The PENTA model: Concepts, use and implications

Yi Xu, Santitham Prom-on & Fang Liu

1 Introduction

Speech is a communication system for transmitting information from one human being to another. The information transmitted is rich and multi-faceted, but 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.

2 The conceptual framework

Beyond the basic facts stated above, 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 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, i.e., 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 will be provided after a brief sketch of the model.

Figure 1 is a schematic of PENTA in its most general form, i.e., representing not only prosody, but also other aspects of speech (Xu & 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 non-hierarchical 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 specified by the

1 There is some overlap in content between this chapter and Xu et al. (2015).
encoding schemes. These parameters control the articulatory process of target approximation represented by the fourth block. This biomechanical process ultimately generates surface acoustics.

**Parallel Encoding**

<table>
<thead>
<tr>
<th>Communicative Functions</th>
<th>Encoding Schemes</th>
<th>Target Approximation parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td></td>
<td>Target height</td>
</tr>
<tr>
<td>Sentential</td>
<td></td>
<td>Target slope</td>
</tr>
<tr>
<td>Focal</td>
<td></td>
<td>Target strength</td>
</tr>
<tr>
<td>Topical</td>
<td></td>
<td>Duration</td>
</tr>
<tr>
<td>Grouping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Target Approximation**

Figure 1. A schematic sketch of the PENTA model (Xu 2005, Xu and Liu 2012, Xu & Wang 2001).

The TA model, as depicted in the rightmost block in Figure 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.

### 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 (since information has to be coded by differential representation), transitional movements between the states are inevitable and each movement takes time. Two empirical questions therefore need to be answered about these movements. A) How much time do they each take? If the amount of time is largely negligible, then there is little need to take it seriously in theoretical considerations. B) 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 F0 contours have to co-occur with supralaryngeal movements that generate segments, and the two movements have to be coordinated in time. For this, the empirical questions are C) what is the basic mechanism of the temporal coordination? and D) how much freedom does the speaker have in this kind of coordination?
With regard to Question A, i.e., how much time the F₀ movements have to take, the empirical findings (Sundberg 1979, Xu & Sun 2002) have shown that the time needed for F₀ movements due to change of laryngeal state is not negligible. The following quasi-linear relations are found between the size of F₀ movement and the mean minimum time it takes to complete the movement:

\[ \text{Rise: } t_r = 89.6 + 8.7 d \]  
\[ \text{Fall: } t_f = 100.04 + 5.8 d \]

where \( d \) is F₀ excursion size in semitones, and \( t \) is duration in milliseconds. These equations show that it takes about 100 ms to make even the smallest pitch movement, be it rising or falling. Given that the average speech rate is about 5-7 syllables per second, meaning that each syllable takes about 143-200 ms, a significant portion of each syllable has to be used for pitch transitions.

Regarding the manner of the transitional movements (question B), systematic examination of lexical tones in Mandarin produced in connected speech has provided relevant clues. As can be seen in Figure 1, when a tone is preceded by different tones, the corresponding F₀ 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 & Wang 2001).

As for the basic mechanism of the temporal coordination of laryngeal and supra-laryngeal movements (question C), Figure 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 F₀ production, namely, syllable-synchronized sequential target approximation (Xu 2005).

![Figure 2: Mandarin tones produced in various tonal contexts.](image)
male speakers of Beijing Mandarin (five repetitions by each). In all plots, vertical lines indicate syllable boundaries. Adapted from Xu (1999). The tone names H, R, L and F stand for High, Rising, Low and Falling tones, respectively.

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 3, to first generate a pitchless syllable (lower left panel) and a carrierless F₀ contour (upper left panel), and then combine them to form the full surface acoustic signal (right panels). It is imaginable that the F₀ contours in the upper left panel and the formant trajectories in the lower left panel could be first formed in the brain and the articulation process then faithfully reproduces them during articulation. However, as discussed above, much of the F₀ 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 & Xu 2013). Thus the F₀ and formant contours in Figure 3 are mostly a by-product of physical inertia. If, however, the transitions were represented in the brain-generated F₀ and formant commands sent to the muscles, the effect of inertia would have been applied twice!

Figure 3. Left: Continuous F₀ (top) and formant (bottom) tracks of the Mandarin utterance “(bi3) ma2 yi2 wei3 (shan4)” [More hypocritical than Aunt Ma]. Right: Waveform, spectrogram and F₀ track of the same utterance. Raw data from Xu (2007).

Thus, it is unlikely that continuous surface F₀ contours are generated independently 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 (i.e., voice quality) targets. This is illustrated in the left block of Figure 4 for pitch and formants. Here the formant patterns are representations of the corresponding
vocal tract shapes, which 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, i.e., during articulatory execution, the final stage in the production chain.

Figure 4. Left: Hypothetical underlying pitch (top) and formant (bottom) targets for the Mandarin utterance shown in Figure 3. Right: The target approximation (TA) model (Xu & 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. Raw data from Xu (2007).

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 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 post-focus domain. But the lowering is not as fast as in a post-focus stressed syllable (Xu & Xu 2005). This indicates that unstressed syllables are assigned post-focus targets, but with a weak target approximation strength. As found in both acoustic
The PENTA model: Concepts, use and implications

analysis (Chen & Xu 2006, Xu & Xu 2005) and computational modelling (Liu et al. 2013, Xu & 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 & Liu 2006, 2012). This assumption is motivated not only by observations like what is shown in Figure 2, but also by findings that the motor system is able to coordinate multiple movements at a fast speed only by fully synchronizing them (Kelso 1984, Kelso, Southard & Goodman 1979, Mechsner et al. 2001). This synchrony constraint could be further 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 rotate only around certain shared axes, under the control of a single steering wheel. This makes the movements of the wheels fully synchronized, and their degrees of freedom merged. In the case of speech, it is possible that the syllable is in fact an evolved motor synchronization mechanism to solve the problem of multi-gestural encoding of information. As captured by the time structure model of the syllable (Xu & 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 et al. 2009), has not yet fully addressed. The first is the delay in target approximation when a target is dynamic. It is 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, consecutive targets also has yet to be implemented in qTA.

The third issue is the phenomenon of post-low bouncing. That is, after a tone with a very low pitch, the F₀ of the following syllables sometimes “bounces up” before returning to the same level as other tones. The bouncing is the most readily observed if the post-low tone is weak, e.g., when it is a neutral tone (Chen & 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 when lowering pitch (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 post-low bouncing by increasing the amount of acceleration (second derivative of F₀) in the transferred state at the junction between the low tone and the following tone. The simulation worked well (Prom-on et al. 2012), but has not yet been included in modeling tools like qTAtrainer and PENTAtrainer.

Closely related to post-low bouncing is the fourth issue — pre-low 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 (Connel & Ladd 1990, Hyman 1993, Gandour, Potisuk & Dechongkit 1994, Laniran 1992, Laniran & Gerfen 1997, Lee, Xu & 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 F₀ lowering movement, which is similar to drawing back the arm in preparation for throwing an object over a long distance. Pre-low raising is not explicitly incorporated in qTAtrainer or PENTAtrainer. However, its effect can be partially simulated by allowing tonal pitch targets to be sensitive to the upcoming tone in computational modeling.

Finally, it is well known that F₀ is temporally perturbed (mostly upwards) at voice onset after an obstruent consonant (Hombert 1977, Silverman 1986). The perturbation consists of a very brief (lasting about 30 ms) aerodynamic effect (Löfqvist & McGowan 1992, Xu & Xu 2003) and an optional longer-lasting vocal tension effect (Stevens 1998, Xu & Wallace 2004). Again, these mechanisms have not yet been incorporated into our computational tools. However, both qTAtrainer and PENTAtrainer allow (and actually recommend) users to set the entire syllable, including any voiceless consonant, as the domain of target approximation. This strategy has generated F₀ contours that better match those of the original than when target approximation domain is limited to the rime only (Xu & Prom-on 2014).

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 by prosody. But questions are open as to
what exactly those meanings are and how each of them is coded in prosody. 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, i.e., 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 (Bailly & Holm, 2005), holds that, when in doubt, it should be function rather than form that is the ultimate 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 & Cooper 1986). It is therefore not easy to identify through direct $F_0$ observations, as would be done in a form-first approach, 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 always tried to establish function-form unities one by one, 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 we will see next, the prosodic means are not limited to temporally separated units as in the case of lexical items. Rather, they often come in the form of modifications of existing forms that are already specified by other functions. For example, lexical tones already specify how the basic shape of a pitch contour may look like in a syllable. 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 function-form relation in prosody, we have always tried to empirically discover, rather than either theory-internally stipulate 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 record and analyze 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 2011a).

This has been facilitated by a method that has allowed us to systematically compare continuous F\textsubscript{0} contours in fine detail (Xu 2011a, 2013). The key component of the method is the extraction of time-normalized continuous F\textsubscript{0} contours. Time-normalization takes the same number of F\textsubscript{0} measurements from each equivalent temporal interval, e.g., 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\textsubscript{0} contours in Figure 2, because the contextual variants of the same tone with complete trajectories are pitted against each other in the same graph.\textsuperscript{2}

These strategies have allowed us to see how it is possible for F\textsubscript{0} to encode multiple communicative functions simultaneously. As can be seen in Figures 5-6, 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\textsubscript{0} contours. These sentences were spoken with sentence-initial (in Figure 5 only), sentence-medial or sentence-final focus, and as either statements or questions. As can be seen, the realization of focus and modality exhibit F\textsubscript{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 post-focus 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 micro-properties of syllabic pitch targets. For on-focus word-final stressed syllables, the target slope is falling in statements, but rising in questions (job in b, d). For on-focus, non-final 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.

\textsuperscript{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.
The PENTA model: Concepts, use and implications

Figure 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. On the left, either focus is on the sentence-initial word (thick lines), or there is no narrow focus (thin lines). On the right, focus is either sentence medial (thick lines) or sentence final (thin lines). The solid lines represent statements, and the dashed lines represent questions. Data from Liu and Xu (2005).

Figure 6. Mean $F_0$ contours of statements (S) and questions (Q) in American English. The word after “/” is focused. Data from Liu et al. (2013).

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, grouping or phrasing, for example, is primarily marked by timing properties, which mainly include syllable and pause duration. In general, a longer preboundary duration signifies greater separation of two adjacent groups (Wagner 2005, Xu & Wang 2009). Toward the end of a sentence, word duration increases exponentially (Yuan, Liberman & Cieri 2006). It is further hypothesized that syllable duration is
combined with silent pauses to serve as an affinity index that iconically encodes relational distance with temporal distance (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, based on a set of hypothetical bio-informational dimensions, and that such encoding is again done in parallel with other, more linguistic functions (Chuenwattanapranithi et al., 2008, Hsu & Xu 2014, Liu & Xu 2014, Noble & Xu 2011, Xu et al. 2013a, Xu et al. 2013b).

There are also various other communicative functions that may be prosodically encoded. But as reviewed in Xu (2011), 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. This general strategy is perhaps where the PENTA approach differs the most from the AM approach. According to Pierrehumbert’s (this volume) comment on the present chapter, AM theory-internally “defines a bottleneck between the phonetic realizations of words and their meanings” in terms of a “small number of tonal elements”. Those tonal elements are all derived from direct observations of F₀ contours (Pierrehumbert, 1980). PENTA, in contrast, only implicitly assumes that there must be articulatory, perceptual and historical bottlenecks that prevent plausible prosodic marking of many communicative functions, and so the existence and properties of these bottlenecks can only be discovered through controlled experiments. They should therefore not be stipulated, a priori, as part of the core of PENTA. In this sense, encoding schemes in PENTA are not free, but restricted by empirically established reality.

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 et al. 2009). The current computational version is quantitative target approximation (qTA), which is a third-order critically damped linear system, as shown in the following formula:

\[ f_0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \]  

Here, \( f_0(t) \) is the surface F₀ of a syllable as a function of time. \( mt + b \), the first term, is the underlying pitch target as a linear function of time, 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 F₀ dynamic state and pitch target of the current syllable. Parameter \( \lambda \) represents the strength of the F₀ movement toward the target.

qTA also explicitly represents the state transfer between adjacent syllables by taking the final F₀ state of the preceding syllable in terms of its final F₀, \( f_0(0) \), 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,

$$c_1 = f_0(0) - b$$
$$c_2 = f_0'(0) + c_1\lambda - m$$
$$c_3 = \left( f_0''(0) + 2c_2\lambda - c_1\lambda^2 \right) / 2$$

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, i.e., 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. Worth particular mentioning is the specification of target slope ($m$), which is not used by any other models that we know. The need for this parameter is based on findings that the final slope is one of the most consistent properties of a tone, as mentioned in 2.1. The third principle is that the model parameters should be learnable from real speech data, so that there will be no missing link 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 (There is no mathematical restrictions, however, on using units other than the syllable. But in our own practice, we have always used the syllable, Xu & Prom-on, 2014), which differs from superpositional models that compute multiple layers of $F_0$ contours and then combine them. This scheme also means that for each syllable, there is only a single target that needs to be estimated (but see discussion in 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.

3.1 Computational modeling tools

Since the proposal of qTA, four computational tools have been developed to enable its conceptual exploration and quantitative testing. qTA_demo1 (www.homepages.ucl.ac.uk/~uclyyix/qTA/) and qTA_demo2 (www.homepages.ucl.ac.uk/~uclyyix/qTA_demo2/) are both web-based interactive programs that demonstrate how qTA works. Their interactive features make them convenient tools for a quick impromptu test of an idea or a prediction based on the TA model.

The other two tools, qTAttrainer (Xu & Prom-On 2010-2015) and PENTAttrainer (Xu & Prom-on 2014), are data-driven modeling programs (available at www.homepages.ucl.ac.uk/~uclyyix/qTAttrainer/ and www.homepages.ucl.ac.uk/~uclyyix/PENTAttrainer/). Both use machine learning algorithms to automatically extract target parameters from real speech data through analysis-by-synthesis. These learning algorithms are based on the analysis-by-synthesis paradigm whereby candidate targets are iteratively tested by using 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 et al. 2009, Prom-on & Xu 2012). The quality of the $F_0$ generation is assessed by three means: a) root mean squared errors (RMSE), which measures the discrepancy of the synthetic contours from the original contours in terms of point-by-point height difference, b) Pearson’s $r$, which assesses how closely the overall shape of the synthetic contours correlates with that of the natural contours, and c) perceptual evaluation in terms of category identification (e.g., tone, focus, etc.) 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 ($b$, $m$, $\lambda$ in equation 1). 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 errors) is selected as its pitch target. An example of such resynthesis is shown in Figure 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, e.g., all the on-focus H tones that occur at the beginning of a sentence (Prom-on et al. 2009). As found in Prom-on et al. (2009) and Liu et al. (2013), good predictive results can be obtained for both English and Mandarin with this approach.

Figure 7. Original (dotted) vs. resynthesised (solid) $F_0$ contours of the English utterance “You’re going to Bloomingdales with Alan”, by qTAtrainer. [X-axis: Time in seconds, Y-axis: $F_0$ in semitones.]
The categorization by averaging strategy employed in qTAt trainer, 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 \( \lambda \) may not be adequately estimated if there is severe target undershoot. Besides, the exhaustive search implemented in qTAt trainer is inefficient and probably ecologically unrealistic as a learning algorithm. These problems are addressed by PENTA trainer, in which function-specific targets are learned from an entire corpus that has been functionally annotated (Prom-on & Xu 2012, Xu & 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 8 shows an example of an annotated utterance (top) and natural F0 and synthetic contours (bottom), where the latter is generated with categorical targets learned from an entire corpus.
The PENTA model: Concepts, use and implications

Figure 8. Snapshots 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 MĂI 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) F0 contours. It also allows users to play the utterance with either synthetic or natural prosody (Prom-on & Xu 2012).

In Xu & Prom-on (2014), good overall numerical results were achieved with PENTAtrainer for English (the same dataset tested with qTAtrainer in Liu et al. 2013), Mandarin and Thai. In Prom-on et al. (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 F0 and duration) was heard to be as natural as natural prosody for English, and only slightly worse for Mandarin.
Interestingly, the total number of function-specific parameters learned from the speech corpora and used in the predictive synthesis was very small. In Xu & Prom-on (2014), 78 parameters (i.e., 26 $b$, $m$, $\lambda$ values each) were used for 960 English sentences (consisting of 8640 syllables), 84 parameters for 1280 Mandarin sentences (consisting of 10240 syllables), and 30 parameters for 2500 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, e.g., 5 parameters per Standard Chinese tone in the Fujisaki model (Fujisaki 1983) and 4 parameters per intonational event in the Tilt model (Taylor 2000).

### 3.2 Modeling encoding schemes of English prosody — An illustration

The application of the computational tools described above has allowed us to model some of the major prosodic encoding schemes in English and Mandarin. Figure 9 provides a summary illustration with modeling data on English from Xu & 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 realizations of multiple encoding schemes of lexical stress, focus and modality in American English, as described in Xu & 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). Secondly, 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).
The PENTA model: Concepts, use and implications

b. You want a **job** with **Mi**cro **soft.**

d. You want a **job** with **Mi**cro **soft?**
Figure 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. For stress, $u =$ unstressed, $S =$ non-final stressed, and $s0 =$ word-final stressed. For syllable position (labeled as Final), $n =$ non-final, $sf =$ semifinal, and $f =$ sentence final. All the graphs are screenshots of the demo window of the Synthesis tool (*synthesize.praat*) in the PENTAtrainer package. Data from Xu and Prom-on (2014).

Note also that the match between the synthetic and original $F_0$ contours in Figure 9 is not nearly as good as that in Figure 7. This is partly because here the synthesis is predictive, based on categorical parameters learned from all the utterances by a speaker in a corpus, as opposed to resynthesis in Figure 7 (by qTAttrainer), but partly also because there is still room for further adjustments in our functional annotations. For example, since 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 *-cros*oft does not show final upstep in Figure 9d. Thus, even if the major characteristics of the encoding schemes have been identified, their detailed properties are still an object of continuous empirical investigations.

## 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 in regard to its quantification, we have tried to ensure the operability of the model, i.e., the ability to take input data in a sufficiently detailed form and generate outputs that are significantly closer to surface acoustics than the input data, based on mechanisms that are biomechanically plausible. This operability is not just about how the model works internally, but also about how it links with 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.

### 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
The PENTA model: Concepts, use and implications

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 faithful quantitative implementation is difficult to falsify, because it cannot generate predictions that are detailed enough to be compared with the natural-occurring patterns or with other theories that do have quantitative implementations.

In practice, however, models can be constructed at different levels and generate predictions with different degrees of details. Theories like autosegmental-metrical phonology and optimality theory are all 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, generates outputs that are detailed enough for numerical acoustic comparisons with real speech. In other words, they lack adequate predictive power. For certain theories, e.g., the AM-theory of intonation (Ladd 2008, Pierrehumbert 1980, Pierrehumbert & Beckman 1988, Pierrehumbert & Hirschberg 1990), it is not even clear what is the input 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 sufficient level of details allows us to see the exact consequences of each assumption we make in a theory. Also, speakers and listeners both 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 acquired by children and operated by adults. Models that leave these details to theory-external processes cannot be considered as cognitively operable, unless plausible interfaces with the external processes have been explicitly proposed.

4.2 Articulation as part of cognition

The above 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, is not part of cognition (Pierrehumbert 1990). What we have learned from computational modeling with qTAttrainer and PENTAttrainer, however, suggests that this may not be the case. Since 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 PENTAttrainer, for example, during target extraction, the articulatory system represented by qTA is used to repeatedly generate F₀ 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
The PENTA model: Concepts, use and implications

...it, the brain would have to replicate all the peripheral processes, which would lead to the kind of paradox discussed in 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 & Doup 2002, Petkov & 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 & Doup 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 & Douglas-Cowie 1992); and children who cannot generate or hear their own voice prior to puberty experience severe delays in their speech development (Doup & Kuhl 1999, Kamen & Watson 1991, Kuhl & Meltzoff 1996, Locke & Pearson 1990). Thus vocal learning in both songbirds and humans requires the use of the very articulatory system that are later used in mature song or speech production. The natural-occurring biological evidence and modeling evidence, therefore, both 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, of course, 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.

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 combine to generate well-formed intonation contours linked both to meaning and to the 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. Since 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, i.e., 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 like pitch accents, phrase accents and boundary tones (Pierrehumbert 1980), they do share some similarities. First, like phonological entities, encoding schemes are also 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. 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 2.1.

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, i.e., whether the representation is 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 dialectical variations. Finally, it needs to be learnable with testable computational algorithms.

The solution found in the PENTA approach that can satisfy all these requirements is parametric representation in the form of underlying articulatory targets. Table 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 the four requirements mentioned above. First, it is abstract because it can represent 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 dialectical 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 that 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 in terms of simulating biological reality with a reasonable level of plausibility.

---

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, as it does not use transcription as a means of analysis.
Table 1. Comparison of PENTA-based parametric and AM-style symbolic representations.

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

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 perplex 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. Since each target is co-specified by multiple functions, learners rarely encounter mono-functional 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, since 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 & Li, 1967).

Beside tonal context, many other frequently occurring functions can also be the conditioning factors, such as boundary marking, focus, sentence type, etc. For example, as discussed in 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 3.2, in PENTA-based computational modeling, these multi-functional targets can be easily learned from and applied to real speech data (Xu & Prom-on 2014). The same principle is
The PENTA model: Concepts, use and implications

applicable to many other cases as well. If these resemble biological reality, we would have identified a core mechanism behind many of the phonological rules.

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 et al. 1985; Eady & Cooper 1986; Eady et al. 1986), we found that Mandarin Chinese shares a common prosodic feature with American English, which we termed post-focus compression (PFC). That is, the pith range of post-focus 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. 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 13,000 years ago (Bombard 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 2011b, Xu et al. 2012). It is also hard 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 2011b). 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 & Guion-Anderson 2014).

Another potential prosodic typology is about how focus interacts with sentence type. In American English, the pitch range of post-focus word is raised well above the pitch median in questions. But this feature is missing in Mandarin (Liu & Xu 2005, Liu et al. 2013) and not seen 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 post-focus domain (Xu & 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 non-stress 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 prosodically 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 2009; Nichols 1996).

4.5 Encoding schemes as prosodic morphemes (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 multi-syllable 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 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 like 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 post-lexical meanings, we may call them prosophemes.

Our description of prosodic focus so far has made it as 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 et al. 2012). Second, as discussed above, post-focus compression of pitch and intensity or PFC has been found to be language specific, and languages with this feature have been hypothesized to be linked to a common proto-language (Xu 2011b). The encoding scheme of focus in languages like Mandarin and English is therefore multi-componential, language-specific, and with likely historical etymologies, thus bearing all the major properties 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 2.2 and 3.2. Modality also interacts with focus to determine the pitch range of all post-focus 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 post-focus raising of pitch range above the reference, neither does it change the pitch targets of individual syllables (Liu & Xu 2005, Liu et al. 2013). So, again we see clear evidence of multi-componentiality, conditional variability and language specificity in the encoding scheme of modality.
Therefore, the multi-componential coding of the prosodic functions demonstrates that it is the communicative functions, rather than the directly observable surface F₀ 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 like focus and modality. Furthermore, some proposed phonetic implementation rules in AM theory (Pierrehumbert 1980, Pierrehumbert & 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 F₀ corresponding to a high boundary tone H% relative to the preceding H- phrase accent, is part of a continuous upshift of post-focus pitch range to mark a question. This extra raising is therefore morpho-phonological, i.e., being part of a prosopheme, rather than being due to a phonetic implementation rule.

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 & Xu 2011, Xu et al. 2013a, Xu et al. 2013b). The few perceptual experiments we have performed on non-emotional prosody were mainly done with the purpose to verify the production patterns found in those studies (Taheri-Ardali, Rahmani & Xu 2014, Xu et al., 2012; Xu, Xu & Sun, 2004) or to evaluate modeling performance (Prom-on et al. 2009, Xu & Prom-on 2014). To test if 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 et al. 2007a, 2007b, 2009). This is consistent with the fact that perception learning precedes production learning in speech acquisition (Kuhl et al. 1992, Werker & Tees 2005). But more research, both behavioral and modeling, is needed to develop better predictive knowledge about prosody perception, and about how it is linked to the production of prosody.

5 Conclusions

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). Yet the empirically established encoding schemes of communicative functions established under PENTA appear to bear close similarities with lexical morphemes, which gives rise to the term prosopheme. Empirically guided search for function-from unities has also led to our finding of prosody-based typological divisions such as those based on PFC, post-focus 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 bio-informational dimensions (Xu et al. 2013a).

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 multi-functional 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.

References
Encoding emotions in speech with the size code — A perceptual investigation. 
*Phonetica* **65**: 210-230.


Peng, S.-h. (2000). Lexical versus ‘phonological’ representations of Mandarin Sandhi
The PENTA model: Concepts, use and implications


The PENTA model: Concepts, use and implications


The PENTA model: Concepts, use and implications


