

The third issue is the phenomenon of postlow bouncing. That is, after a tone with a very low pitch, the F_0 of the following syllables sometimes "bounces up" before returning to the same level as other tones. The bouncing is most readily observed if the postlow tone is weak, as when it is a neutral tone (Chen and Xu 2006). When the following tone is not weak, the effect is observable only when the low-pitched tone is focused. We have speculated that such bouncing is due to a temporary loss of antagonistic muscle balance when the extrinsic laryngeal muscles, which are engaged mainly in pitch lowering (sternohyoids, sternothyroids, thyrohyoids), suddenly stop contracting, resulting in an abrupt increase of vocal fold tension since the cricothyroids, the only muscles that lengthen the vocal folds (Zemlin 1988), are in contraction. We have been able to simulate postlow bouncing by increasing the amount of acceleration (second derivative of F_0) in the transferred state at the junction between the low tone and the following tone. The simulation worked well (Prom-on, Liu, and Xu 2012) but has not yet been included in modeling tools such as qTAtrainer and PENTAtrainer.

Closely related to postlow bouncing is the fourth issue: prelow raising. This is a phenomenon (also known as anticipatory raising, anticipatory dissimilation or H raising) whereby a tone or a pitch accent with a low pitch component raises the pitch of the preceding syllable (Connell and Ladd 1990; Hyman 1993; Gandour, Potisuk, and Dechongkit 1994; Laniran 1992; Laniran and Gerfen 1997; Lee, Xu, and Prom-on 2017; Xu 1997). Given that it is found in a number of unrelated languages, the phenomenon is likely due to a universal articulatory mechanism, although its exact nature is not yet clear. Our current theory is that it is a preparation for a sharp F_0 lowering movement, which is similar to drawing back the arm in preparation for throwing an object over a long distance. Prelow raising is not explicitly incorporated in qTAtrainer or PENTAtrainer either. However, its effect can be partially simulated by allowing tonal pitch targets to be conditioned by the upcoming tone in computational modeling.

Finally, it is well known that F_0 is temporally perturbed (mostly upward) at the voice onset after an obstruent consonant (Hombert 1977; Silverman 1986). The perturbation consists of a very brief (lasting about thirty milliseconds) transient effect (Xu and Wallace 2004; Xu and Xu 2003) and an optional longer-lasting vocal tension effect

(Stevens 1998; Xu and Wallace 2004). Again, these mechanisms have not yet been incorporated into our computational tools.

However, both qTAtrainer and PENTAtrainer do allow (and actually recommend) users to set the entire syllable, including any voiceless consonant, as the domain of target approximation. This strategy has generated F_0 contours that better match those of the original than when target approximation domain is limited to the rime only (Xu and Prom-on 2014).

11.2.2 Communicative Functions

Again following the principle of starting from basic facts, we first recognize that it is indisputable that many meanings are conveyed through prosody. But questions are open as to what exactly those meanings are and how each of them is prosodically coded. One of the key issues is whether meanings are coded directly, or through the intermediary of a prosodic phonology. For this there is a need to go at least as far back as the Saussurean view of linguistic sign (de Saussure 1916). The Saussurean view emphasizes the distinction between meanings and their bearers, which are meaningless, that is, between signifier and signified. Linguistic units in speech are therefore signified-signifier or function-form unities. Although this view is already deeply entrenched in many aspects of modern linguistic theories, the difficulty of prosody research suggests that further theoretical clarifications are needed. The first is that, of the two sides of the signified-signifier coin, is the signified—the communicative function, or the signifier—the phonetic form, that should provide the primary definition of a prosodic category? This point may not appear critical in the case of lexical items, because the function-form unities are relatively clear, given the easy delineation of lexical items. In prosody, however, the function-form unities are almost never clearly delineated. In this case, should function or form be given primacy when defining prosodic units? PENTA, similar to some other function-oriented models (e.g., Bailly and Holm 2005), holds that, when in doubt, it should be function rather than form that is the primary guide to the delineation of prosodic units.

A further complication of prosody is the multiplicity of meanings it conveys, which the classic Saussurean view again cannot easily address. For example, in the case of tone languages like Mandarin or Swedish or a pitch accent language like Japanese, pitch is used to form tones or pitch accents, which serve to distinguish words that are segmentally identical. But pitch is also used to encode focus that highlights a particular component of an utterance (Bolinger 1972), and sentence type that indicates whether the speaker is making a statement or asking a question (Eady and Cooper 1986). It is therefore not easy to identify through direct F_0 observations, as typically done in form-first approaches (e.g., Pierrehumbert and Hirschberg 1990), which parts are the signifiers of the three meanings, respectively.

Based on our recognition of these inherent difficulties, the development of PENTA has followed a two-pronged strategy. First, we have persistently tried to establish function-form unities one at a time by starting from a reasonably clear functional definition of each category and then empirically discover (or rule out) the prosodic means used by each language to encode it. As will be seen next, the prosodic means are not limited to temporally separated units as in the case of lexical items. Rather, they are often encoded by modifying existing forms that are already specified by other functions. For example, the basic shapes of a pitch contours of individual syllables are already specified by lexical tones in a tone language. So focus and sentence type can only be encoded by modifying these basic shapes in one way or another.

Second, because of the obscurity of the function-form relation in prosody, we have always tried to empirically discover, rather than either stipulate for theory-internal reasons or conceptually assume, the exact means with which each communicative function is encoded in prosody. For each function, reliable methods of triggering its occurrence and controlling its categorical contrasts are established, and experiments are conducted to uncover the phonetic realization of the functional contrasts, following all the necessary precautions needed for proper empirical studies, including, in particular, the use of minimal pairs and attention to details (Xu 2011b).

This has been facilitated by a method that has allowed us to systematically compare continuous F_0 contours in fine detail (Xu 2011b, 2013). The key component of the method is the extraction of time-normalized continuous F_0 contours. Time normalization takes the same number of F_0 measurements from each equivalent temporal interval, for example, the syllable, regardless of the interval duration. The time-normalized contours can then be averaged across repetitions and even across speakers. Such averaging removes variations from individual utterances that may mask the common characteristics that we look for. For example, the target approximation characteristic of tonal realization is brought to a clear view by the averaged F_0 contours in figure 11.2 because the contextual variants of the same tone with complete trajectories are pitted against each other in the same graph.[2]

These strategies have allowed us to see how it is possible for F_0 to encode multiple communicative functions simultaneously. As can be seen in figures 11.5 and 11.6, in both Mandarin and English, focus and sentence type can be realized on top of the lexical functions of tone and word stress by modifying the height and shape of local F_0 contours. These sentences were spoken with sentence-initial (in figure 11.5 only), sentence-medial, or sentence-final focus and as either statements or questions. As can be seen, the realization of focus and modality exhibits F_0 patterns that are best described in terms of their interactions both with each other and with lexical stress:

1. Focus is characterized by a robust postfocus pitch-range shift, with the direction of the shift dependent on modality as well as language. It is downward in statements in both languages. In questions, the shift is upward in English but downward in Mandarin. In the latter case, however, the downward shift is less sharp in questions than in statements, thus allowing Mandarin to still distinguish statements from questions under focus.
2. In English, focus and modality interact with lexical stress to determine the microproperties of syllabic pitch targets. For on-focus word-final stressed syllables, the target slope is falling in statements but rising in questions (*job* in figure 11.9b and d). For on-focus, nonfinal stressed syllables, the target slope is rising in both statements and questions.
3. In Mandarin, the contribution of modality is most clearly seen from the location of focus, except in the sentence-final word, where the modality difference can be seen even when the sentence is focus-neutral.

The principle of empirical guidance in establishing the encoding schemes has also allowed us to see that not all prosodic functions have to be chiefly encoded by F_0. The function of boundary marking (or grouping, phrasing), for example, is primarily marked by timing properties, which include both syllable and pause duration. In general, a longer preboundary duration signifies greater separation of two adjacent groups (Wagner 2005; Xu and Wang 2009). Toward the end of a sentence, word duration

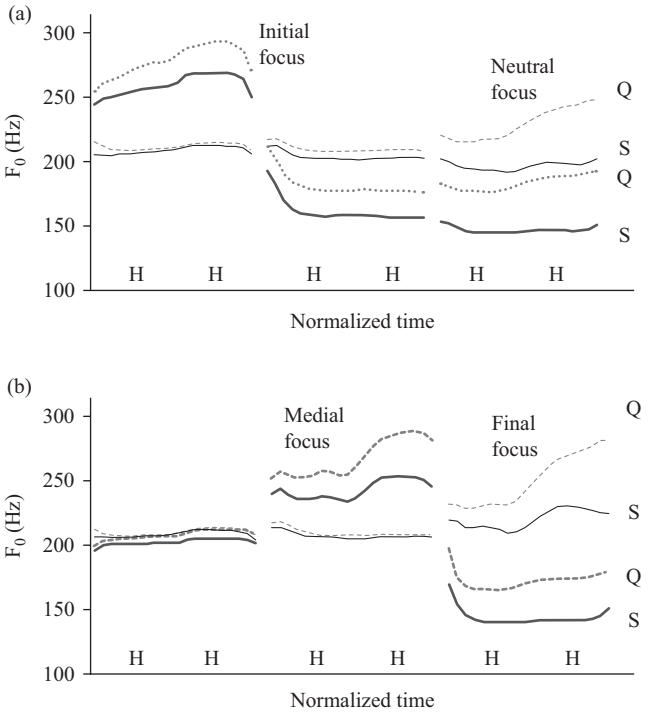


Figure 11.5
Mean F_0 contours of Mandarin sentence "Zhāng Wēi dānxīn Xiāo Yīng kāichē fāyūn" (张威担心肖英开车发晕; Zhang Wei is concerned that Xiao Ying may get dizzy when driving) spoken as either a statement or a question. (a) Either focus is on the sentence-initial word (thick lines) or there is no narrow focus (thin lines). (b) Focus is either sentence medial (thick lines) or sentence final (thin lines). The solid lines represent statements, and the dashed lines represent questions. S, statement; Q, question. *Source*: Data courtesy of Liu and Xu (2005).

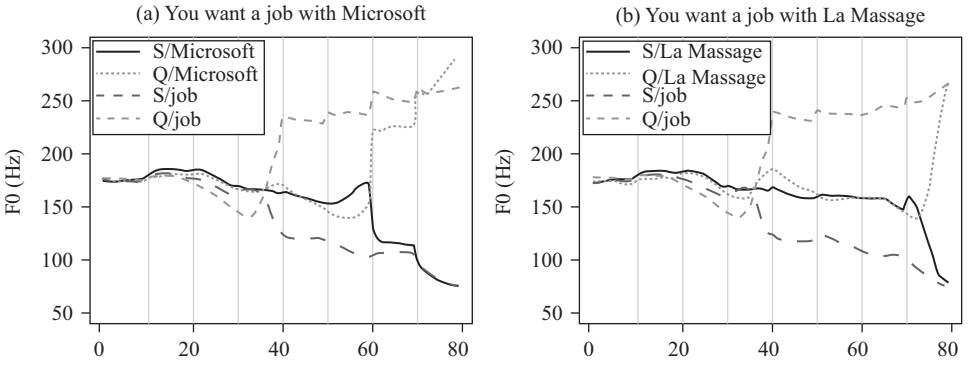


Figure 11.6
Mean F_0 contours of statements and questions in American English. The word after the virgule / is focused. *Source*: Data courtesy of Liu et al. (2013).

increases exponentially (Yuan, Liberman, and Cieri 2006). It is further hypothesized that syllable duration is combined with duration of silent pauses to serve as an affinity index that iconically encodes relational distance with temporal distance (Wang, Xu, and Zhang 2019; Xu 2009). It is also found, through a series of experimental studies, that emotion, attitude, and vocal attractiveness are encoded with a combination of pitch, spectral density, voice quality, and speech rate properties, based on a set of hypothetical bioinformational dimensions, and that such encoding is again done in parallel with other, more linguistic functions (Chuenwattanapranithi et al. 2008; Hsu and Xu 2014; Liu and Xu 2014; Noble and Xu 2011; Xu, Kelly, and Smillie 2013; Xu et al. 2013).

There are also various other communicative functions that may be prosodically encoded. But as reviewed in Xu (2011a), there may be more conceptually plausible communicative functions than there are consistent encoding schemes. The establishment of the latter, therefore, can only be done one by one through empirical studies.

11.3 PENTA as a Research Tool

As a theory of prosody, PENTA can be used in practice as a framework under which empirical studies can be conducted. Such usage may involve setting up hypotheses to be tested, selecting experimental factors to be manipulated, and interpreting data and results through systematic analyses. But PENTA can be also used as a research tool in its own right, thanks to its quantification into a computational model (Prom-on, Xu, and Thipakorn 2009). The current computational version is qTA, which is a third-order critically damped linear system, as shown in the following formula:

$$f_0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \tag{11.3}$$

Here, $f_0(t)$ is the surface F_0 of a syllable as a function of time. $mt + b$, the first term, is the underlying pitch target as a linear function of time t, with m representing the slope and b representing the height of the target. The second term represents the natural response of the system, in which the transient coefficients c_1, c_2, and c_3 are calculated based on the initial F_0 dynamic state and pitch target of the current syllable. Parameter λ represents the strength of the F_0 movement toward the target.

qTA also explicitly represents the state transfer between adjacent syllables by taking the final F_0 state of the preceding syllable in terms of its final F_0 $f_0(0)$, velocity $f_0'(0)$, and acceleration $f_0''(0)$ as the initial F_0 dynamic state of the current syllable. With this initial state, the three transient coefficients are computed with the following formulae:

$$c_1 = f_0(0) - b \; c_1 = f_0(0) - b \tag{11.4}$$

$$c_2 = f_0'(0) + c_1\lambda - m \tag{11.5}$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2 \tag{11.6}$$

A number of principles are behind the development of qTA. The first is to use as few free parameters as possible, and every free parameter should be communicatively meaningful, that is, usable by one or more encoding schemes. The second principle is that all the critical components of the model should be quantitatively implemented, so

as to faithfully reflect the theoretical framework. Particularly worth mentioning is the specification of target slope *m*, which, as far as we know, is used only by qTA. The need for this parameter is based on findings that the final slope is one of the most consistent properties of a dynamic tone, as mentioned in 11.2.1. The third principle is that the model parameters should be learnable from real speech data, so that there are no missing links between the theoretical framework and the reality of continuous speech. This would also potentially allow the simulation of the acquisition of tone and intonation.

Note that qTA computes F_0 contours of only a single sequence of syllables, which is different from superpositional models (Fujisaki 1983; Bailly and Holm 2005) that compute multiple layers of F_0 contours and then combine them. There are no mathematical restrictions, however, on using units other than the syllable. But in our own practice, we have always used the syllable (Xu and Prom-on 2014) as the domain of local pitch targets. This scheme also means that for each syllable, there is only a single target that needs to be estimated (but see discussion in section 11.2.1 of the possibility of having consecutive targets within a single syllable), which has turned out to be a critical property for our understanding of the operation of speech prosody, as will be discussed later.

11.3.1 Computational Modeling Tools

Since the proposal of qTA, three computational tools have been developed to enable its conceptual exploration and quantitative testing. qTA_demo (www.homepages.ucl.ac.uk/~uclyyix/qTA/) is a web-based interactive programs that demonstrates how qTA works. Its interactive features make it a convenient tool for a quick impromptu test of an idea or a prediction based on the TA model.

The other two tools, qTAtrainer (www.homepages.ucl.ac.uk/~uclyyix/qTAtrainer/) and PENTAtrainer (www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer/), both use machine learning algorithms to automatically extract target parameters from real speech data through analysis-by-synthesis. In the analysis-by-synthesis paradigm, candidate targets are iteratively tested by applying them in the qTA function to generate continuous F_0 contours. The goodness of fit between the synthetic and natural contours is then used as the criterion in the selection of the targets (Prom-on, Xu, and Thipakorn 2009; Prom-on and Xu 2012). The quality of the F_0 generation is assessed by three means: (i) root mean squared error, which measures the discrepancy of the synthetic contours from the original contours in terms of point-by-point height difference; (ii) Pearson's *r*, which assesses how closely the overall shape of the synthetic contours correlates with that of the natural contours; and (iii) perceptual evaluation in terms of category identification (e.g., tone, focus) and naturalness.

Both qTAtrainer and PENTAtrainer allow predictive synthesis of F_0 contours using categorical parameters learned from training. They differ only in terms of how function-specific targets are obtained. qTAtrainer takes a two-phase approach. In phase 1, an optimal target is obtained for each syllable of each utterance by comparing the performance of all possible combinations of the three target parameters (*b*, *m*, and λ in equation (11.3)). The parameter set that achieves the best fit to the natural F_0 contour of a specific syllable (i.e., with the smallest sum square error) is selected as its pitch target. An example of such resynthesis is shown in figure 11.7, where the short dashed lines are the learned targets. The F_0 contours generated with these learned targets (solid curves) seem to fit the natural F_0 contours (dotted curves) quite well. In phase 2, categorical targets are obtained by averaging over the parameters of all the syllables in the corpus that belong to the same categorical combination, for example, all the on-focus

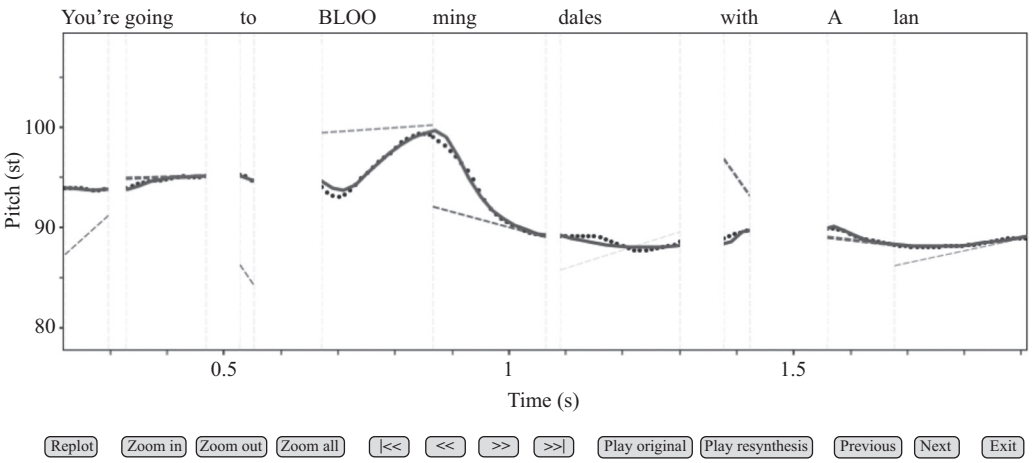


Figure 11.7
Original (dotted) versus resynthesized (solid) F_0 contours of the English utterance "You're going to Bloomingdales with Alan" by qTAtrainer. Pitch is F_0 given in semitones.

H tones that occur at the beginning of a sentence (Prom-on, Xu, and Thipakorn 2009). As found in Prom-on, Xu, and Thipakorn (2009) and Liu et al. (2013), good predictive results can be obtained for both English and Mandarin with this approach.

The categorization by averaging strategy employed in qTAtrainer, despite its reasonable performance, cannot satisfactorily estimate all qTA parameters. In particular, locally estimated parameters may not be globally optimal. For example, in some cases, the rate of target approximation (λ) may not be adequately estimated if there is severe target undershoot. In addition, the exhaustive search implemented in qTAtrainer is inefficient and probably ecologically unrealistic as a learning algorithm. These problems are addressed by PENTAtrainer, in which function-specific targets are learned from an entire corpus that has been functionally annotated (Prom-on and Xu 2012; Xu and Prom-on 2014). This is achieved with *simulated annealing*, an optimization algorithm that performs stochastic parameter sampling to avoid local minima in parameter estimation. Figure 11.8 shows an example of an annotated utterance (top) and natural F_0 and synthetic contours (bottom), where the latter is generated with categorical targets learned from an entire corpus.

In Xu and Prom-on (2014), good overall numerical results were achieved with PENTAtrainer for English (the same data set tested with qTAtrainer in Liu et al. 2013), Mandarin, and Thai. In Prom-on, Xu, and Thipakorn (2009), which applied categorization by averaging, the perceptual identification rates for tone in Mandarin and focus in both Mandarin and English were similar for synthetic and natural speech. Just as importantly, synthetic prosody (in terms of F_0 and duration) was heard to be as natural as natural prosody for English and only slightly worse for Mandarin.

The total number of function-specific parameters that need to be learned from the speech corpora to perform predictive synthesis is fairly small. In Xu and Prom-on (2014), seventy-eight parameters (twenty-six b, m, λ values each) were used for 960 English sentences (consisting of 8,640 syllables), eighty-four parameters for 1,280 Mandarin sentences (consisting of 10,240 syllables), and thirty parameters for 2,500 Thai disyllabic phrases. This suggests that a high level of abstraction can be achieved with

PENTA-based computational approaches. The abstraction level is comparable to other models, for example, five parameters per Standard Chinese tone in the Fujisaki model (Fujisaki 1983) and four parameters per intonational event in the Tilt model (Taylor 2000).

11.3.2 Modeling Encoding Schemes of English Prosody: An Illustration

The application of the computational tools described has allowed us to model some of the major prosodic encoding schemes in English and Mandarin. Figure 11.9 provides a summary illustration with modeling data on English from Xu and Prom-on (2014). Each graph shows the original F_0 of an American English utterance, pitch targets learned by PENTAtrainer, and synthetic F_0 contours generated with the learned targets. These sentences were spoken with either sentence-medial or sentence-final focus and as either statements or questions (Liu et al. 2013). As can be seen, the predicted F_0 contours closely simulate the interactive realization of multiple encoding schemes of lexical stress, focus, and modality in American English (as described in Xu and Prom-on 2014). These include, first, the robust post-focus pitch range shift, whose direction depends on modality: downward in a statement (a, b), but upward in a question (c, d). Second, it has also simulated the interactive determination of target slope by lexical stress, focus, and modality: negative slope for on-focus word-final stressed syllables in a statement, but positive in a question (*job* in b, d).

Note also that the match between the synthetic and original F_0 contours in figure 11.9 is not nearly as good as that in figure 11.7. This is partly because here, the synthesis is based on categorical parameters learned from all the utterances by a speaker in a corpus, as opposed to resynthesis in figure 11.7 (by qTAtrainer), but partly also because there is still room for further adjustments in our functional annotations. For example, because the relative position of unstressed syllables within an initial-stressed word is not annotated in this simulation, the pitch targets of the unstressed syllables are the same regardless of their positions in the word. As a result, the synthetic F_0 in *-crosoft* does not show the final upstep in figure 11.9d. Thus, even if the major characteristics of the encoding schemes have been identified, their detailed properties are still an object of continuous investigations.

11.4 Broader Significance

PENTA is a conceptual framework for characterizing speech prosody as a process of articulatorily encoding communicative functions. Thus, its central concern is how prosody operates as part of the speech communication process. At each stage of the development of PENTA, especially during its quantification, we have tried to ensure the operability of the model, that is, the ability to take input data in a sufficiently detailed form and generate outputs that are sufficiently close to surface acoustics, based on mechanisms that are biomechanically plausible. This operability is not just about how the model works internally, but also about how it is linked to external processes. As a result, PENTA is relevant not only for the direct characterization of prosody, but also for understanding many broader issues, including speech acquisition, language change, typology, and phonological representation. As will be shown in the following discussion, the implications for all these issues are most clearly seen through computational modeling.

(a)

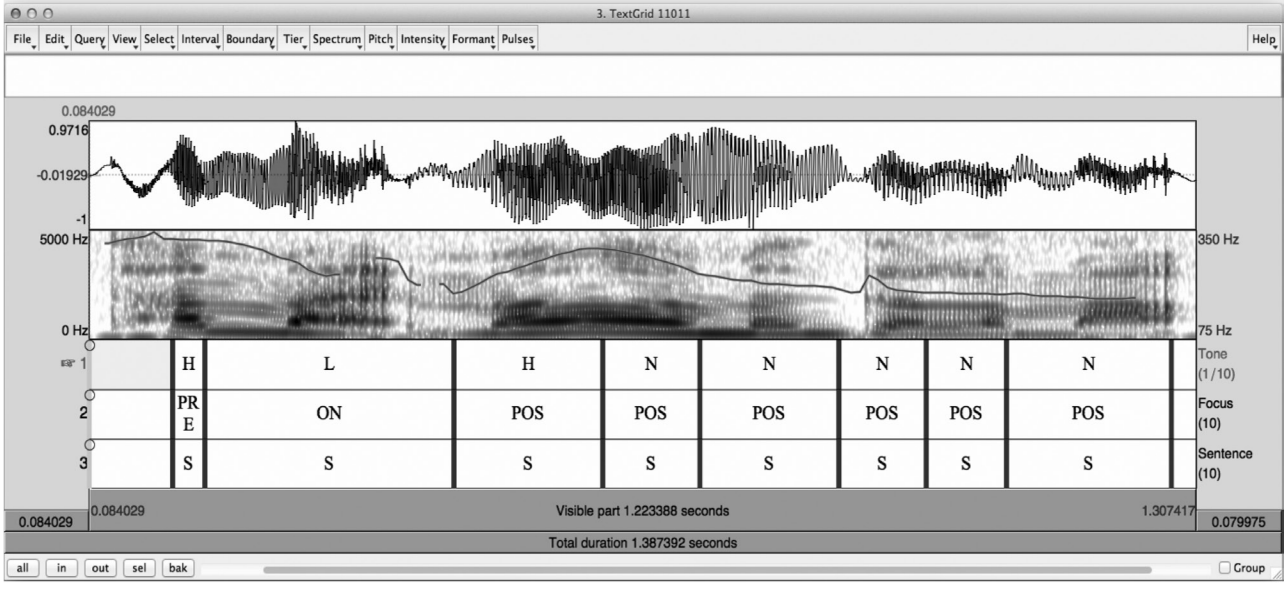


Figure 11.8

Screenshots of PENTAtrainer (www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer/) interfaces. The annotation interface (top) allows users to mark functional units and their temporal domains. In this example (Mandarin sentence "tā MǍI māma men de la ma?" [Did he BUY what mother has?], with focus on mai3), the annotated functions are lexical tone, focus, and sentence type (modality). All the boundaries are set to coincide with syllable boundaries. The temporal domain of a functional region covers syllables with identical labels. The output interface (bottom) displays learned pitch targets (dashed lines), synthetic (dotted), and natural (solid) F_0 contours. It also allows users to play the utterance with either synthetic or natural prosody (Prom-on and Xu 2012).

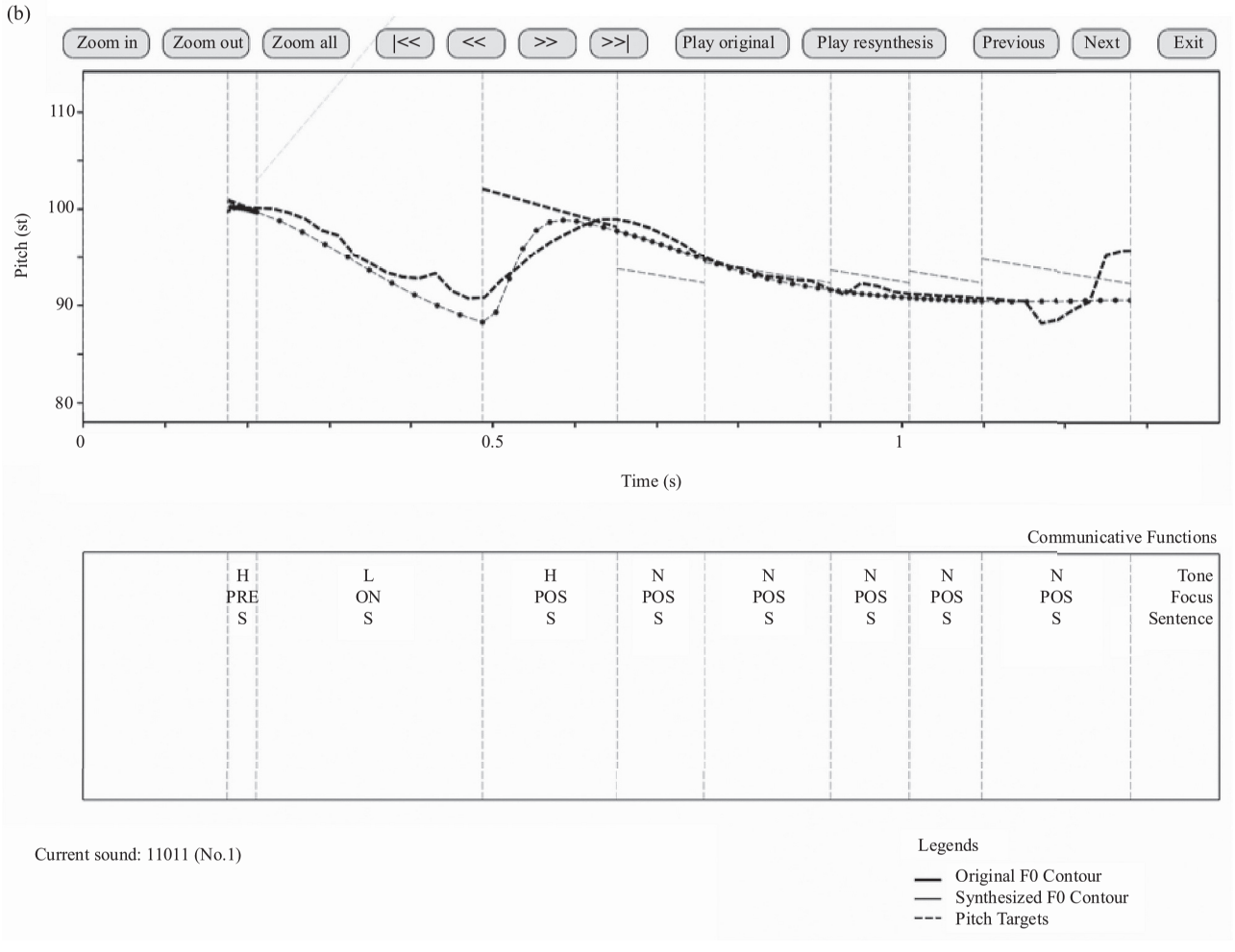


Figure 11.8
(continued)

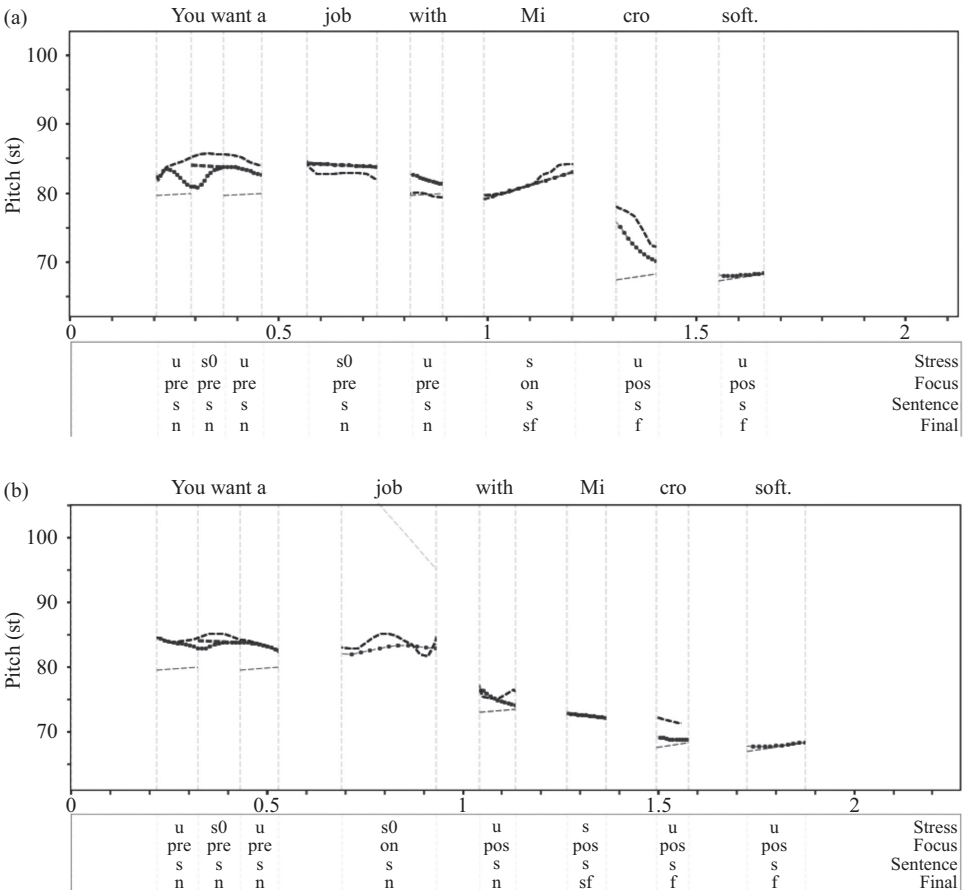


Figure 11.9
Original (dashed) and synthetic (dotted) F_0 contours of the sentence *You want a job with Microsoft*, spoken by a male American English speaker as either a statement (a, b) or a question (c, d), with focus on either *job* (b, d) or *Microsoft* (a, c). Also displayed are the pitch targets (straight dashed lines) learned by PENTAtrainer based on the functional annotations shown at the bottom of each graph. Stresses are categorized as unstressed (u), nonfinal stressed (S), and word-final stressed (s0). Syllable positions (labeled as final) are categorized as nonfinal (n), semifinal (sf), and sentence-final (f). All the graphs are screenshots of the demo window of the Synthesis tool (synthesize.praat) in the PENTAtrainer package. *Source*: Data courtesy of Xu and Prom-on (2014).

11.4.1 Role of Computational Modeling

The explanatory power of a scientific theory can be measured by the number of falsifiable predictions it can make and the level of specificity of these predictions. A theory of speech prosody, for example, can be evaluated in terms of the number of prosodic patterns it can predict and how closely the predicted surface forms match the natural ones in fine detail. A computational model, especially if it is able to generate continuous surface acoustic patterns, would offer an effective means of testing the predictive power of its corresponding prosodic theory. In contrast, a theory with no quantitative implementation is difficult to falsify, because it cannot generate predictions that are

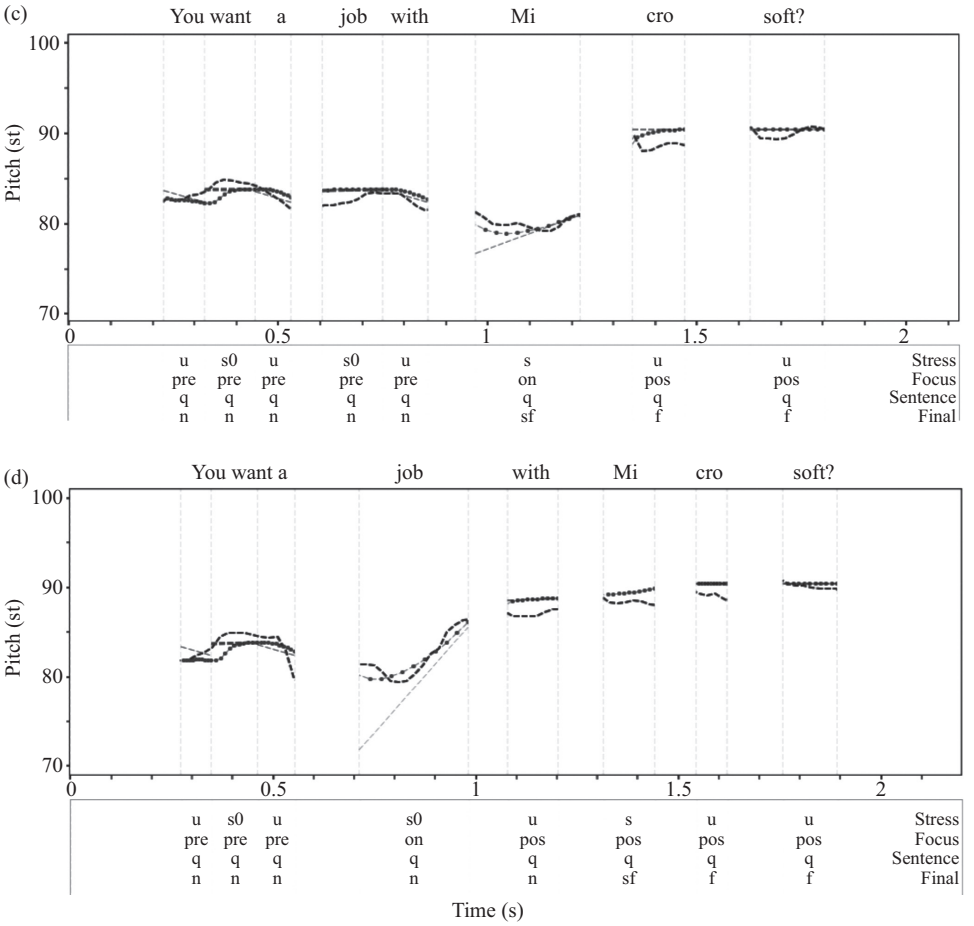


Figure 11.9
(continued)

detailed enough to be compared with the naturally occurring patterns or with other theories that do have quantitative implementations.

In practice, however, models can be constructed at different levels and generate predictions in different degrees of details. Theories such as autosegmental-metrical (AM) phonology and optimality theory are models that take certain forms of data (underlying segments or features) as input and generate predictions as output (narrow phonetic transcription) based on hypothetical mechanisms. None of these theories, however, generate outputs that are detailed enough for numerical acoustic comparisons with real speech. In other words, they lack adequate predictive power. For certain theories, for example, the AM theory of intonation (Ladd 2008; Pierrehumbert 1980; Pierrehumbert and Beckman 1988; Pierrehumbert and Hirschberg 1990), it is not even clear what the input is to the model-internal prosodic grammar. This, plus the fact that the output is largely symbolic, makes the theory hard to falsify.

Hence, computational modeling with fine prosodic details is not just a fancy and dispensable way of doing prosody. Modeling at a sufficient level of detail allows us to

see the exact consequences of each assumption we make in a theory. Also, both speakers and listeners have to deal with fine phonetic details. So, any theory that claims to be relevant about the cognitive process of prosody eventually has to have an account of how prosody is articulated by adults and how it is acquired by children. Models that leave these details to theory-external processes cannot be considered cognitively operable, unless plausible interfaces with the external processes have been explicitly proposed.

11.4.2 Articulation as Part of Cognition

The preceding discussion also brings up an issue that so far has received little attention. It is generally assumed that anything cognitive happens only in the brain, and that motor movements, because they are external to the brain, are not part of cognition (Pierrehumbert 1990). What we have learned from computational modeling with qTAtrainer and PENTAtrainer suggests, however, that this may not be the case. Because the qTA model is a simulator of the articulatory process of pitch generation, extracting underlying pitch targets with both trainers mimics the use of articulators as part of a cognitive learning process. In PENTAtrainer, for example, during target extraction, the articulatory system represented by qTA is used to repeatedly generate F_0 contours with randomly selected pitch targets, and the contours are then compared to the natural ones. If the match is good enough, the target is accepted for the current stage of learning. If the match is bad, the hypothetical target is discarded and the random target selection continues. In this kind of learning, a biomechanical device (here simulated by qTA) is indispensable. This is because, without it, the brain would have to replicate all the peripheral processes, which would lead to the kind of paradox discussed in section 11.2.1.

More interestingly, what we have learned from modeling actually matches well with what we now know about vocal acquisition by songbirds and human children. Like human babies, young (male) songbirds have to learn their songs from adult males. But their learning takes a number of stages. In an early sensory learning stage, they need to hear the adult songs and store the songs as templates. Otherwise they would never learn to sing properly (Brainard and Doupe 2002; Petkov and Jarvis 2012). In the later sensorimotor learning stage, they no longer need to hear the adult songs, but need to hear their own song practices (Brainard and Doupe 2002). Deprived of such opportunity, they would again never sing properly. There is also evidence that human children follow a similar order of bootstrapped learning. Acquisition of normal speech is severely affected not only in children who lose hearing at a very young age, but also in those who are unable to vocalize during a later practice stage (Cowie and Douglas-Cowie 1992). And children who cannot generate or hear their own voice prior to puberty experience severe delays in their speech development (Doupe and Kuhl 1999; Kamen and Watson 1991; Kuhl and Meltzoff 1996; Locke and Pearson 1990).

Thus vocal learning in both songbirds and humans requires the use of the very articulatory system that is later used in mature song or speech production. Both the naturally occurring biological evidence and modeling evidence, therefore, point to the use of a biomechanical system as part of the cognitive process of vocal learning. This in turn demonstrates the usefulness of computational modeling in our theoretical understanding of speech, provided that biological plausibility is taken into full consideration during the development of the models. It also shows the relevance of simulating learning as a crucial task of computational modeling. After all, mature vocalization is but a

behavior at the later stage of an incessant vocal learning process. Thus, demonstration of learnability should go a long way toward demonstrating full operability.

11.4.3 What about Phonology?

Another critical issue to be addressed in this chapter is whether PENTA sees a role for a phonological level of representation in prosodic theory. More specifically, does PENTA posit a set of abstract, symbolic primitives that can be combined to generate well-formed intonation contours linked to both meaning and a range of acceptable phonetic instantiations?

This issue goes back to the earlier discussion of the Saussurean view on linguistic symbols as unities of the signifiers and the signified. Because the symbols concerned in the earlier days of linguistics are structures like words and morphemes, it is natural to take it for granted that the symbols are relatively certain while the meanings are relatively vague. But the difficulty encountered in the study of prosody suggests that it is crucial to consider as early as possible the meanings conveyed by prosody, given the obscurity of the prosodic forms. This has led to our view that it is the signified, namely, the communicative functions, that should be the defining properties of prosodic units, while the signifiers, that is, the prosodic forms, should be empirically discovered for each of the communicative functions. This is the basis of the notion of encoding schemes in PENTA.

Although the encoding schemes are not the same as phonological entities such as pitch accents, phrase accents, and boundary tones (Pierrehumbert 1980), they do share some similarities. First, like phonological entities, encoding schemes are abstract, as they do not exactly resemble what is observable in surface prosody. Unlike phonological entities, however, encoding schemes are not symbolic, because they do not rely on symbolic values to define their identity.[3] Rather, their identity is defined by the respective communicative functions. Also, unlike pitch accents and boundary tones, encoding schemes are not the primitive units in PENTA. Rather, the prosodic primitives are the syllable-bound pitch targets. Their primitive status comes from the fact that they are obligatory in the production of the syllable, as discussed in section 11.2.1. But pitch targets are each associated to multiple functions, which makes them inappropriate as phonological units.

Most importantly, to PENTA, it is critical to identify the kind of representation that enables the transmission of communicative meanings from the speaker to the listener, that is, representations that are operable. An operable representation needs to be sufficiently abstract so as not to require too much memory resource. It also needs to be able to account for fully continuous surface forms, leaving as few details unexplained as possible. And it should allow adequate representation of individual and dialectal variations. Finally, it needs to be learnable with testable computational algorithms.

The solution found in the PENTA approach that can satisfy all of these requirements is parametric representation in the form of underlying articulatory targets. Table 11.1 shows a list of properties of such parametric representation as compared to symbolic representation. As can be seen, the parametric representation can satisfy all four requirements. First, it is abstract because it can represent an infinite number of contextual variants by translating underlying targets into surface acoustics based on model-simulated articulatory mechanisms. Thus, the representation itself is free of redundant surface details. It is also gradient, because the targets are numerically specified and so are not categorical themselves. This allows them to represent numerous

Table 11.1
Comparison of PENTA-based parametric and AM-style symbolic representations

Properties	Parametric	Symbolic
Abstract (able to represent multiple surface variants)	√	√
Gradient (allowing for individual and dialectal variations/representation itself is not categorical)	√	
Continuous (with built-in time-varying patterns)	√	
Data-driven (trainable, learnable)	√	?
Functionally defined	√	

dialectal and individual variations of the same linguistic categories. The targets are also continuous with built-in time-varying patterns. Given that the continuity is an intrinsic property of the target approximation model, the representation does not leave the filling of the detailed contours to other, unspecified mechanisms, as is the case with theories like AM. Furthermore, the specific values of this representation are data-driven, so they can be obtained from real speech data. More importantly, this way of parameter extraction allows the simulation of real-life learning of articulatory targets. This makes the representation computationally operable in terms of simulating biological reality with a reasonable level of plausibility.

Finally, even more directly relevant to the core concerns of phonology, the model-based parametric representation offers not only an operable way of linking meanings to articulatory targets, but also a biologically plausible mechanism for the emergence and evolution of rule-like phonological variations. This can be seen in two examples of our recent modeling study. The first one is about the highly perplexing phenomenon of tone sandhi (Chen 2000). As a well-known example, the Mandarin tone 3 assumes a T2-like surface form before another tone 3: T3 T2 / _ T3. From a functional perspective, such a "rule" makes little sense because it leads to homophony and so reduces categorical contrasts. But rules of this kind are commonplace in tone languages (Chen 2000), which is puzzling. So, there must be some strong biological constraint that makes the emergence of such communication-harming rules inevitable.

Our modeling simulation of tone and intonation with PENTAtrainer seems to offer a suggestion as to what this constraint might be. That is, it is probably because there is no other way of conveying communicative meanings than to load all the functions onto syllable-bound articulatory targets that are realized in succession. Because each target is co-specified by multiple functions, learners rarely encounter monofunctional targets. Thus, a tone is hardly ever learned as having a single-category target. Instead, for example, a specific version of T3 may be learned with a particular target when followed by another T3. If, for some unknown articulatory, perceptual, social, or historical reason, this version of T3 happened to sound more like T2 in surface form, a T2-like context-specific target could be learned for T3. But this context-specific T3 target does not need to be identical to the T2 target in the same context, because the functional combinations are not the same. As found in Xu and Prom-on (2014), the best modeling result was obtained when the sandhi T3 was allowed to learn its own target, rather than when it was forced to share the same target with T2. This result is in line with the empirical finding of subtle yet consistent differences between the original and sandhi-derived T2 in Mandarin (Peng 2000; Xu 1997) despite their full perceptual merger (Peng 2000; Wang and Li 1967)

Beside tonal context, many other frequently occurring functions can also be the conditioning factors, such as boundary marking, focus, sentence type, and so on. For example, as discussed in section 11.2.2, syllable-bound pitch targets are conditioned in American English by the interaction of lexical stress, focus, and sentence type (Liu et al. 2013). And, as shown in section 11.3.2, in PENTA-based computational modeling, these multifunctional targets can be easily learned from natural speech data (Xu and Prom-on 2014). The same principle is applicable to countless other cases. If these resemble biological reality, we would have identified a core mechanism behind many of the phonological rules.

11.4.4 Prosodic Typology

Through our empirical studies, the PENTA approach has been shown to be also relevant for prosodic typology. Specifically, typological phenomena, as we have found, are best described in terms of variations in the way specific communicative functions are encoded. The model itself, however, does not stipulate what prosodic form a language should take. Instead, it allows researchers to empirically discover for each function what its encoding is in a particular language. This function-oriented approach has led to our discovery of an interesting typological distribution of a prosodic pattern, namely, *focus*. Following a series of classic studies of focus realization in American English (Cooper, Eady, and Mueller 1985; Eady and Cooper 1986; Eady et al. 1986), we found that Mandarin Chinese shares a common prosodic feature with American English, which we termed *postfocus compression* (PFC). That is, the pitch range of postfocus words is lowered relative to the neutral-focus reference.

In later studies, we were surprised to find that Taiwanese (Southern Min) and Cantonese do not have this feature. From that, we hypothesized that this typological pattern is not accidental; rather, it is an indication that the origin of Mandarin and probably other northern Chinese dialects is different from that of southern Chinese languages such as Taiwanese and Cantonese. Furthermore, PFC has been found in a number of languages in language families that have been described as belonging to the putative Nostratic macro-family. We have therefore hypothesized that PFC originated from proto-Nostratic over thirteen thousand years ago (Bomhard 2008). One of the bases of this hypothesis is the finding that PFC is almost impossible to pass on from one language to another (Xu 2011a; Xu, Chen, and Wang 2012). It is also hard for it to emerge on its own, as none of the non-Nostratic languages examined for focus has shown evidence of PFC, regardless of whether they have lexical stress, tone, or any other prosodic features (Xu 2011a). PFC is also hard to acquire in a second language, as it seems to require the learner to speak the language more than their first language before PFC is consistently seen in their production (Chen 2015; Chen, Xu and Guion-Anderson 2014).

Another potential prosodic typology is about how focus interacts with sentence type. In American English, the pitch range of a postfocus word is raised well above the pitch median in questions. But this feature is missing in Mandarin (Liu and Xu 2005; Liu et al. 2013) and not reported in other languages. It is possible that both patterns are shared by many other languages, which is worth exploring in future studies.

Yet another typological divide is in terms of the interaction between focus and phrasing. In languages like English and Mandarin, focus can operate on the syllable level. When focus falls on a multisyllabic word, any syllable after the stressed syllable is treated as belonging to the postfocus domain (Xu and Xu 2005). In contrast, in languages like French, due to the constraint of phrase structure, or due to lack of lexical

stress, auxiliaries do not show consistent pitch-accent–like patterns. In this case, the interaction of focus and phrasing is partially influenced by the lexical function, which differs across stress versus nonstress languages.

Finally, an even deeper typological divide has been suggested by Rialland (2009), who discovered that in a number of languages in a geographically restricted area in the Sudanic belt of Africa, prosodic means other than final pitch raising are used to indicate interrogation. The fact that the final pitch raising seems to be universal among languages outside Africa may suggest that this divide existed before *Homo sapiens* first left Africa over tens of thousands of years ago. All these patterns, if proven reliable by further studies, suggest that prosody is likely to be more stable than segmental, lexical, or even syntactic features during language change. If so, prosodic features could be used as indicators of language affinity that may have greater time depths than traditional indicators (Longobardi and Guardiano 2009; Nichols 1996).

11.4.5 Encoding Schemes as Prosophemes

In the lexical domain, the smallest meaning-bearing units are recognized as morphemes, which can be as small as a segment or as large as a multisyllable word or word root. In the prosodic domain, our work with PENTA has suggested that it is the encoding schemes that bear the closest resemblance to lexical morphemes (Liu et al. 2013). First, like lexical morphemes, each encoding scheme consists of multiple prosodic components. These components are meaningless by themselves, but they act jointly to mark both intra- and interfunctional contrasts. Second, similar to lexical morphemes, encoding schemes have allomorph-like variants whose occurrence is conditioned by factors such as location in sentence and interaction with other prosodic functions. Finally, similar to lexical morphemes, encoding schemes are language-specific, and their patterns likely have historical origins. Given that prosodic encoding schemes use prosody to carry postlexical meanings, we may call them *prosophemes*.

Our description of prosodic focus so far has made it a clear case of prosopheme. First, focus is realized not only with specific pitch patterns, but also with specific patterns of duration, intensity, and even voice quality (as reviewed in Xu, Chen, and Wang 2012). Second, PFC of pitch and intensity has been found to be language-specific, and languages with this feature have been hypothesized to be linked to a common protolanguage (Xu 2011a). The encoding scheme of focus in languages like Mandarin and English is therefore multicomponential, language-specific, and with likely historical etymologies, thus bearing all the major hallmarks of a lexical morpheme.

Another case is modality or sentence type, which encodes whether a sentence is a statement or yes/no question. In American English, for example, modality determines not only sentence-final F_0 (which is treated as boundary tone in AM theory), but also the underlying pitch targets of all stressed syllables throughout the sentence (Liu et al. 2013), as discussed in sections 11.2.2 and 11.3.2. Modality also interacts with focus to determine the pitch range of all postfocus words: well above the neutral-focus reference in a question, but well below the reference in a statement. Both of these features are missing in Mandarin. Question intonation in Mandarin does not involve postfocus raising of pitch range above the reference, and neither does it change the pitch targets of individual syllables (Liu and Xu 2005; Liu et al. 2013). So, again we see clear evidence of multicomponentiality, conditional variability, and language specificity in the encoding scheme of modality.

Therefore, the multicomponential coding of the prosodic functions demonstrates that it is the communicative functions, rather than the directly observable surface F_0

events, that bear the most resemblance to lexical morphemes. This prosopheme notion is an alternative to the tonal morpheme proposed by Pierrehumbert and Hirschberg (1990). As discussed in detail in Liu et al. (2013), many of the morpheme-like meanings proposed by Pierrehumbert and Hirschberg for the phonological intonational components are similar to those associated with prosodic functions such as focus and modality. Furthermore, some proposed phonetic implementation rules in AM theory (Pierrehumbert 1980; Pierrehumbert and Hirschberg 1990) are part of the morpheme-like characteristics of focus and modality. For example, the upstep rule in English, which is said to raise the portion of F_0 corresponding to a high boundary tone H% relative to the preceding H- phrase accent, is part of a continuous upshift of postfocus pitch range to mark a question. This extra raising is therefore morphophonological, that is, part of a prosopheme rather than being due to a phonetic implementation rule.

11.4.6 The Perceptual Perspective

So far we have had little discussion of how PENTA can account for the perception of prosody. A main reason is that we have not yet conducted extensive investigation of the perception of prosody, with the exception of studies on emotional prosody, which were mainly perception-based (Chuenwattanapranithi et al. 2008; Noble and Xu 2011; Xu, Kelly, and Smillie 2013; Xu et al. 2013). The few perceptual experiments we have performed on nonemotional prosody were mainly done with the purpose to verify the production patterns found in those studies (Taheri-Ardali, Rahmani, and Xu 2014; Xu, Chen, and Wang 2012; Xu, Xu, and Sun 2004) or to evaluate modeling performance (Prom-on, Xu, and Thipakorn 2009; Xu and Prom-on 2014). To test whether PENTA can also serve as a proper theory of prosody perception, especially in terms of being able to make precise predictions on how communicative functions are perceptually decoded from prosody, additional research is needed. Nonetheless, there is already evidence from modeling studies using self-organizing maps (Kohonen 1982) that the perception of tone and intonation is based on mechanisms that do not require full knowledge of production (Gauthier, Shi, and Xu 2007a, 2007b; Gauthier, Shi, and Xu 2009). This is consistent with the fact that perception learning precedes production learning in speech acquisition (Kuhl et al. 1992; Werker and Tees 2005). But more research, both behavioral and modeling, is needed to develop better predictive knowledge about prosody perception and how it is linked to the production of prosody.

11.5 Conclusion

PENTA was proposed based on the premise that prosody is a system of encoding communicative meanings with an articulatory system. From this basis, we have identified syllable-synchronized sequential target approximation as the core articulatory mechanism of prosodic encoding. In PENTA, therefore, syllable-bound pitch targets and their articulatory approximation are the phonetic primitives. On the meaning side, PENTA assumes that categories and dimensions of communicative meanings are defined by function rather than form. Thus, there are no theory-intrinsic units in PENTA that are equivalent to pitch accents, phrase accents, and boundary tones (Pierrehumbert 1980), or accent and phrase commands (Fujisaki 1983), or any units that are defined primarily by their phonetic properties. Yet the empirically established encoding schemes of communicative functions established under PENTA appear to bear close similarities to lexical morphemes, which gives rise to the term *prosopheme*. Empirically guided search for function-form unities has also led to our finding of prosody-based typological divisions

such as those based on PFC, postfocus pitch upshift in questions, and differential interaction of focus and phrasing. The principle of parallel encoding in PENTA also makes it easy to incorporate more gradient and universal functions such as emotional codes based on the hypothetical bioinformational dimensions (Xu, Kelly, and Smillie 2013).

The quantization of PENTA has given us tools for computational modeling of prosody that can significantly increase the predictive power of the theory. Most interestingly, modeling has made it clear to us that the only way to acquire an encoding scheme is through the learning of syllable-bound multifunctional targets. Such a process, because it requires the learning of different targets for the same functional category conditioned by interactions with other functions, is a likely breeding ground for phonological rules such as tone sandhi that seem to make little sense from a purely functional perspective. Much more research is needed, nevertheless, to explore the full potential of computational modeling not only for speech prosody, but also for the segmental aspect of speech.

Notes

1. There is some overlap in content between this chapter and Xu et al. (2015).

2. This is facilitated by the software tool that we have developed, namely, ProsodyPro, a Praat-based script, available at www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/ (Xu 2013). Also, a similar tool for segmental analysis in the form of FormantPro (www.homepages.ucl.ac.uk/~uclyyix/FormantPro/). It facilitates systematic comparison of continuous formant trajectories, which is still rare in segmental studies.

3. Even if symbols sometimes are used in our approach, they are only for convenience of discussion. This also rules out PENTA as a transcription system because it does not use transcription as a means of analysis.

References

Bailly, G., and B. Holm. 2005. "SFC: A Trainable Prosodic Model." *Speech Communication* 46:348–364.

Bernstein, N. A. 1967. *The Coordination and Regulation of Movements*. Oxford: Pergamon Press.

Bolinger, D. L. 1972. "Accent Is Predictable (If You're a Mind Reader)." *Language* 48:633–644.

Bomhard, A. R. 2008. *Reconstructing Proto-Nostratic: Comparative Phonology, Morphology, and Vocabulary*. Leiden, the Netherlands: Brill.

Brainard, M. S., and A. J. Doupe. 2002. "What Songbirds Teach Us about Learning." *Nature* 417 (6886): 351–358.

Chen, M. Y. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press.

Chen, Y., and Y. Xu. 2006. "Production of Weak Elements in Speech: Evidence from F0 Patterns of Neutral Tone in Standard Chinese." *Phonetica* 63:47–75.

Chen, Ying. 2015. "Post-focus Compression in English by Mandarin Learners." In *Proceedings of the 18th International Congress of Phonetic Sciences*, edited by The Scottish Consortium for ICPhS 2015.

Chen, Y., Y. Xu, and S. Guion-Anderson. 2014. "Prosodic Realization of Focus in Bilingual Production of Southern Min and Mandarin." *Phonetica* 71:249–270.

Cheng, C., and Y. Xu. 2013. "Articulatory Limit and Extreme Segmental Reduction in Taiwan Mandarin." *Journal of the Acoustical Society of America* 134 (6): 4481–4495.

Chuenwattanapranithi, S., Y. Xu, B. Thipakorn, and S. Maneewongvatana. 2008. "Encoding Emotions in Speech with the Size Code: A Perceptual Investigation." *Phonetica* 65:210–230.

Connell, B., and D. R. Ladd. 1990. "Aspects of Pitch Realization in Yoruba." *Phonology* 7:1–29.

Cooper, W. E., S. J. Eady, and P. R. Mueller. 1985. "Acoustical Aspects of Contrastive Stress in Question-Answer Contexts." *Journal of the Acoustical Society of America* 77:2142–2156.

Cowie, R., and E. Douglas-Cowie. 1992. *Postlingually Acquired Deafness*. New York: Mouton de Gruyter.

de Saussure, F. 1916. "Nature of the Linguistics Sign." In *Cours de linguistique générale*, edited by C. Bally and A. Sechehaye, 66–70. New York: McGraw-Hill.

Doupe, A. J., and P. K. Kuhl. 1999. "Birdsong and Human Speech: Common Themes and Mechanisms." *Annual Review of Neuroscience* 22:567–631.

Eady, S. J., and W. E. Cooper. 1986. "Speech Intonation and Focus Location in Matched Statements and Questions." *Journal of the Acoustical Society of America* 80:402–416.

Eady, S. J., W. E. Cooper, G. V. Klouda, P. R. Mueller, and D. W. Lotts. 1986. "Acoustic Characteristics of Sentential Focus: Narrow vs. Broad and Single vs. Dual Focus Environments." *Language and Speech* 29:233–251.

Fujisaki, H. 1983. "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing." In *The Production of Speech*, edited by P. F. MacNeilage, 39–55. New York: Springer-Verlag.

Gandour, J., S. Potisuk, and S. Dechongkit. 1994. "Tonal Coarticulation in Thai." *Journal of Phonetics* 22:477–492.

Gauthier, B., R. Shi, and Y. Xu. 2007a. "Learning Phonetic Categories by Tracking Movements." *Cognition* 103:80–106.

Gauthier, B., R. Shi, and Y. Xu. 2007b. "Simulating the Acquisition of Lexical Tones from Continuous Dynamic Input." *JASA Express Letters* 121:EL190–195.

Gauthier, B., R. Shi, and Y. Xu. 2009. "Learning Prosodic Focus from Continuous Speech Input: A Neural Network Exploration." *Language Learning and Development* 5:94–114.

Gay, T. J. 1968. "Effect of Speaking Rate on Diphthong Formant Movements." *Journal of the Acoustical Society of America* 44:1570–1573.

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

Hombert, J.-M. 1977. "Consonant Types, Vowel Quality, and Tone." *Studies in African Linguistics* 8:73–190.

Hsu, C., and Y. Xu. 2014. "Can Adolescents with Autism Perceive Emotional Prosody?" In *Proceedings of Interspeech 2014*, edited by Haizhou Li, Helen Meng, Bin Ma, Eng Siong Chng, and Lei Xie, 1924–1928.

Hyman, L. M. 1993. "Register Tones and Tonal Geometry." In *The Phonology of Tone*, edited by H. v. d. Hulst and K. Snider, 75–108. New York: Mouton de Gruyter.

Kamen, R. S., and B. C. Watson. 1991. "Effects of Long-Term Tracheostomy on Spectral Characteristics of Vowel Production." *Journal of Speech and Hearing Disorders* 34:1057–1065.

Kelso, J. A. S. 1984. "Phase Transitions and Critical Behavior in Human Bimanual Coordination." *American Journal of Physiology: Regulatory, Integrative and Comparative* 246:R1000–R1004.

Kelso, J. A. S., D. L. Southard, and D. Goodman. 1979. "On the Nature of Human Interlimb Coordination." *Science* 203:1029–1031.

Kohonen, T. 1982. "Self-Organized Formation of Topologically Correct Feature Maps." *Biological Cybernetics* 43:59–69.

Kuhl, P. K., and A. N. Meltzoff. 1996. "Infant Vocalizations in Response to Speech: Vocal Imitation and Developmental Change." *Journal of the Acoustical Society of America* 100 (4): 2425–2438.

Kuhl, P. K., K. A. Williams, F. Lacerda, F., K. N. Stevens, and B. Lindblom. 1992. "Linguistic Experience Alters Phonetic Perception in Infants by Six Months of Age." *Science* 255:606–608.

Ladd, D. R. 2008. *Intonational Phonology*. Cambridge: Cambridge University Press.

Laniran, Y. 1992. "Intonation in Tone Languages: The Phonetic Implementation of Tones in Yorùbá." PhD diss., Cornell University.

Laniran, Y., and C. Gerfen. 1997. "High Raising, Downstep and Downdrift in Igbo." (Presentation, *Seventy-First Annual Meeting of the Linguistic Society of America*).

Latash, M. L. 2012. *Fundamentals of Motor Control*. Orlando, FL: Academic Press.

Lee, A., S. Prom-on, and Y. Xu. 2017. "Pre-low Raising in Japanese Pitch Accent." *Phonetica* 74 (4): 231–246.

Liu, F., and Y. Xu. 2005. "Parallel Encoding of Focus and Interrogative Meaning in Mandarin Intonation." *Phonetica* 62:70–87.

Liu, F., Y. Xu, S. Prom-on, and A. C. L. Yu. 2013. "Morpheme-Like Prosodic Functions: Evidence from Acoustic Analysis and Computational Modeling." *Journal of Speech Sciences* 3 (1): 85–140.

Liu, X., and Y. Xu. 2014. "Body Size Projection and Its Relation to Emotional Speech: Evidence from Mandarin Chinese." In *Proceedings of Speech Prosody 2014*, edited by N. Campbell, D. Gibbon, and D. Hirst, 974–977.

Locke, J. L., and D. M. Pearson. 1990. "Linguistic Significance of Babbling: Evidence from a Tracheostomized Infant." *Journal of Child Language* 17:1–16.

Longobardi, G., and C. Guardiano. 2009. "Evidence for Syntax as a Signal of Historical Relatedness." *Lingua* 119 (11): 1679–1706.

Mechsner, F., D. Kerzel, G. Knoblich, and W. Prinz. 2001. "Perceptual Basis of Bimanual Coordination." *Nature* 414:69–73.

Nichols, J. 1996. "The Comparative Method as Heuristic." In *The Comparative Method Revised*, edited by M. Durie and M. Ross, 29–71. Oxford: Oxford University Press.

Noble, L., and Y. Xu. 2011. "Friendly Speech and Happy Speech: Are They the Same?" In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 1502–1505.

Peng, S.-H. 2000. "Lexical versus 'Phonological' Representations of Mandarin Sandhi Tones." In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, edited by M. B. Broe and J. B. Pierrehumbert, 152–167. Cambridge: Cambridge University Press.

Petkov, C. I., and E. Jarvis. 2012. "Birds, Primates and Spoken Language Origins: Behavioral Phenotypes and Neurobiological Substrates." *Frontiers in Evolutionary Neuroscience* 4: 12. https://doi.org/10.3389/fnevo.2012.00012.

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

Pierrehumbert, J. 1990. "Phonological and Phonetic Representation." *Journal of Phonetics* 18:375–394.

Pierrehumbert, J., and M. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, J., and J. Hirschberg. 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Prom-on, S., F. Liu, and Y. Xu. 2012. "Post-Low Bouncing in Mandarin Chinese: Acoustic Analysis and Computational Modeling." *Journal of the Acoustical Society of America* 132:421–432.

Prom-on, S., and Y. Xu. 2012. "PENTATrainer2: A Hypothesis-Driven Prosody Modeling Tool." In *Proceedings of Exling 2012*, edited by A. Botinis, 93–100.

Prom-on, S., Y. Xu, and B. Thipakorn. 2009. "Modeling Tone and Intonation in Mandarin and English as a Process of Target Approximation." *Journal of the Acoustical Society of America* 125:405–424.

Rialland, A. 2009. "African 'Lax' Question Prosody: Its Realisations and Its Geographical Distribution." *Lingua* 119:928–949.

Silverman, K. 1986. "F0 Segmental Cues Depend On Intonation: The Case of the Rise after Voiced Stops." *Phonetica* 43:76–91.

Stevens, K. N. 1998. *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Sundberg, J. 1979. "Maximum Speed of Pitch Changes in Singers and Untrained Subjects." *Journal of Phonetics* 7:71–79.

Taheri-Ardali, M., Rahmani, H., and Y. Xu. 2014. "The Perception of Prosodic Focus in Persian." In *Proceedings of Speech Prosody 2014*, edited by N. Campbell, D. Gibbon, and D. Hirst, 515–519.

Taylor, P. 2000. "Analysis and Synthesis of Intonation Using the Tilt Model." *Journal of the Acoustical Society of America* 107:1697–1714.

Wagner, M. 2005. "Prosody and Recursion." PhD diss., MIT.

Wang, C., Y. Xu, and J. Zhang. 2019. "Mandarin and English Use Different Temporal Means to Mark Major Prosodic Boundaries." In *Proceedings of the 19th International Congress of Phonetic Sciences*.

Wang, W. S.-Y., and K.-P Li. 1967. "Tone 3 in Pekinese." *Journal of Speech and Hearing Research* 10:629–636.

Werker, J. F., and R. C. Tees. 2005. "Speech Perception as a Window for Understanding Plasticity and Commitment in Language Systems of the Brain." *Developmental Psychobiology* 46:233–251.

Xu, C. X., and Y. Xu. 2003. "Effects of Consonant Aspiration on Mandarin Tones." *Journal of the International Phonetic Association* 33:165–181.

Xu, Y. 1997. "Contextual Tonal Variations in Mandarin." *Journal of Phonetics* 25:61–83.

Xu, Y. 1998. "Consistency of Tone-Syllable Alignment across Different Syllable Structures and Speaking Rates." *Phonetica* 55:179–203.

Xu, Y. 1999. "Effects of Tone and Focus on the Formation and Alignment of F0 Contours." *Journal of Phonetics* 27:55–105.

Xu, Y. 2001. "Fundamental Frequency Peak Delay in Mandarin." *Phonetica* 58:26–52.

Xu, Y. 2005. "Speech Melody as Articulatorily Implemented Communicative Functions." *Speech Communication* 46:220–251.

Xu, Y. 2007. "Speech as Articulatory Encoding of Communicative Functions." In *Proceedings of the Sixteenth International Congress of Phonetic Sciences*, edited by J. Trouvain and W. J. Barry, 25–30.

Xu, Y. 2009. "Timing and Coordination in Tone and Intonation: An Articulatory-Functional Perspective." *Lingua* 119 (6): 906–927.

Xu, Y. 2011a. "Post-Focus Compression: Cross-Linguistic Distribution and Historical Origin." In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 152–155.

Xu, Y. 2011b. "Speech Prosody: A Methodological Review." *Journal of Speech Sciences* 1:85–115.

Xu, Y. 2013. "ProsodyPro: A Tool for Large-Scale Systematic Prosody Analysis." In *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, edited by B. Bigi and D. Hirst, 7–10.

Xu, Y. 2020. Syllable is a synchronization mechanism that makes human speech possible. *PsyArXiv*. https:// doi:10.31234/osf.io/9v4hr.

Xu, Y., S.-W. Chen, and B. Wang. 2012. "Prosodic Focus with and without Post-Focus Compression (PFC): A Typological Divide within the Same Language Family?" *Linguistic Review* 29:131–147.

Xu, Y., A. Kelly, and C. Smillie. 2013. "Emotional Expressions as Communicative Signals." In *Prosody and Iconicity*, edited by S. Hancil and D. Hirst, 33–60. Philadelphia: John Benjamins.

Xu, Y., A. Lee, S. Prom-on, and F. Liu. 2015. "Explaining the PENTA Model: A Reply to Arvaniti and Ladd (2009)." *Phonology* 32:505–535.

Xu, Y., A. Lee, W.-L. Wu, X. Liu, and P. Birkholz. 2013. "Human Vocal Attractiveness as Signaled by Body Size Projection." *PLoS ONE* 8 (4): e62397.

Xu, Y., and F. Liu. 2006. "Tonal Alignment, Syllable Structure and Coarticulation: Toward an Integrated Model." *Italian Journal of Linguistics* 18:125–159.

Xu, Y., and F. Liu. 2012. "Intrinsic Coherence of Prosodic and Segmental Aspects of Speech." In *Understanding Prosody: The Role of Context, Function, and Communication*, edited by O. Niebuhr, 1–26. New York: Walter de Gruyter.

Xu, Y., and S. Prom-on. 2014. "Toward Invariant Functional Representations of Variable Surface Fundamental Frequency Contours: Synthesizing Speech Melody via Model-Based Stochastic Learning." *Speech Communication* 57:181–208.

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

Xu, Y., and A. Wallace. 2004. "Multiple Effects of Consonant Manner of Articulation and Intonation Type on F0 in English." *Journal of the Acoustical Society of America* 115 (part 2): 2397.

Xu, Y., and M. Wang. 2009. "Organizing Syllables into Groups: Evidence from F0 and Duration Patterns in Mandarin." *Journal of Phonetics* 37:502–520.

Xu, Y., and Q. E. Wang. 2001. "Pitch Targets and Their Realization: Evidence from Mandarin Chinese." *Speech Communication* 33:319–337.

Xu, Y., and C. X. Xu. 2005. "Phonetic Realization of Focus in English Declarative Intonation." *Journal of Phonetics* 33:159–197.

Xu, Y., C. X. Xu, and X. Sun. 2004. "On the Temporal Domain of Focus." In *Proceedings of International Conference on Speech Prosody 2004*, edited by K. Hirose, 81–84.

Yuan, J., M. Liberman, and C. Cieri. 2006. "Towards an Integrated Understanding of Speaking Rate in Conversation." In *Proceedings of Interspeech 2006*, edited by R. M. Stern, 541–544.

Zemlin, W. R. 1988. *Speech and Hearing Science: Anatomy and Physiology*. Englewood Cliffs, NJ: Prentice Hall.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data is available.
Names: Barnes, Jonathan, 1970– editor. | Shattuck-Hufnagel, Stefanie, editor.
Title: Prosodic theory and practice / edited by Jonathan Barnes and Stefanie Shattuck-Hufnagel.
Description: Cambridge, Massachusetts : The MIT Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021000764 | ISBN 9780262543170 (paperback)
Subjects: LCSH: Prosodic analysis (Linguistics)
Classification: LCC P224 .P739 2022 | DDC 414/.6—dc23
LC record available at https://lccn.loc.gov/2021000764

Prosodic Theory and Practice

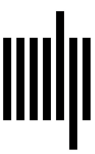
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel

Citation:

Prosodic Theory and Practice
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel
DOI: 10.7551/mitpress/10413.001.0001
ISBN (electronic): 9780262543194
Publisher: The MIT Press
Published: 2022

OA Funding Provided By:

OA Funding from MIT Press Direct to Open



The MIT Press

Author Response to the Commentary: Multiple Layers of Meanings Can Be Linked to Surface Prosody without Direct Mapping

Yi Xu, Santitham Prom-on, and Fang Liu

PENTA Is Not a Direct Mapping Model

We are delighted to see Pierrehumbert's characterization of parallel encoding and target approximation (PENTA) as a third-generation model of prosody and intonation. Indeed, much of the refinement PENTA may potentially bring to our understanding of prosody has benefited from knowledge gained from empirical research since the earlier models. One of the key insights from empirical findings is that surface prosodic forms, such as F0 peaks, valleys, elbows, whole contours, and so on, cannot be mapped to underlying units, be it tone, stress, pitch accents, or prominence. This insight is instrumental in the conceptualization of PENTA and is expressed explicitly in the presentation of the model. Figure 11r.1 is a reproduction of the schematic of PENTA, now with the addition of optional mappings (indicated by curved arrows) to various underlying levels that are more direct than those assumed in the model. Also added is a representation (the cloud on the far left) of all the meanings that could potentially, but not necessarily, be conveyed by speech. As indicated by the crosses, surface prosody (solid curve on the far right) not only cannot be mapped directly to meanings (longest curved arrow), but also cannot be directly linked to communicative functions, encoding schemes, underlying articulatory targets, or even the target parameters. In fact, at least three degrees of separation were recognized when PENTA was first proposed: articulatory implementation, target assignment, and parallel encoding (Xu 2004a, 2004b). In other words, the very premise of PENTA is that surface "phonetic outcomes" are not mapped directly to meanings. Of course, it is not enough to just point out the mismatches between meaning and phonetic outcomes. PENTA is about how meanings can be ultimately mapped to surface prosody through specific connection mechanisms so that there are no missing conceptual links. This means that each of the three degrees of separation needs to be explicitly represented in the model. Very broadly, as shown in figure 11r.1, meanings are first conventionalized into communicative functions, each having an encoding scheme that has been developed through many rounds of conversational interactions. The encoding schemes of all functions work in parallel to jointly determine a single sequence of targets. These targets are then articulatorily implemented through nonoverlapping sequential target approximation to generate continuous surface acoustic events.

This conceptualization indeed deviates from what Pierrehumbert, in her commentary, calls "modern linguistic theories" of prosody in various ways. In particular, two ideas offered by PENTA, which are mentioned in the main essay of this chapter, are worth recapitulating. The first is that the function-form relation, as formulated by de Saussure (1916), needs a major refinement. The second is that parametric

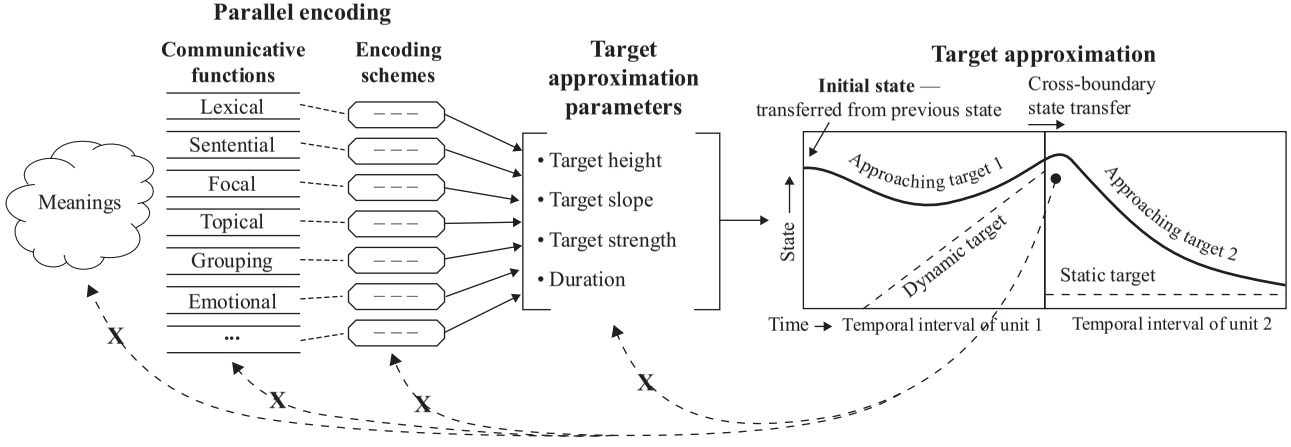


Figure 11r.1

representations should replace symbolic representations as the final link to surface phonetics. These points are elaborated in this response.

Why Function First?

De Saussure's (1916) notion that linguistic units are unities of signified and signifier does not make it clear what to do if there are uncertainties about both the signifier and the signified. This vagueness has not been a major problem for segmental phonemes because their function is relatively straightforward: to differentiate words. Thanks to people's strong intuition about words, the only major uncertainty is whether a particular segment does or does not distinguish certain words in a particular language. In prosody, both the form of the contrastive units and their functions are often ambiguous, as can be seen in the lack of consensus on both after decades of research. It is thus tempting, and has been tried many times, to first develop a descriptive account of easily observable surface prosodic features such as peaks, valleys, shapes, contours, and overall trends (Bolinger 1986; Crystal 1969; Grabe, Kochanskiand, and Coleman 2007; 't Hart, Collier, and Cohen 1990) with the hope that their meaning associations can be determined by further research. Likewise, units such as pitch accents, phrase accents, and boundary tones were originally summarized from "observed features of F0 contours" without explicit association with meanings, as is made clear in Pierrehumbert (1980, 59). Although there have been later efforts to link them to pragmatic meanings such as truth condition and common ground (Pierrehumbert and Hirschberg 1990), the proposed prosodic units remain primarily defined by their forms, as is evident from the fact that transcription of pitch tracks is used as a major means of prosody analysis (Silverman et al. 1992).

What is overlooked in these approaches is that this is *not* how segmental phonemes are determined. While it is true, as Pierrehumbert notes in her commentary, that "each language has a relatively small inventory of phonological units" ("Introduction"), whether a particular segment should be considered as a phoneme has to be determined by whether it serves to make any specific lexical contrasts rather than whether it sounds sufficiently different from other segments (Swadesh 1934). In other words, a highly specific functional contrast is the primary determinant of the phonemic status of the segment.

What may have made the segmental phonology different from prosody is what is known as *duality of patterning* (Hockett 1960), which is the essence of phonology as a bottleneck that, as Pierrehumbert notes, "helps the language learner to acquire a large vocabulary by allowing articulatory and perceptual patterns exhibited in one word to be reused in other words" ("Introduction"). Here the key word is the *reuse* of the same phoneme in different words, for example, the vowel /i/ in *bin*, *pin*, and *tin*, and the consonants /b/ and /n/ in *bin*, *ban*, and *bun*. Note, however, that the reuse is within the same function, that is, lexical contrast. An appropriate comparison in prosody would be the reuse of on-focus expansion and postfocus compression (PFC) of pitch range in foci at different sentence locations (Xu, Chen, and Wang 2012. But the reuse of the same phonetic feature would not work across functions. It would be hard to claim, for example, that because a postfocus high (H) tone has the same pitch level as a prefocus low (L) tone, the [low] feature is shared between the focus function and the lexical function. In other words, it is unlikely that there is a function-independent phonological /Low/ floating around in its own right, because the [low] is only relative to other tones within the same lexical contrast function.

As recognized by Hockett (1960), duality of patterning is due to heavy crowding in the lexical contrast function, as the number of words that need to be encoded massively exceeds the number of possible distinct segmental categories. Prosody, in contrast, confronts a different kind of crowding, that is, each prosodic dimension, for example, F0, is shared by many functions: lexical, focal, phrasal, topical, sentential, attitudinal, emotional, social-indexical, and so on. To make things worse, the identity and nature of these functions are not clear, given the lack of reference in the form of words, either spoken or written. Faced with this difficulty, PENTA-based research has followed a function-first principle that goes beyond the simple function-form relation envisaged by de Saussure. That is, the task of prosody modeling is to find out whether a particular set of meanings has been conventionalized into a communicative function, and what the encoding scheme of this function is like in terms of how the various prosodic dimensions are utilized to encode its internal categories. Following this principle, observable prosodic forms are always treated as a secondary property, that is, a means of encoding the function-internal categories. This is why PENTA-based studies never use prosodic transcription as a method of prosodic analysis.

Hypothesis Testing by Controlled Experiments

Identifying communicative functions and their encoding schemes is by no means a trivial task. The multiple degrees of separation depicted in figure 11r.1 means that not only are surface acoustic events not directly mapped to meanings, but also no two adjacent levels are linearly related to each other to allow analysis by inversion, that is, deriving the underlying form directly from surface properties. Starting from the right end of figure 11r.1, target approximation, implemented as a generative model in the form of quantitative target approximation (Prom-on, Xu, and Thipakorn 2009), cannot be mathematically inversed to derive the underlying targets. So our modeling work has always used analysis-by-synthesis to estimate the underlying targets (Prom-on, Xu, and Thipakorn 2009; Xu and Prom-on 2014). And even with this approach, the quality of the target estimation is correlated with the size of the training corpus. This means that it is simply impossible to derive authentic underlying targets from single utterances.

Moving leftward to the link between underlying targets and the encoding schemes, any single target is the end result of joint contributions by multiple encoding schemes, which makes it impossible to derive all the contributing encoding schemes from an estimated target, no matter how accurate the estimation may be. Even within an encoding scheme, a large portion of it consists of conventions that stipulate arbitrary context-sensitive assignment of the target parameters (referred to by Pierrehumbert as "language-specific constraints"). For Mandarin, for example, the low tone would assume a rising-tone–like target if it is followed by another low tone. This means that even if a contour is correctly recognized as related to a rising tone, the underlying morpheme could be either one with the low tone or with the rising tone. For English, as found in Liu et al. (2013), whether a stressed syllable is assigned a high or low-rising target depends on its position in word, focus status, and the modality (question or statement) of the sentence. This again means that it is impossible to derive individual functions even from the estimated targets.

Finally, as indicated at the far left of the figure, not all possible meanings have conventionalized functions. It is therefore impossible to know, a priori, whether a potential meaning, no matter how useful it may seem (e.g., truth condition and common ground), can be mapped to a specific encoding scheme. For example, seven different

types of focus have been suggested in Gussenhoven (2007). But so far, not even the two most obviously different types, namely, information focus and contrastive focus, have been demonstrated to be consistently distinct from each other in their prosodic realizations (Hanssen, Peters, and Gussenhoven 2008; Hwang 2012; Katz and Selkirk 2011; Kügler and Ganzel 2014; Sityaev and House 2003).

In the face of so many levels of indirect and nonunique mappings, the only viable method of discovering whether a potential meaning has developed a conventionalized function, and what the encoding scheme of that function is like, is hypothesis testing by controlled experiments. In this paradigm, both the function and the encoding schemes are treated as hypothetical, and experiments designed to systematically manipulate the functional content are performed. In the end, it is the outcome of the experiments, which often requires multiple studies, that can inform us, with various levels of certainty, of the presence of a function and the internal structure of its encoding scheme. It is with this approach, for example, that it is determined that the most salient encoding feature of prosodic focus is PFC of pitch range and intensity in many languages and that PFC is nevertheless fully absent in many other languages (Xu, Chen, and Wang 2012).

Even with controlled experiments, however, there is an issue of whether function- or form-defined units should be the target of testing. For example, when pitch accent is targeted in some controlled studies (e.g., Grabe et al. 2000; Shue et al. 2010; Turk and White 1999), the method of elicitation is the same as those used in studies of focus, that is, question-answer or negation paradigms (Cooper, Eady, and Mueller 1985; Eady and Cooper 1986; Liu et al. 2013; Patil et al. 2008; Wang and Xu 2011; Xu and Xu 2005). Due to the presumption of pitch accents as phonological units, these studies either examine phonetic properties of the focused words only or treat those of postfocus components as due to phrase accent or boundary tones that are independent of the nuclear pitch accents.

From the perspective of the function-first principle, pitch accents are merely a phonetic property, as they are identified by the presence of local F0 peaks, valleys, or movements that sound and/or look prominent, which may or may not be due to focus. For example, a prominent F0 peak may occur at the beginning of an utterance even in the absence of an initial focus (Wang and Xu 2011). Or, a prominent pitch movement may occur near the end of a sentence, which would, by definition, be treated as a nuclear pitch accent. But both production and perception studies have shown that these peaks would neither be always intended nor perceived as a sentence-final focus (Cooper, Eady, and Mueller 1985; Rump and Collier 1996; Xu and Xu 2005). Furthermore, focus may not always be marked by an F0 peak more prominent than that in a neutral-focus sentence, as found in Turkish (Ipek 2011). This is not surprising, because the presence of PFC (which is attributed to deaccenting and/or an L-phrase accent in the autosegmental-metrical [AM] theory) already enables successful perception of focus (Ipek 2011; Rump and Collier 1996; Xu, Xu, and Sun 2004). Focus, therefore, is empirically attested as a communicative function marked by multiple phonetic cues, including on-focus increase of pitch range, intensity, and duration, and postfocus reduction of pitch range and intensity (Xu 2011), with a temporal domain that expands even across a silent phrasal pause within a sentence (Wang, Xu, and Ding 2018). In contrast, pitch accent, even when seemingly obvious, is only one of such cues, which may not even be the most critical cue, because the presence of an F0 peak later in the utterance would effectively block the perception of an early focus (Rump and Collier 1996). It would therefore be difficult for PENTA to equate focus with nuclear accent in the phrase, as suggested in Pierrehumbert's commentary.

By the same token, boundary tone, as a cue to sentence modality (question versus statement), is also only one of the phonetic markers of the contrast, rather than being a phonological unit in its own right. For American English, at least, the marking of modality involves not only a sentence-final F0 rise or fall, but also a drastic raising or lowering of postfocus F0 register (treated as due to an independent phrase accent in the AM theory), and a change of target height and target slope of all stressed syllables throughout the sentence (Liu et al. 2013).

Economy of Representation and Degrees of Freedom

The kind of controlled experiments involved in typical empirical studies, however, can go only so far as identifying the functions and the gross patterns of their encoding schemes. To be able to account for the full details of surface prosody, a further step is needed to establish a form of representation that can generate real speech–like continuous prosodic events. This ultimate goal is attempted in PENTA through parametric representation. In this regard, however, PENTA is often criticized for being uneconomical in representation (see Arvaniti, chapter 1, this volume; Arvaniti and Ladd 2009, 2015), given its insistence on (i) pitch target for every syllable even if it is unstressed or bearing the neutral tone, and (ii) full specification of all targets in terms of not only target height (register), but also target slope and target strength, with no allowance for any underspecifications. But we fully agree with Pierrehumbert's remark that "the human cognitive system can learn very detailed patterns and often represents them with a great deal of redundancy" ("Conclusion"). The redundancy is not only in terms of the multiple cues for any specific communicative function, as we've discussed, but also in terms of detailed continuous trajectories that carry massive variability due to articulatory mechanisms, dialectical differences, and idiosyncrasies of individual speakers.

The solution to the redundancy problem explored in the PENTA approach, as detailed in the main essay of this chapter, is model-based parametric representation. *Model-based* means that the representation is meaningful only with respect to a specific computational model. *Parametric* means that targets are specified by numerical parameters rather than symbolic features. The representation of F0, for example, is by numerical specifications of target height, target slope, and target strength, as shown in figure 11r.1. The parameter values are obtained neither by transcription nor by direct acoustic measurement, but by training the computational model on real speech data. Depending on the nature of the training data, the learned targets can be language-, dialect-, or speaker-specific. Our computational studies so far have shown that the approach is able to generate pitch contours that are both natural sounding and functionally contrastive (Prom-on, Thipakorn, and Xu 2009; Xu and Prom-on 2014). And our pilot results based on speech corpora that are less well controlled than typical experimental data have also been encouraging.

Overall, whether a representation is sufficiently economical cannot be measured by the number of representational units assumed by a theory, but by the total specifications needed to generate detailed continuous prosodic events that resemble those of natural speech. If a unit is specified only in terms of H or L, as is the case with pitch accents, phrase accent, and boundary tones, somewhere down the line, there have to be specifications of the exact pitch height, the onset time and offset time of the unit, and how exactly the unit is connected to adjacent units. If underspecification is assumed, sooner or later there has to be a mechanism to generate surface acoustics for

the underspecified units. Without including all these specifications, it is impossible to compare degrees of freedom between different models.

Another way of assessing the economy of a model is to see how many redundant parameters are required. PENTA uses only three free parameters: height, slope, and strength of targets. None of them is redundant, because they are all independently motivated. *Target height* is motivated by its universal recognition; *target slope* is motivated by the consistency of final velocity in dynamic tones (Wong 2006; Xu 1998); and *target strength* is motivated by the sluggish realization of a mid target in the neutral tone in Mandarin (Chen and Xu 2006) and unstressed syllable in English (Xu and Xu 2005). In comparison, the equivalent of target strength in the Fujisaki and the task dynamic models (stiffness) is mostly fixed (Fujisaki 1983; Saltzman and Munhall 1998) and so is largely redundant. On the other hand, the temporal domain of target approximation is fixed to the entire syllable in PENTA (Xu and Prom-on 2015), so that there are virtually no temporal degrees of freedom. This also contrasts with the Fujisaki model (Fujisaki 1983) and articulatory phonology/task dynamic model (Browman and Goldstein 1992; Saltzman and Munhall 1989), where the onset and offset of the commands and gestural scores are free parameters, which means many more degrees of freedom in the temporal domain than PENTA. Given that the AM theory has no strict specifications of tonal alignment, it would also face the problem of degrees of freedom in the temporal domain.

Conclusion

PENTA is part of an effort to develop a new way of conceptualizing the mapping between meanings and continuous acoustic signals in speech, starting from the prosodic aspect. The multifold complexity of prosody has forced us to go back to the first principles to reconsider the phonetic-phonology interface in light of the function-form dichotomy. As a result, PENTA is one of the most indirect models of prosody, as it explicates multiple degrees of separation between meaning and continuous surface prosody. At the same time, it also insists that there be no broken links in the theoretical conceptualization of prosody and intonation and has implemented this tenet by proposing specific connection mechanisms in its computational implementation. What has also emerged from this effort is that model-based parametric representation could be the key to understanding not only the mapping of meaning to continuous phonetic output, but also how the acquisition of speech production is achieved (Xu and Prom-on 2014, 2015).

References

Arvaniti, A., and D. R. Ladd. 2009. "Greek wh-Questions and the Phonology of Intonation." *Phonology* 26 (1): 43–74.

Arvaniti, A., and D. R. Ladd. 2015. "Underspecification in Intonation Revisited: A Reply to Xu, Lee, Prom-on and Liu." *Phonology* 32:537–541.

Bolinger, D. 1986. *Intonation and Its Parts: Melody in Spoken English*. Palo Alto: Stanford University Press.

Browman, C. P., and L. Goldstein. 1992. "Articulatory Phonology: An Overview." *Phonetica* 49:155–180.

Chen, Y., and Y. Xu. 2006. "Production of Weak Elements in Speech: Evidence from f0 Patterns of Neutral Tone in Standard Chinese." *Phonetica* 63:47–75.

Cooper, W. E., S. J. Eady, and P. R. Mueller. 1985. "Acoustical Aspects of Contrastive Stress in Question-Answer Contexts." *Journal of the Acoustical Society of America* 77:2142–2156.

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

de Saussure, F. 1916. "Nature of the Linguistics Sign." In *Cours de linguistique générale*, edited by C. Bally and A. Sechehaye, 66–70. New York: McGraw-Hill.

Eady, S. J., and W. E. Cooper. 1986. "Speech Intonation and Focus Location in Matched Statements and Questions." *Journal of the Acoustical Society of America* 80:402–416.

Fujisaki, H. 1983. "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing." In *The Production of Speech*, edited by P. F. MacNeilage, 39–55. New York: Springer-Verlag.

Grabe, E., G. Kochanski, and J. Coleman. 2007. "Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency." *Language and Speech* 50:281–310.

Grabe, E., B. Post, F. Nolan, and K. Farrar. 2000. "Pitch Accent Realization in Four Varieties of British English." *Journal of Phonetics* 28:161–185.

Gussenhoven, C. 2007. "Types of Focus in English." In *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, edited by C. Lee, M. Gordon and D. Büring, 83–100. New York: Springer.

Hanssen, J., J. Peters, and C. Gussenhoven. 2008. "Prosodic Effects of Focus in Dutch Declaratives." In *Proceedings of Speech Prosody*, edited by P. A. Barbosa, S. Madureira, and C. Reis, 609–612.

Hockett, C. F. 1960. "The Origin of Speech." *Scientific American* 203:88–96.

Hwang, H. K. 2012. "Asymmetries between Production, Perception and Comprehension of Focus Types in Japanese." In *Proceedings of Speech Prosody 2012*, edited by Q. Ma, H. Ding, and D. Hirst, 326–329.

Ipek, C. 2011. "Phonetic Realization of Focus with no On-Focus Pitch Range Expansion in Turkish." In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 140–143.

Katz, J., and E. Selkirk. 2011. "Contrastive Focus vs. Discourse-New: Evidence from Phonetic Prominence in English." *Language* 87 (4): 771–816.

Kügler, F., and S. Genzel. 2014. On the Elicitation of Focus: Prosodic Differences as a Function of Sentence Mode of the Context? *Proceedings of the 4th International Symposium on Tonal Aspects of Languages*, edited by C. Gussenhoven, Y. Chen, and D. Dediu, 71–74.

Liu, F., Y. Xu, S. Prom-on, and A. C. L. Yu. 2013. "Morpheme-Like Prosodic Functions: Evidence from Acoustic Analysis and Computational Modeling." *Journal of Speech Sciences* 3 (1): 85–140.

Patil, U., G. Kentner, A. Gollrad, F. Kügler, C. Féry, and S. Vasishth. 2008. "Focus, Word Order and Intonation in Hindi." *Journal of South Asian Linguistics* 1:55–72.

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

Pierrehumbert, J., and J. Hirschberg. 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Prom-on, Y., S. Xu, and B. Thipakorn. 2009. "Modeling Tone and Intonation in Mandarin and English as a Process of Target Approximation." *Journal of the Acoustical Society of America* 125:405–424.

Rump, H. H., and R. Collier. 1996. "Focus Conditions and the Prominence of Pitch-Accented Syllables." *Language and Speech* 39:1–17.

Saltzman, E. L., and K. G. Munhall. 1989. "A Dynamical Approach to Gestural Patterning in Speech Production." *Ecological Psychology* 1:333–382.

Shue, Y.-L., S. Shattuck-Hufnagel, M. Iseli, S.-A. Jun, N. Veilleux, and A. Alwan. 2010. "On the Acoustic Correlates of High and Low Nuclear Pitch Accents in American English." *Speech Communication* 52 (2): 106–122.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. "ToBI: A Standard for Labeling English Prosody." In *Proceedings of the 1992 International Conference on Spoken Language Processing*, edited by J. J. Ohala, T. Nearey, B. Derwing, M. Hodge, and G. Wiebe, 867–870.

Sityaev, D., and J. House. 2003. "Phonetic and Phonological Correlates of Broad, Narrow and Contrastive Focus in English." In *Proceedings of the Fifteenth International Congress of Phonetic Sciences*, edited by D. Recasens, M.-J. Solé, 1819–1822.

Swadesh, M. 1934. "The Phonemic Principle." *Language* 10:117–129.

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

Turk, A. E., and L. White. 1999. "Structural Influences on Accentual Lengthening." *Journal of Phonetics* 27:171–206.

Wang, B., and Y. Xu. 2011. "Differential Prosodic Encoding of Topic and Focus in Sentence-Initial Position in Mandarin Chinese." *Journal of Phonetics* 39 (4): 595–611.

Wang, B., Y. Xu, and Q. Ding. 2018. "Interactive Prosodic Marking of Focus, Boundary and Newness in Mandarin." *Phonetica* 75 (1): 24–56.

Wong, Y. W. 2006. "Realization of Cantonese Rising Tones under Different Speaking Rates." In *Proceedings of Speech Prosody 2006*, edited by R. Hoffmann and H. Mixdorff, PS3-14-198.

Xu, Y. 1998. "Consistency of Tone-Syllable Alignment across Different Syllable Structures and Speaking Rates." *Phonetica* 55:179–203.

Xu, Y. 2004a. "The PENTA Model of Speech Melody: Transmitting multiple Communicative Functions in Parallel." In *Proceedings of From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, edited by J. Slifka, S. Manuel and M. Matthies, C-91–96.

Xu, Y. 2004b. "Transmitting Tone and Intonation Simultaneously: The Parallel Encoding and Target Approximation (PENTA) Model." In *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, edited by B. Bel and I. Marlien, 215–220.

Xu, Y. 2011. "Post-Focus Compression: Cross-Linguistic Distribution and Historical Origin." In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 152–155.

Xu, Y., S.-W. Chen, and B. Wang, B. 2012. "Prosodic Focus with and without Post-Focus Compression (PFC): A Typological Divide within the Same Language Family?" *Linguistic Review* 29:131–147.

Xu, Y., and S. Prom-on. 2014. "Toward Invariant Functional Representations of Variable Surface Fundamental Frequency Contours: Synthesizing Speech Melody via Model-Based Stochastic Learning." *Speech Communication* 57:181–208.

Xu, Y., and S. Prom-on. 2015. "Degrees of Freedom in Prosody Modeling." In *Speech Prosody in Speech Synthesis—Modeling, Realizing, Converting Prosody for High Quality and Flexible speech Synthesis*, edited by K. Hirose and J. Tao, 19–34. Berlin: Springer.

Xu, Y., and C. Xu. 2005. "Phonetic Realization of Focus in English Declarative Intonation." *Journal of Phonetics* 33 (2): 159–197.

Xu, Y., C. X. Xu, and X. Sun. 2004. "On the Temporal Domain of Focus." In *Proceedings of International Conference on Speech Prosody*, edited by K. Hirose, 81–84.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data is available.
Names: Barnes, Jonathan, 1970– editor. | Shattuck-Hufnagel, Stefanie, editor.
Title: Prosodic theory and practice / edited by Jonathan Barnes and Stefanie Shattuck-Hufnagel.
Description: Cambridge, Massachusetts : The MIT Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021000764 | ISBN 9780262543170 (paperback)
Subjects: LCSH: Prosodic analysis (Linguistics)
Classification: LCC P224 .P739 2022 | DDC 414/.6—dc23
LC record available at https://lccn.loc.gov/2021000764