Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 46 (2005) 220–251
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech melody as articulatorily implemented communicative functions

Yi Xu *

Haskins Laboratories, New Haven, CT, USA
Department of Phonetics and Linguistics, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK

Received 23 September 2004; received in revised form 31 December 2004; accepted 4 February 2005

Abstract

The understanding of speech melody, i.e., pitch variations related to tone and intonation, can be improved by simultaneously taking into consideration two basic facts: that speech conveys communicative meanings, and that it is produced by human articulators. Communicative meanings, as I will argue, are conveyed through a set of separate functions that are realized by an articulatory system with various biophysical properties. These properties make it unlikely that the melodic functions are encoded directly in terms of invariant surface acoustic forms. Rather, the encoding is likely done through the manipulation of a limited number of articulatorily operable parameters that may be considered as the phonetic primitives. Four such primitives can be recognized for speech melody: *local pitch targets*, *pitch range*, *articulatory strength* and *duration*. The values of the melodic primitives are specified by a set of encoding schemes, each associated with a particular communicative function. The encoding schemes are distinct from each other in the manner of controlling the melodic primitives, which allows multiple communicative functions to be conveyed in parallel. The communicative functions are ultimately converted to continuous, detailed surface acoustic patterns through an articulatory process of *syllable-synchronized sequential target approximation*, which takes the melodic primitives specified by the encoding schemes as the control parameters. This view of speech melody is summarized into a comprehensive model of tone and intonation, namely, the parallel encoding and target approximation (PENTA) model.

Keywords: Intonation; Tone; Pitch accents; Pitch target; Intonation model

* Address: Department of Phonetics and Linguistics, University College London, Wolfon House, 4 Stephenson Way, London NW1 2HE, UK. Tel.: +44 20 7679 5011.
E-mail address: yi@phon.ucl.ac.uk

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.014

1. Introduction

The pitch patterns of speech related to lexical tone and intonation can be jointly referred to as speech melody. Despite extensive research over the decades, many issues about speech melody are still far from clear. In this paper, I will argue that a better understanding of speech melody may be achieved by jointly consider two basic facts: that speech conveys communicative information, and that it is produced by an articulatory system. In other words, as an information transmission system, speech should be viewed as consisting of two subsystems: (a) the information being transmitted, and (b) the tools being used to perform the transmission. Thus a good understanding of speech melody can be obtained only with clear knowledge about both the structure of the melodic information, namely, communicative functions, and the properties of the tools, namely, the articulatory system.

Current approaches to speech melody, however, do not typically make a clear separation between the functional components and the articulatory attributes, although some kind of division is often assumed. In many phonological approaches, a division is made between phonological units, such as pitch accents, phrase accents and boundary tones, and their phonetic implementations. But the phonological units are typically defined as well as organized in terms of forms. That is, the proposed units are specified in terms of their height, shape, perceptual prominence, etc., as well as the location and condition of their occurrences. For example, in the widely accepted intonational phonology known as the autosegmental and metrical approach (AM) or the Pierrehumbert model (Ladd, 1996; Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988), pitch accents are specified as H, L and different combinations of H and L, and there is a grammar of intonation that specifies the occurrence and transformation of these phonological units. These phonological units are considered to be eventually linked to communicative meanings. However, the links are formulated separately from the definition of the accents, and are not treated as part of the core of intonation. In other words, it is assumed that we can have a rather explicit intonational grammar without explicit knowledge about how the individual phonological units are linked to communicative meanings (see Pierrehumbert, 1980: p. 59 for a clear statement about this characteristic of the AM approach).[1]

Current intonational phonology does have an explicit assumption about how tonal units are linked to surface acoustic forms, however. Specifically, H and L are assumed (at least in practice) to correspond directly to F_0 peaks and valleys, i.e., turning points in the observed F_0 trajectories. The heights of these points are given specific values through a number of phonetic implementation rules such as downstep, upstep, final lowering, etc. The F_0 contours between the turning points are assumed to be generated through *interpolation* across the flanking accentual tones. In other words, the final generation of surface pitch trajectories is to connect adjacent F_0 turning points with F_0 lines that are either straight or "sagging."

In many experimental approaches, a different division is often assumed, namely, there is a direct link between communicative functions and surface acoustic forms. Thus the quest is typically to find the *acoustic correlates* of certain communicative functions, such as focus, stress, newness, questions, etc. (e.g., Cooper et al., 1985; Cooper and Sorenson, 1981; Fry, 1958). Such approaches have met criticisms from phonologists, who argue that prosodic meanings are not directly mapped onto acoustic correlates (Ladd, 1996; Liberman and Pierrehumbert, 1984). Instead, as they argued, intonational meanings should be first mapped onto phonological structures, which should in turn be linked to surface acoustic forms through phonetic implementation rules.

In the following discussion, I will argue that these two general approaches are both only

[1] There are efforts within the AM approach to strengthen the link between the formally defined tonal units and communicative meanings, e.g., Pierrehumbert and Hirschberg (1990). Nevertheless, in those efforts meanings are sought for tones that are *pre-established* based on direct formal observations. This is rather different for the establishment of segmental phonemes, for which the ability to distinguish *specific* meanings is a prerequisite rather than an afterthought.

partially right. On the one hand, prosodic functions should indeed be defined and organized mainly in terms of communicative meanings rather than in terms of autonomous phonological structures. On the other hand, there are no direct mappings between communicative meaning and invariant acoustic correlates. And this is true whether the acoustic correlates are described in terms of gradient parameters (e.g., Cooper and Sorenson, 1981) or directly recognizable forms such as F_0 peaks and valleys (Pierrehumbert, 1980). I will also show that the lack of direct link is caused not just by a single factor, but by at least three separate mechanisms: (a) articulatory implementation, (b) language-specific target assignment, and (c) parallel encoding of multiple communicative functions. I will start my discussion with a case where functional and phonological definitions of the melodic units *do* largely coincide.

1.1. The case of lexical tone—Where function and phonology coincide

While not all issues about lexical tones are fully resolved, there is little doubt about either their functional importance or their general phonological status. It is well known that in a tone language the same syllable may represent entirely different words or morphemes when said with different pitch patterns. The syllable [ma] in Mandarin, for example, can stand for "mother", "hemp", "horse" or "to scold" when said with high-level, rising, low-dipping, or falling contours, respectively. Thus the lexical identities associated with the tonal aspect of the syllables leave little doubt as to the functional status of the lexical tones. Because of this, it is widely accepted that tonal distinctions are phonological and tones are phonemic just as consonants and vowels are. This level of agreement allows us to focus directly on the acoustic realization of tone in connected speech.

Fig. 1 shows time-normalized F_0 contours of a set of five-syllable sentences in Mandarin. In each panel the target tones of all syllables are held constant except those of syllable 2, which alternate across High (H), Rising (R), Low (L) and Falling (F) (corresponding to conventional tone names Tone 1–4, respectively). Overlaying these tone sequences in the same plot as done in Fig. 1 allows us to see what remains constant and what varies with the surrounding tones. Several observations can be directly made.[2]

1. Although the tone of the third syllable in each panel remains the same, its F_0 trajectory is robustly different when preceded by four different tones in syllable 2.
2. There is nevertheless some consistency about each tone. In each panel, the four F_0 curves in syllable 3 gradually and asymptotically converge to a linear configuration that seems to be appropriate for the tone based on its traditional description (e.g. Chao, 1968): *high-level* for H, *rising* for R, and *falling* for F, as illustrated by the short straight dashed lines.
3. Assuming that these linear trajectories in some way resemble the desired goals for the tones, then the approximation of each goal starts roughly from the onset of syllable 3 and ends around the offset of syllable 3, regardless of whether the converging configuration is a slope or a horizontal line. Thus for Mandarin at least, the approximation of the underlying pitch target of a tone appears to coincide with the syllable.
4. Coinciding tones strictly with syllables seems to have led to cases of apparent "undershoot" at the tone-internal F_0 turning points (peaks in F and valleys in R) in syllable 3 when the initial F_0 is very different from the targeted extreme points. For example, the peaks in F become rather low in the L F and F F sequences in (c), and the valleys in R become rather high in the H R and R R sequences in (b). Furthermore, the location of these tone-internal turning points also varies: the greater the F_0 excursion in the early part of the syllable, the later the turning point.
5. In (d) the F_0 rise of R in syllable 3 continues for a while into the beginning of syllable 4, where the tone is L. That is, there is a "peak delay" (Xu, 2001). If the goal of the L in syllable 4 is *low-level* as suggested by (a)–(c), then the onset

[2] The L tone was not used in syllable 3 in (Xu, 1999) to avoid the known L-tone sandhi which changes the first L into R, as detailed in 4.1.1.

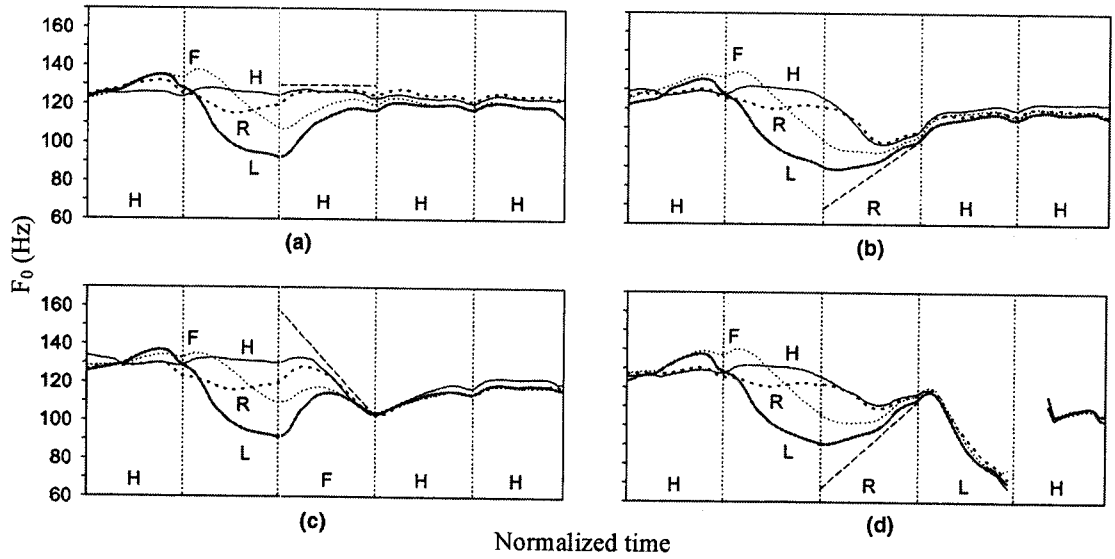

Fig. 1. In (a)–(c) Mandarin H, R and F (in syllable 3) are preceded by four different tones and followed by H. In (d) R in syllable 3 is followed by L. Vertical lines indicate syllable boundaries. The short dashed lines depict hypothetical underlying pitch targets. All plots adapted from Xu (1999).

of the approximation of that goal appears to be delayed by the sharp F_0 rise resulting from the implementation of the previous R.

6. If we view the F_0 variations of tones in terms of contextual influences from adjacent tones, then there is extensive carryover assimilation but little anticipatory variation. In other words, the contextual influences are asymmetrical, with the carryover assimilation fully dominating. There is nevertheless a small amount of anticipatory *dissimilation*: The F_0 of H in syllable 1 is higher when the tone of syllable 2 has a low F_0 point.

The above observations are backed by results of systematic quantitative analyses in (Xu, 1997, 1998, 1999, 2001) for Mandarin. They are also consistent with findings of Gandour et al. (1994) for Thai and those of Li and Lee (2002) for Cantonese. At least two conclusions seem obvious from these observations. First, because neither the overall F_0 height, nor the location and height of the tone-internal F_0 turning points remain constant for any tone in connected speech, *there are no invariant surface acoustic correlates of tones*. Second, what remains constant about each tone across different tonal contexts is an underlying trajectory that is continually approximated throughout the syllable to which the tone is associated.

Thus by systematically examining a melodic function whose communicative meaning and phonological status are both relatively unambiguous, the basic manner of its implementation seems to have been made to emerge. As I will show next, this manner of implementation is likely rooted in the articulatory mechanisms of speech production.

2. Basic biophysical mechanisms

Speech is produced by a biomechanical system consisting of articulators and a nervous system that controls them. As in any motor system, its biophysical properties determine its capabilities. In Section 1 will examine some of the properties that have been found critical for our understanding of speech melody.

2.1. Maximum speed of pitch change

The continually varying melodic patterns in speech are generated by the human larynx, the

organ that produces the fundamental frequency of voice (F_0). If likened to a musical instrument, the larynx belongs to the type that can produce only one note at a time and shift to a new note only after the previous note is over. This means that how quickly two adjacent notes can be produced is dependent on how quickly the vibratory property of the vocal folds, their tension in particular (Fujisaki, 1983; Titze and Talkin, 1979), can be changed. Several attempts have been made to assess this speed (Ohala and Ewan, 1973; Sundberg, 1979; Fujisaki, 1983; Xu and Sun, 2002). One consistent finding of these studies is that the maximum speed of pitch change increases as the size of the change becomes bigger. At the same time, however, the time taken to achieve a pitch change also increases with the size of the change (with the possible exception of lowering F_0 by professional singers (Sundberg, 1979)). Such relation is in general agreement with the kinematic principles proposed by Nelson (1983). The following linear equations were obtained by Xu and Sun (2002) for the mean speed of pitch change and time of pitch change averaged across 36 native speakers of American English and Mandarin.

$$s = 10.8 + 5.6\,d \quad \text{(raising)} \tag{1}$$

$$s = 8.9 + 6.2\,d \quad \text{(lowering)} \tag{2}$$

$$t = 89.6 + 8.7\,d \quad \text{(raising)} \tag{3}$$

$$t = 100.4 + 5.8\,d \quad \text{(lowering)} \tag{4}$$

where s is the average maximum speed of pitch change in semitones per second (st/s), t is the amount of time (ms) it takes to complete the pitch shift, and d is the size of pitch shift in semitone. With these equations, given the magnitude of a particular pitch change, one can calculate both the mean maximum speed of the pitch change, and the average minimum time of the pitch change.

Xu and Sun (2002) also found that, when measured in semitones, male and female speakers do not differ much in the maximum speed of pitch change, nor do American English and Mandarin speakers. Also, pitch falls were found to be a bit faster than pitch rises, but only at magnitudes larger than four semitones. Below four semitones, a pitch rise is faster than a fall. Xu and Sun (2002) speculated that this might have to do with the fact that the pitch raising muscles such as the cricothyroids (CT) are faster but less powerful than the pitch lowering muscles such as the strap muscles.

The findings about the maximum speed of pitch change have many implications for our understanding of speech melody, of which I will mention just two here. First, according to Eqs. (3) and (4), it takes about 100 ms for an average speaker to change pitch by even the smallest amount. This means that a large amount of the contextual variability of a tone as seen in Fig. 1 is articulatorily unavoidable but also predictable, and as such it is neither deliberate nor random. Second, because of the substantial amount of time it takes to make any pitch change, it would be quite hard for the articulatory system to directly control the temporal location of F_0 turning points as has been suggested recently (Arvaniti et al., 1998; Atterer and Ladd, 2004; Ladd et al., 1999; Ladd et al., 2000), for that would entail constant re-calculations just to determine when to start the laryngeal movement toward a turning point based on tonal context as well as many other F_0-affecting factors, as will be discussed later.

2.2. Synchronization of tone and syllable

As discussed in Section 1.1, the implementation of an underlying tonal target seems to start at the onset of the host syllable and end at the offset of the syllable, or, in other words, tone and syllable are implemented in synchrony. Also discussed in Section 1.1 is that such synchrony is often achieved at the expense of full implementation of the tonal targets. As found in (Xu, 1994), such "undershoot" due to the "conflicting" tonal contexts, when severe, lowers the rate of tone recognition. Thus full synchrony is not always desirable so far as tonal transmission is concerned. This further suggests that the synchrony is obligatory rather than optional.

Further evidence for tone-syllable synchrony comes from two sets of findings regarding the F_0 contour variation in both the initial and final portions of a syllable. First, in Mandarin the alignment of F_0 contours does not change depending on whether the syllable has a nasal coda (Xu,

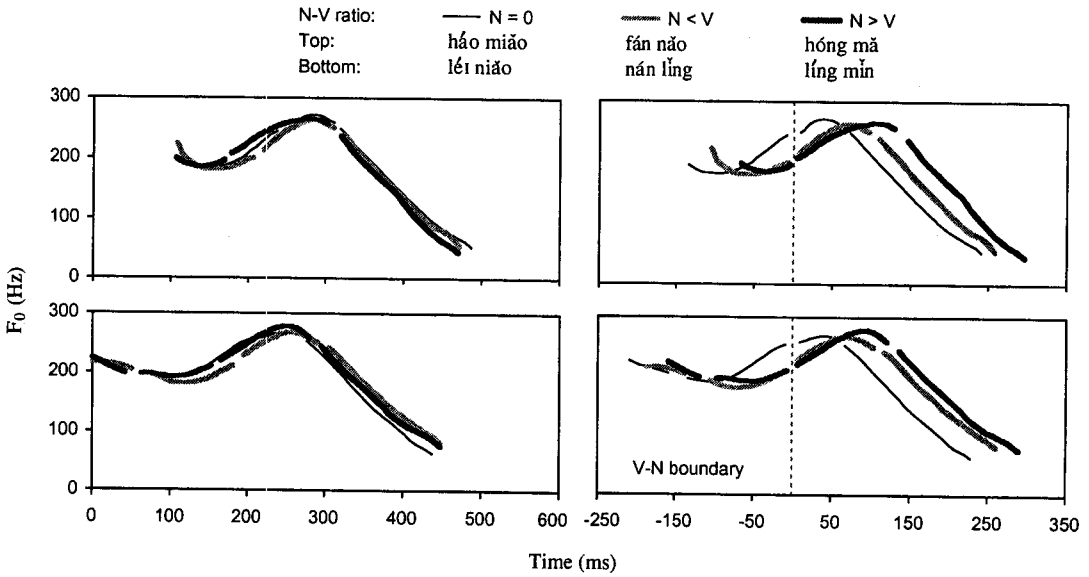

Fig. 2. Mean F_0 contours of disyllabic words in Mandarin in which the first syllable has no nasal coda ($N = 0$), a short nasal coda ($N < V$), or a long nasal coda ($N > V$). The F_0 curves in the left two graphs are aligned to the onset of the first syllable (i.e., at time 0). The curves in the right two graphs are aligned to the offset of the nuclear vowel in the first syllable. Data from Xu (1998).

1998). This remains true even when the nasal coda is longer than the nuclear vowel in cases where the vowel is mid or high. Fig. 2 shows mean F_0 contours of disyllabic words in Mandarin in which the first syllable has no nasal coda, a short nasal coda, or a long nasal coda. The F_0 curves in the left two graphs are aligned to the onset of the first syllable, and all three curves in each graph are very similar to each other in shape and alignment. The curves in the right two graphs are aligned to the offset of the nuclear vowel in the first syllable. There the relative location of the curves becomes increasingly later relative to the nuclear vowel as the nasal coda becomes longer. Since all the tone sequences in the figure are R L, the greater consistency in the left graphs indicates that the alignment there is closer to the underlying alignment of the tone relative to the syllable. Similar alignment patterns have been reported for Thai, where the duration of the nasal coda varies compensatorily with the phonological length of the vowel. Just as the Mandarin case seen in Fig. 2, in Thai "more of the tonal contour falls on the vowel when it is long whereas more falls on the nasal when the vowel is short." (Ohala and Roengpitya, 2002: p. 2285).

The second set of evidence for tone-syllable synchrony is that the underlying alignment of F_0 with the syllable does not change in Mandarin or English even if the initial or final (in English) consonant of the syllable is voiceless. As found in (Xu et al., 2003), the effect of voiceless consonants such as stops and fricatives is to introduce rather local perturbations without changing the carryover or anticipatory tonal variations reported in previous studies such as Xu (1997, 1999). Fig. 3 displays the F_0 contours of Mandarin syllables [ma], [ta], [t^ha] and [ʂa] with the tones R and F. Compared to the F_0 contours in [ma], in which the transition toward the current tonal target is visible, the F_0 curves in syllables with initial voiceless consonants start late and have various amounts of local perturbations at the voice onset. Nonetheless, if these local effects are put aside, the F_0 curves in [ta], [t^ha] and [ʂa] look very similar to those of [ma]. As a result, by the time the local effects are over, F_0 is already quite low in R (left panels) but quite high in F (right panels). So, the implementation of the tonal targets always starts from the beginning of the syllable in Mandarin, whether or not voicing continues through the initial consonant.

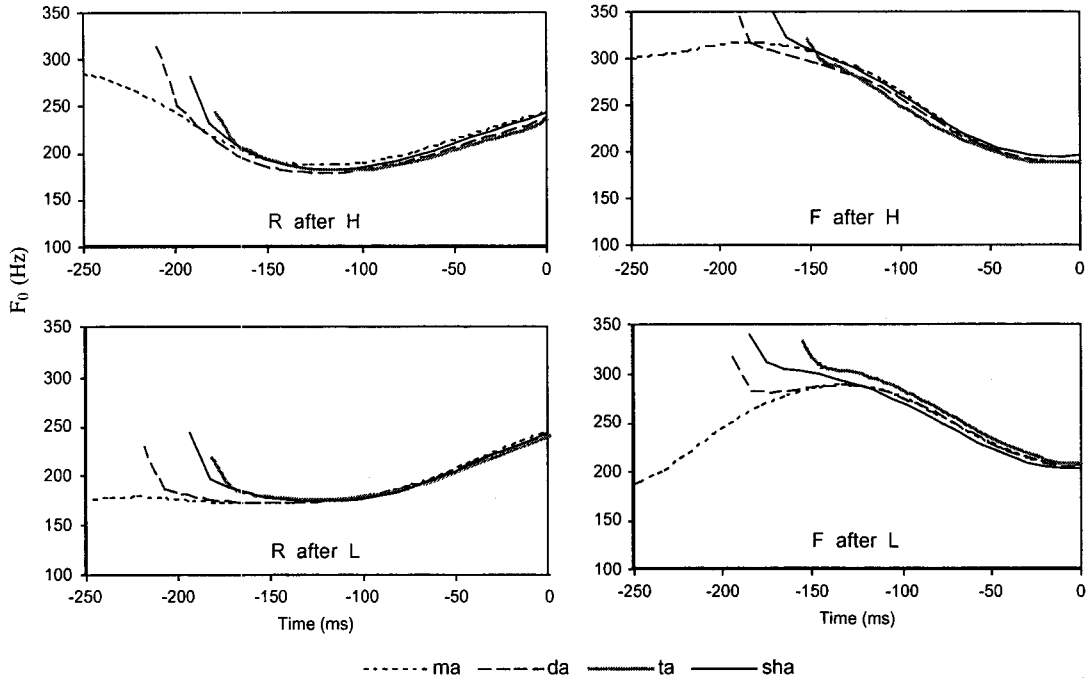

Fig. 3. Effects of voiceless consonants on the F_0 contours of Mandarin R and F produced after H and L. Each curve is an average across five repetitions, two carrier sentences and seven female speakers. All curves are aligned to the syllable offset (i.e., at time 0). "ma", "da", "ta" and "sha" are Pinying spellings of [ta], [t^ha] and [ʂa], respectively. Data from Xu et al. (2003).

For English, Xu and Wallace (2004) compared F_0 contours of syllables with voiced and voiceless obstruents with those with sonorant consonants at word initial, medial and final positions. It was found that F_0 changes across the voiceless interval of the obstruents followed the same trajectory during the sonorant consonants despite apparent local perturbations by the consonants. It was also found that F_0 contours were more consistently aligned to the entire syllable rather than to only the vocalic portion of the syllable. These results suggest that, in English, too, the F_0 movement toward a possible underlying pitch target starts from the syllable onset and ends at the syllable offset.

Why, then, is there such seemingly strict synchrony? In speech, laryngeal movement for producing F_0 patterns and supralaryngeal movements that generate spectral patterns have to be separately controlled, for it makes the independent variations of the melodic and segmental components of speech possible. Separation of controls, however, does not necessarily mean total independence from each other, because there seem to be limited degrees of freedom in executing several movements concomitantly. Kelso (1984) asked human subjects to perform a simple task of wagging two fingers (one in each hand) together. At low speed, they could start the movement cycles of the two fingers either simultaneously, i.e., with 0° phase shift, or with one finger starting earlier than the other by half a cycle, i.e., with a 180° phase shift. At a high speed, however, they could move the two fingers together only with 0° phase shift. Schmidt et al. (1990) further found that the same happened when two people were asked to oscillate their legs while watching each other's movement. Based on such findings, these authors suggest that (a) there is a deep-rooted biological tendency to coordinate one's movement with the environment whenever pertinent, regardless of whether the environment is within the same person or between persons, (b) the 0° phase angle is the most stable phase relation between two coordinated movements, and (c) at high speed, the only way to temporally

coordinate two movements is to lock their phase angle at 0°, i.e., implementing them in full synchrony. If such coordination constraint is a fundamental mechanism in motor movement control, it must apply to speech as well. That is, the synchrony constraint may force the articulatory movements toward local pitch targets to coincide with certain recurrent articulatory cycles. There has been evidence that the syllable serves as such a coordinative structure to which many articulatory movements are aligned (Krakow, 1999; Fujimura, 2000; MacNeilage, 1998). As discussed earlier, pitch movements in speech is often as fast as possible (Xu and Sun, 2002). This means that it would be difficult for a speaker to maintain any phase relation between pitch movement and the syllable other than full synchrony.

2.3. The total pitch range

In the description of tone languages, it is often assumed that lexical tones take up the entire normal pitch range. Thus the H tone is assumed to be at the top of the pitch range and the L tone at the bottom of the pitch range. Data from Xu (1999) indicate, however, that at any particular sentence position, the pitch range across the four Mandarin tones spans no more than one octave, which is only half of a speaker's total pitch range. According to Fairbanks (1959) and Honorof and Whalen (in press), a speaker's conversational pitch range can span as much as *two* octaves. Thus even in a tone language like Mandarin, a full octave of the speakers' total pitch range, mostly in the upper region, is unused by lexical tones, and hence available for other purposes. As we will see later, this is an important fact for the understanding of how multiple communicative functions can be concurrently transmitted.

3. Syllable-synchronized sequential target approximation

The foregoing discussion has introduced several crucial facts about tonal implementation in speech. First, it takes a significant amount of time to complete even the smallest pitch shift, and the time needed increases quasi-linearly with the size of pitch change. Second, no time intervals are preserved just for making transitions *between* tones. Instead, the transition toward each tone occurs in situ, i.e., during the syllable to which the tone is associated. In fact, the F_0 of an entire syllable constitutes a continuous transition toward an ideal pitch pattern of the tone. These facts therefore suggest *syllable-synchronized sequential target approximation* as the basic mechanism of tonal implementation. This mechanism has been summarized by Xu and Wang (2001) into the pitch target approximation (henceforth TA) model. At the core of the model is the assumption that associated with each tone is an ideal *pitch target*, which has a simple form such as static [high], [low] or [mid], or dynamic [rise] or [fall]. The process of realizing each tone is to implement its pitch target by approaching it asymptotically. A schematic illustration of the model is shown in Fig. 4. The vertical lines in the figure represent the onset and offset of two adjacent syllables. The straight dashed lines represent two adjacent pitch targets: a dynamic [rise] and a static [low]. The solid curve represents the surface F_0 contour, which is assumed to be the result of implementing the pitch targets under various articulatory constraints, including the maximum speed of pitch change. During target approximation, at each moment in time, the present articulatory state is compared only to the desired state of the *current* target, and the difference between the two determines the direction and speed of further F_0 movement. Nevertheless, an extensive influence from the last target is naturally entailed because the implementation of that target gives rise to the initial laryngeal state of the current syllable. This "carryover" influence, also naturally, diminishes over time as the current target is being approached. Finally, the approximation of the next target starts as soon as that of the current one is over, but not any time sooner.

As can be seen in Fig. 4, the synchronization of pitch target with the syllable does not always lead to perfect alignment of F_0 turning points with the onset or offset of the syllable. For example, the F_0 valley in syllable 1 resulting from the implementation of the [rise] target may vary in temporal location depending on the F_0 at the syllable onset, as can be seen in syllable 3 in Fig. 1b and d. The

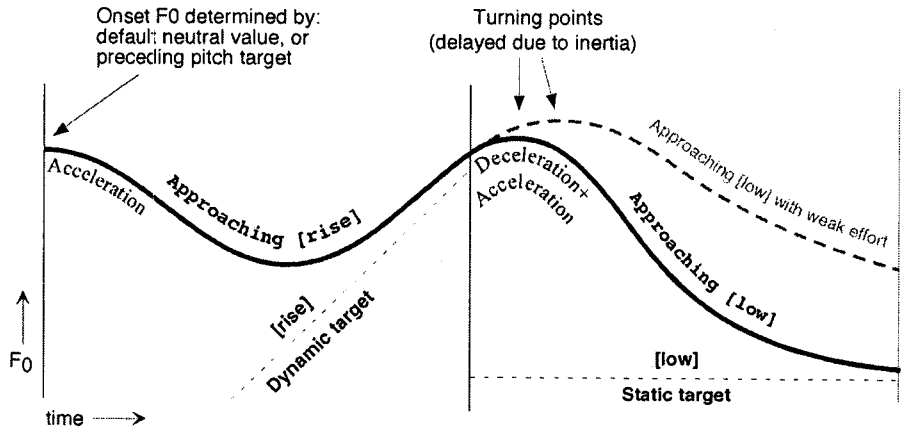

Fig. 4. The target approximation (TA) model. The vertical lines represent syllable boundaries. The dashed lines represent the underlying targets. The thick curve represents the F_0 contour that results from asymptotic approximation of the targets. The dashed curve simulates the effect of weak effort. Adapted from Xu and Wang (2001).

same is true for F_0 peak when the target is [fall] as can be seen in syllable 3 in Fig. 1c. Also, the implementation of [rise] in syllable 1 in Fig. 4 results in a sharp rise by the syllable offset. This sharp rise carries a momentum that takes time to overcome. As a result, although the implementation of the next [low] presumably starts from the onset of syllable 2, F_0 does not start to go down until sometime *after* the syllable onset. Examples of such "peak delay" can be seen in Fig. 1d, and its underlying mechanisms have been discussed in detail in (Xu, 2001, 2002, 2003).

Note that under the TA model, pitch targets serve as some kind of *melodic segments* whose presence and implementation are obligatory. On the other hand, various aspects of the target approximation process may also be independently specified, and they may further determine the additional details of the output F_0. At least four of them, including the target itself, can be identified: (1) pitch target (specified both in terms of shape and height), (2) strength (which determines the approximation speed), (3) pitch range, and (4) syllable duration. In the following discussion, I will refer to these controllable aspects of the TA model as the *melodic primitives*. As is shown in Fig. 5, each of these primitives can have significant impact on the final output F_0, and thus is potentially useful as encoding elements for various communicative functions.

The impact of pitch target is seen in all four panels in Fig. 5. The target sequence in all panels is [high] [low] [high] [low], i.e., four static targets with two alternating heights. The asymptotic approximation of this particular target sequence produces similar up and down patterns in all panels. (The initial F_0 is arbitrarily set at a lower-middle level in all the plots.) In the upper left panel all the other three primitive are assumed to have normal values and the output F_0 there can therefore serve as a reference. The effects of pitch range adjustment are shown in the upper right panel. The pitch range of the first two syllables is both high and wide, while that of the second two syllables low and narrow. Note that the pitch range changes are applied through direct adjustment of target heights.[3] As a result, the local shapes of the F_0 contours remain the same but the magnitudes of F_0 movements are changed considerably. The impact of strength is illustrated in the lower left panel, where the strength of the first and last syllables is strong while that of the middle two syllables weak. As a

[3] This apparently bears the assumption that the adjustment is done before the neural commands are issued to the laryngeal muscles. This is different from the command response models by Öhman (1967) and Fujisaki (1983), in which two continuous pitch curves resulting from muscle responses to two streams of neural commands are generated first and then added together to form surface F_0 contours.

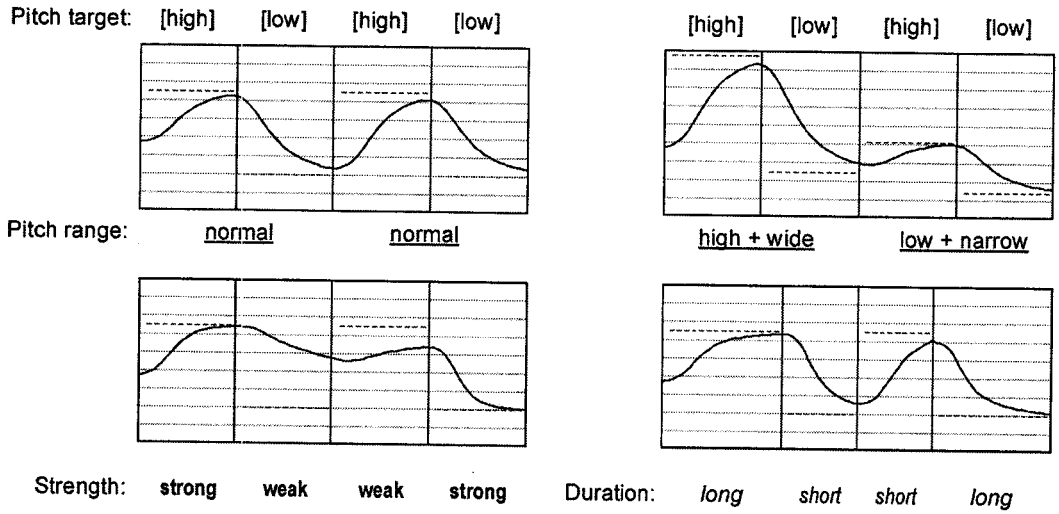

Fig. 5. Illustration of the effects of pitch targets, pitch range, strength and duration.

consequence, the final F_0 in the first and last syllables virtually reaches the targets, but that of the middle two syllables falls far short of their targets, resulting in severe "undershoot." The effect of syllable duration is illustrated in the lower right panel. There the first and last syllables have long durations while the middle two syllables have short durations. Note that long duration gives rise to quasi-steady-state F_0 trajectories. Short duration, in contrast, leads to more dynamic-looking F_0 contours and greater "undershoot."

When the effects of all the melodic primitives are combined, the resulting F_0 contours can be quite complicated. Nevertheless, because their manipulations are all applied to the core mechanism of syllable-synchronized sequential target approximation, the effects of the melodic primitives are predictable and likely recoverable in perception. As I will discuss next, these melodic primitives can be effectively used as the basic encoding elements in transmitting multiple communicative meanings.

4. Encoding multiple communicative functions in parallel

It has been long recognized that many layers of information need to be transmitted through speech melody. This need is probably most strikingly obvious in a tone language. As pointed out by Yip (2002: p. 129), exclamation mark original) "... tonal phonology does not take place in a vacuum, but is frequently affected by the morphology and syntax of the language. As a result, the tonal output may provide the listener with cues that signal, directly or indirectly, such things as phrase boundaries, compound structure, or syntactic category, as well as grammatical information such as tense or mood and of course lexical information!" As we will see later, the list can actually be much longer. Various suggestions, often in the form of analogies, have been made as to how it is possible to transmit multiple communicative functions simultaneously with a single medium, namely, F_0. Bolinger (1964) and Chao (1968) likened the local and global components to small ripples riding on big waves in the ocean. Liberman and Pierrehumbert (1984) and Chao (1932) likened them to contours drawn on elastic graphing paper that can be stretched or compressed. What we need to recognize is that analogies are just that, analogies. The actual mechanisms of intonation are not exactly the same as those of the systems they are likened to. F_0 is not really a wave. Nor is it drawn on graphing paper. Instead, as explained earlier, it has its own specific mechanisms which are in general biophysical. The core mechanism is likely

syllable-synchronized sequential target approximation, as is captured by the TA model. As we have just seen, manipulation of various aspects of the target approximation process, i.e., through the control of the melodic primitives, may have different impacts on the output F_0. What I will discuss in the following is that languages have developed distinct *encoding schemes* that encode individual communicative functions by specifying the values of one or more of the melodic primitives. I will also argue, however, that the encoding schemes can be language specific or universal, and simple or complex, and sometimes even non-unique in their mappings to the communicative functions.

4.1. Encoding schemes involving local pitch targets

A local pitch target (or simply "pitch target") refers to the underlying ideal pitch trajectory associated with a segmental unit, typically a syllable. As discussed earlier, pitch targets may be viewed as the *melodic segments* because their assignment and implementation are obligatory. From a functional point of view, pitch targets have to be linked to some communicative functions. In a tone language, they are directly linked to lexical tones which serve to distinguish words or morphemes (although the mapping is not always one-to-one, as will be discussed next). In some languages, tones may also serve certain syntactic or discourse functions (cf. Yip, 2002 for a review). In so-called pitch accent languages such as Japanese, Lithuanian (Blevins, 1993), Swedish (Bruce, 1977; Gårding, 1982), etc. (also see Yip, 2002 for a detailed review), not every syllable carries a tone. However, as argued by Yip (2002), these languages are not fundamentally different from those considered truly tonal. From the perspective of the present paper, it suffices to recognize that pitch targets are also lexically assigned in these languages. The functional role of pitch targets is much less clear when it comes to nontone languages like English. In the two most influential frameworks of English intonation, the British nuclear tone analysis (cf. Cruttenden, 1997 for a comprehensive overview) and the Autosegmental and Metrical approach (Pierrehumbert, 1980), the basic intonational units are pitch accents, which are linked to words rather than syllables. And, pitch accents are considered to be directly linked to prominence. There is also an attempt to link accent types directly to discourse meanings (Pierrehumbert and Hirschberg, 1990).

4.1.1. Non-unique assignment of pitch targets in Mandarin

To make the issue of local target even more complicated, in many cases the association of the pitch target with a function is not one-to-one. That is, the assignment of targets is not only conventional and language-specific, but also often *non-unique*.

The non-unique assignment of pitch targets is most clearly seen in the case of *tone sandhi*, the phenomenon that the realization of a tone varies with adjacent tones. Chao (1968) described an extensive set of sandhi patterns in Mandarin. Chen (2000) devoted an entire volume to the documentation of various sandhi phenomena in a large variety of languages and dialects. The existence of many of these patterns has been confirmed by instrumental studies, (e.g., Lin et al., 1980; Lin and Yan, 1991; Shen, 1990, 1992; Shih, 1988; Wu, 1982, 1984; Xu, 1997). The mechanisms behind these phenomena, however, remain mostly unclear. In light of the TA model, we may distinguish between two types of contextual tonal variations: those due to *target alternation* and those due to articulatory implementation (*Implementational variation*). Target alternation occurs when the pitch target of a tone is changed *before* being implemented in articulation. Implementational variation, on the other hand, does not involve change of tonal targets. Instead, it occurs when the realization of the *same* target is varied due to articulatory constraints. In the following, I will briefly discuss only the L-tone sandhi patterns in Mandarin. A more detailed discussion of other sandhi patterns can be found in (Xu, 2004a).

Fig. 6 displays Mandarin L produced in different tonal contexts. Three types of L variations can be seen in the figure. In Fig. 6a are the average F_0 curves of the four Mandarin tones said in isolation by eight male speakers. Note that L in this graph has a final rise. Fig. 6b displays disyllabic sequences produced in a sentence-medial position. Here syllable 2 always carries L while syllable 1

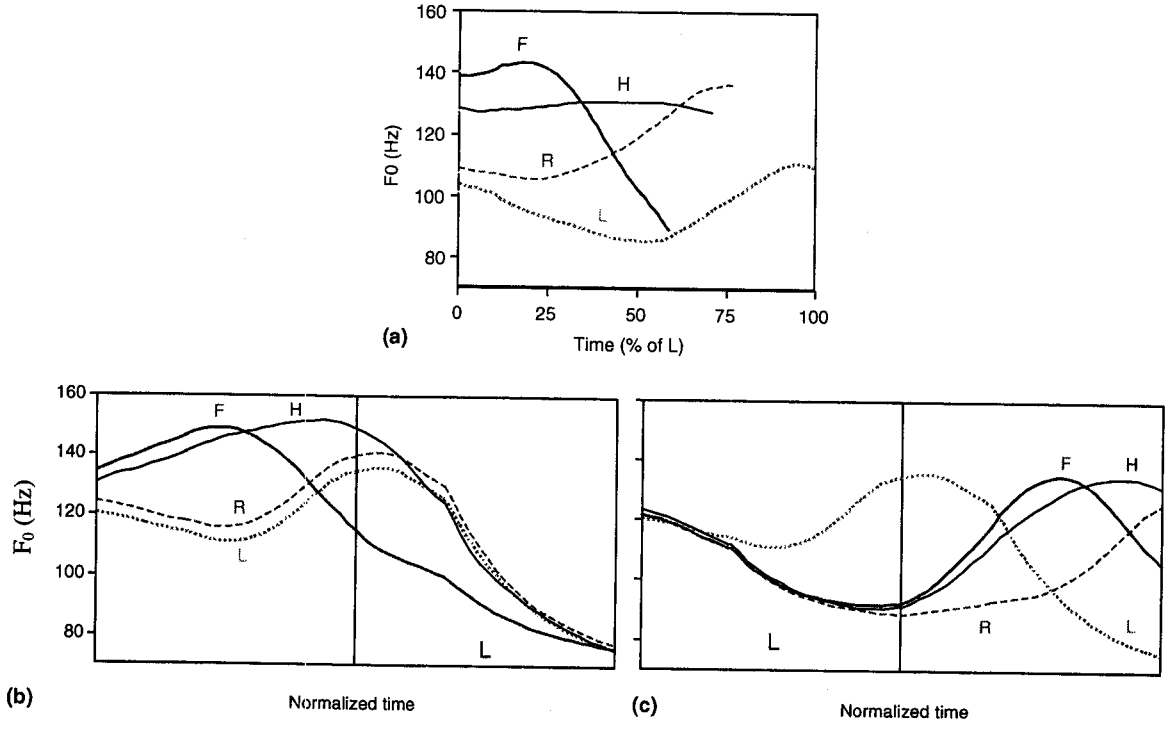

Fig. 6. (a) Four Mandarin tones produced in isolation. (b) Mandarin L after four different tones, produced in carrier phrases. (c) Mandarin L before four different tones, also produced in carrier phrases. All F_0 curves are averages of eight male speakers of Mandarin; adapted from Xu (1997).

carries four different tones. We can see that L in syllable 2 has no trace of the final rise as that in the left graph. As explained in (Xu, 2004a), this absence of final rise cannot be attributed to articulatory constraint. Rather, the pitch target implemented for L in a nonfinal position probably has no final rise to begin with. Hence, the alternation between versions of L with and without a final rise probably involves changes in the pitch targets prior to actual articulatory implementation.

The second type of L variation can be also seen in Fig. 6b. In the L L sequence, the first L is not very different in shape from R in the same syllable, although the two differ somewhat in overall height. Wang and Li (1967) found that Mandarin listeners could not distinguish words and phrases with L L sequence from those with R L sequence. This is further confirmed by Peng (2000) with Taiwan Mandarin speakers. Although acoustic studies have noticed that F_0 values in the L L sequence are not exactly the same as those in the R L sequence (Peng, 2000; Xu, 1993, 1997; Zee, 1980), as is also apparent in Fig. 7b, it is quite clear that the F_0 contour corresponding to the first L in the L L sequence cannot be explained in terms of articulatory implementation of a [low] pitch target according to the TA model, because there is no mechanism in the model for generating a falling-rising contour by asymptotically approaching a [low] target. Fig. 6c shows the L–L sandhi case as compared to L followed by other tones. Note here that the F_0 contour of the first L in the L–L sequence is drastically different from those of other sequences, where there is virtually no F_0 movement in anticipation of an upcoming tone (which is consistent with the absence of anticipatory assimilation discussed in Section 1.1). Given that the F_0 contour of the L L sequence closely resembles that of the R L sequence as seen in Fig. 6b, it is reasonable to assume that a pitch target similar to that of

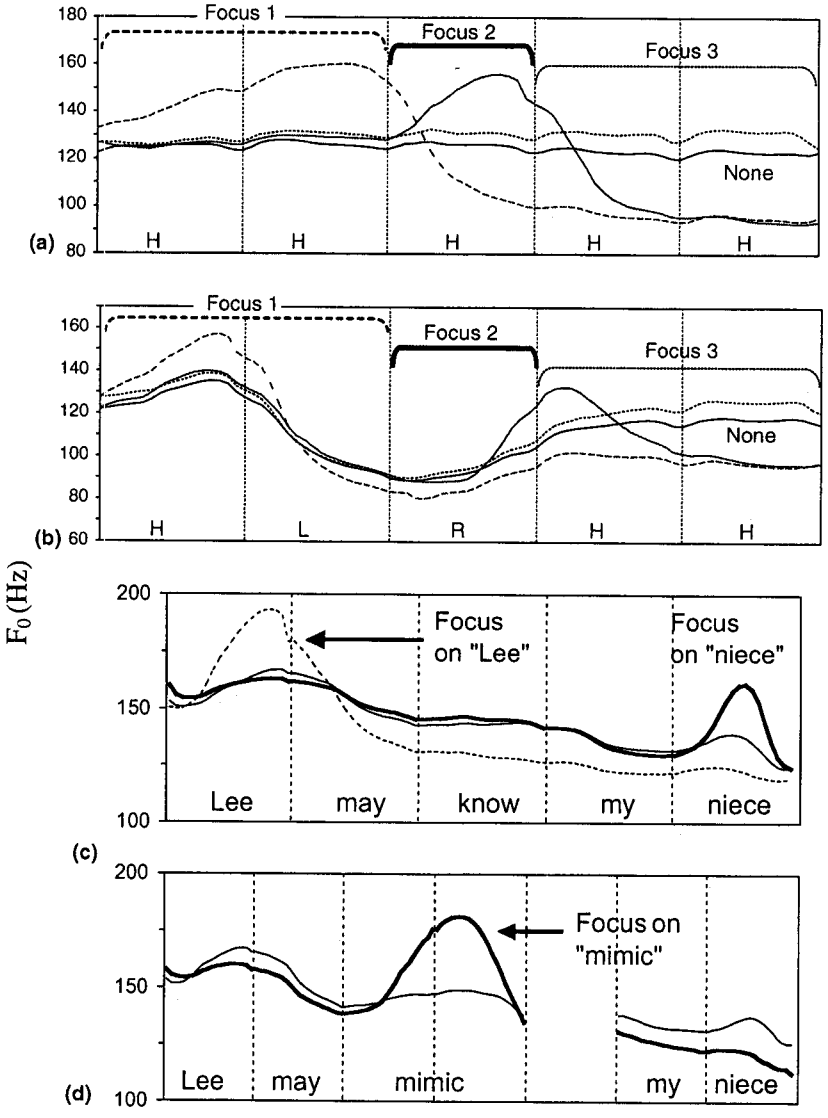

Fig. 7. (a,b) Pitch range variations in Mandarin due to focus in different positions. The curves are time-normalized F_0 averaged across 24 repetitions by four speakers (data from Xu, 1999). (c, d) Pitch range variations in English due to focus in different positions. The curves are time-normalized F_0 averaged across 49 repetitions by seven speakers (data from Xu and Xu, in press).

R is reassigned to the first L in L L *before* its articulatory execution, and that after the reassignment, the execution of the target is just the same as that of any other target.

The third type of L variation can be seen in syllable 2 in Fig. 6b. When syllable 1 has different tones, L in syllable 2 has rather different onset F_0. These variations, because they can be readily explained by asymptotic approximation of the same [low] target when having different initial F_0, are likely directly due to the sluggishness of the F_0 changing process, as assumed by the TA model. They should therefore be considered as cases of *implementational variation.*

Target alternation may have evolved originally due to certain interaction between articulatory

and perceptual mechanisms, including those proposed by Ohala (1981, 1992). However, in many cases the original triggers are no longer recognizable. For example, there have been many attempts to find a plausible mechanism behind the Mandarin L-tone sandhi. However, according to Wu (1982, 1984) and Jiao (2001), similar shifts in tone category involving the same phonological tone occur in many northern Chinese dialects, but the phonetic values of this tone category vary extensively from dialect to dialect. It is thus likely that those dialects, Mandarin included, have inherited the same target alternation that may be originally triggered by an articulatory/perceptual mechanism, but the trigger is no longer present in the contemporary dialects. In some cases, however, the trigger may be still present. For example, Yoruba is known to have two tone spreading rules: (a) H → R/L_, and (b) L → F/H_. Being assimilatory in nature, they could be cases of implementational variation. However, my informal observations suggest that the rules apply even at very low speaking rate and across some pauses. This was confirmed by personal communication with Ian Maddieson. So, it is possible that the pattern has already fossilized into a target alternation rule. A more systematic investigation in which speaking rate is manipulated may help verify whether the variation involves true target alternation.

4.1.2. Pitch target assignment in English

For lack of space, I will mention only three target related issues in English that are important from the perspective of the TA model and the functional view of speech melody. The first is about the functional source of the pitch targets. There is little dispute that within-word pitch height specifications in English do not come from the lexicon. The AM model assumes that pitch accents are directly linked to prominence. Xu and Xu (in press) found, however, that the presence and gross pattern of local F_0 peaks are independent of focus. I will revisit this issue in Section 4.2.1 when discussing focus.

The second issue is about the density of pitch targets in English, i.e., are all or only some syllables assigned pitch targets in the language? In the AM model, only syllables considered to bear pitch accents are given tonal targets, while the F_0 of the unaccented syllables is assumed to derive from interpolation between adjacent pitch accents. Also related to this issue, partly based on the similarity among the F_0 profiles of words like "limb", "limo" and "limousine," it has been argued that the temporal domain of a pitch accent, such as H*L or L*H is a (prosodic) word rather than a single syllable (Pierrehumbert, 2000). Some recent findings have raised questions about the low density assumption, however. I will discuss them in detail in Section 4.3.1. But just briefly here, there is a virtual lack of anticipatory tonal influence in English just like in Mandarin as discussed in Section 1.1. Such lack of anticipatory influence would render interpolation an impossible mechanism for generating surface F_0 trajectories, which in turn suggests the possibility that every syllable in English is assigned a pitch target at the moment of its articulation.

The third issue is about whether pitch target assignment by a function is unique in English, i.e., whether there is an invariant pitch target assignment by a function in all contexts. As will be discussed in Section 4.2.1, Xu and Xu (in press) have argued that the underlying pitch target of a stressed syllable in American English alternates between [high] and [fall] depending on whether the syllable is word final and whether the word is under focus. Thus pitch target assignment in English does not seem to be uniquely related to a single function.

4.1.3. Pitch targets can be both simple and complex

The discussion in Section 1.1 has established that what remains constant about a lexical tone in Mandarin is an underlying target that is continually approximated during its host syllable. The shapes of these targets suggest simple, linear functions. While these unitary linear functions seem to suffice for the four Mandarin tones produced in nonfinal positions, they may not be always fully adequate, even for Mandarin. Also from the perspective of articulation, there is no strong reason why underlying pitch targets cannot have shapes other than those of simple linear functions, as long as the approximation of the target shapes is possible given the maximum speed of pitch change. For example, it is quite likely that the Convex tone in Zhenhai, China, as described in detail by Rose

(1988), has a convex pitch target as opposed to a simple [fall] that may underlie the Falling tone in the dialect. It is also possible to assign two successive pitch targets to a single tone. The Beijing Mandarin L said in isolation or in a pre-pausal position may have either a static [mid] or [mid-high] or a dynamic [rise] following the early [low]. In general, as long as it is recognized that surface F_0 contours do not directly map onto the underlying pitch targets, the exact property of each pitch target in a particular language or dialect is an empirical issue. What is critical is that the contribution of asymptotic target approximation has to be teased out when considering the possible underlying properties of the pitch target.

4.2. Encoding schemes involving pitch range

Pitch range specifies the pitch interval within which local pitch targets are implemented. It can be defined by two parameters: height and span (Ladd, 1996). As discussed in Section 2.3, for an average speaker, the entire exploitable pitch range is quite large, about 2 octaves (Fairbanks, 1959; Honorof and Whalen, in press). This is more than twice as much as needed for lexical tones (Xu, 1999). Thus even in a tone language like Mandarin there is much room left in the pitch range for functions other than lexical tone. As illustrated in Fig. 5, it is possible to shift the height of a pitch target without changing anything else about it. This property of the target approximation process has apparently been put to good use by languages. In the following I will first discuss the use of pitch range specifications by focus and new topic/turn taking. I will then discuss a third function, namely, sentence type, which may also use pitch range as its major encoding element.

4.2.1. Focus

Focus has been increasingly recognized in recent research as an independent linguistic function with robust acoustic manifestations. When the sentence "Mary saw John" is said in response to the question "Who did Mary see?" the word "John" is naturally emphasized, hence, "focused." Focus is therefore emphasis motivated by discourse/pragmatics. A finer distinction is sometimes made between emphatic focus and contrastive focus. The aforementioned sentence would be an example of the emphatic focus. An example of the contrastive focus would be "*Mary* saw John, not Linda" in response to the statement "Linda saw John."[4] Note that in both cases the speaker's choice of whether and where to place focus is based on the assessment of the information flow in the discourse rather than on other concerns (see Bolinger, 1972, 1989; van Heuven, 1994 for more detailed arguments). In other words, the location of focus is *independent* of lexical tone, lexical stress, syntax, and prosodic structure of the sentence, although there may be interactions between focus and these factors. Furthermore, it has been demonstrated that units as small as a single segment can be put under focus (van Heuven, 1994). Thus, neither the location nor the scope of focus is fully predictable solely on the basis of the utterance *containing* the focus, although many sentences can conceivably have a default focus pattern. A wh-question, for example, often attracts a default focus on the wh-component, especially when there is no wh-movement (Ishihara, 2002).

Also, by definition, focus is an utterance-wide function. With focus, the speaker tries to indicate that a particular component of an utterance is singled out against all other components. It follows, then, what is being emphasized is given special articulatory/acoustic treatments, including larger pitch range, longer duration, greater intensity, more expanded vowel space, more clearly enunciated consonants and vowels, and more forcefully implemented pitch targets, etc. What may also follow is that portions of the utterance not being emphasized should be *deemphasized.* As can be seen in Fig. 7a and b, however, deemphasis does not happen evenly in Mandarin in all nonfocused regions as far as F_0 is concerned. Pitch range of post-focus words is compressed extensively. Pitch range of pre-focus words, in contrast, remains largely the same as in utterances without narrow focus. Similar patterns have been found in other tone or pitch accent languages: Shanghai Chinese

[4] See Gussenhoven (in press) for detailed discussion of subtle distinctions of different types of focus and how they are related to syntax and discourse.

(Selkirk and Shen, 1990), Cantonese (Man, 2002), Japanese (Pierrehumbert and Beckman, 1988) and Swedish (Bruce, 1977). The same pattern can be also seen in English in Fig. 7c and d. Although some pre-focus lowering can be also seen there, the lowering was inconsistent and statistically non-significant as found by Eady and Cooper (1986) and Xu and Xu (in press). Also it can be seen in Fig. 7c and d that focus neither introduces F_0 peaks that are totally absent without focus nor completely eliminates peaks that are originally there without focus. Rather, as found in (Xu and Xu, in press), focus only modifies the pitch ranges of the existing F_0 peaks and valleys in English.

There is increasing evidence that the asymmetrical pitch range modification as seen in Fig. 7 is a universal trend, though the trend seems to manifest in different ways in different languages due to its interaction with other aspects of the language. If a language has lexical tones, focus only modifies the *pitch range* of individual syllables (Jin, 1996; Xu, 1999). If the language has lexical stress but no tone or pitch accent, focus may modify both the pitch ranges and the pitch target of the focused syllables. In English, for example, the F_0 of a nonfinal monosyllabic or iambic word peaks after the syllable offset when not under focus but before the syllable offset when under focus (Silverman and Pierrehumbert, 1990; Steele, 1986; Xu and Xu, in press). This is in contrast with a trochaic word where the F_0 usually peaks after the stressed syllable (Xu and Xu, in press), as can be seen in Fig. 7d. Xu and Xu (in press) attribute the earlier peak under focus to a [fall] pitch target. Furthermore, post-focus pitch range suppression has been found to start from the first unstressed syllable in a trochaic word in English rather than from the first post-focus word (Xu and Xu, in press). If the language has neither pitch accent nor morphemic lexical stress, focus may modify not only the pitch ranges and shapes of pitch targets, but also the location of the prominent pitch target within the word. In some languages, e.g., French, the initial syllable of a multi-syllabic word is given a high pitch under focus (Carton et al., 1976). Note that all these modifications tend to favor a sharp F_0 decline as early as possible within the focused item.

The importance of the asymmetrical pitch range modification by focus has also been demonstrated in various perception studies (Botinis et al., 2000; Hasegawa and Hata, 1992; Krahmer and Swerts, 2001; Mixdorff, 2004; Rump and Collier, 1996; Xu et al., 2004b). Furthermore, Xu et al. (2004b) found that subjects could correctly *imitate* the global focus profile of a sentence even if they could only hear the post-focus portion of an utterance when the focused word was replaced by noise. Xu et al. (2004a,b) therefore argued that a narrow focus is melodically encoded through a tri-zone pitch range control: the pitch range of the focused region is expanded; that of the post-focus region compressed; and that of the pre-focused region left largely neutral. According to this understanding, pitch range variations in all three regions are components of the encoding scheme of focus, or *intrinsic properties* of focus. Note that this view is different from that of the AM theory, in which the focus-related pitch range variations discussed in this section are attributed to separate sources: on-focus pitch range expansion to a *nuclear accent*, and post-focus pitch range suppression to a *phrase accent* (Grice et al., 2000; Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988). The notion of phrase accent is particularly complicated, because, in order to account for the entire temporal span of post-focus pitch range suppression, the phrase accent is required to have both a primary association to the right edge of the post-focus region and a secondary association to the right edge of the focused word (Grice et al., 2000; Pierrehumbert and Beckman, 1988).

4.2.2. Sentence type: Statement versus question

The distinction between statement and question, yes–no question in particular, has always been considered a core function of intonation. And yet the exact formation of question intonation is still very much an open issue. Aside from the matter of whether yes-no questions always have a rising intonation, which I will briefly discuss later, even when a question sounds clearly different from a statement, its exact acoustic manifestation and especially its underlying representations are in much dispute. The most obvious F_0 pattern associated with a yes–no question is the

final rise, which has been attributed to a high boundary tone H% in the AM theory (Pierrehumbert, 1980; Ladd, 1996). There is evidence, however, that question intonation involves not only local F_0 variations, but also more global patterns. Thorsen (1980) conducted a gating experiment in which Danish subjects heard both statements and yes–no questions with increasingly longer initial or final portions of the sentence removed. She found that when the initial portions of the sentence were removed, listeners were always able to distinguish the two sentence types; when the final portions were removed, listeners continued to hear the question/statement distinction until only the first stress group was left. Eady and Cooper (1986) also found, through acoustic analysis, that not only is the F_0 of the final syllable raised in English yes–no questions, but also is that of all prominent words on and after focus. Myers (1999a) reported that in Chichewa, a Bantu language (spoken in Malawi) with three level tones, in addition to a final rise in a yes–no question, the F_0 of non-final H tones in the sentence are also raised, and the amount of raising is greater the closer the tone is to the end of the sentence.

The findings of Thorsen (1980) and Eady and Cooper (1986) demonstrated that in both Danish and English, the global F_0 raising is conditioned by focus, such that the significant F_0 increase occurs only from the focused element onward. A recent study by Liu and Xu (2004) found similar focus-conditioned F_0 raising in Mandarin. Fig. 8a and b show mean F_0 contours of Mandarin yes–no question versus statement in sentences consisting of only H tones. In Fig. 8a focus is either on the sentence initial word, or there is no narrow focus (neutral focus). In Fig. 8b, focus is either sentence medial or sentence final. It is apparent from Fig. 8 that the divergence between statement and question starts from the focused word. Neither does the divergence always start from the beginning of the sentence, nor does it always occur only in the final syllable or word. F_0 of sentences with other tones showed comparable patterns, although the details are varied depending on the properties of the lexical tones (Liu and Xu, 2004).

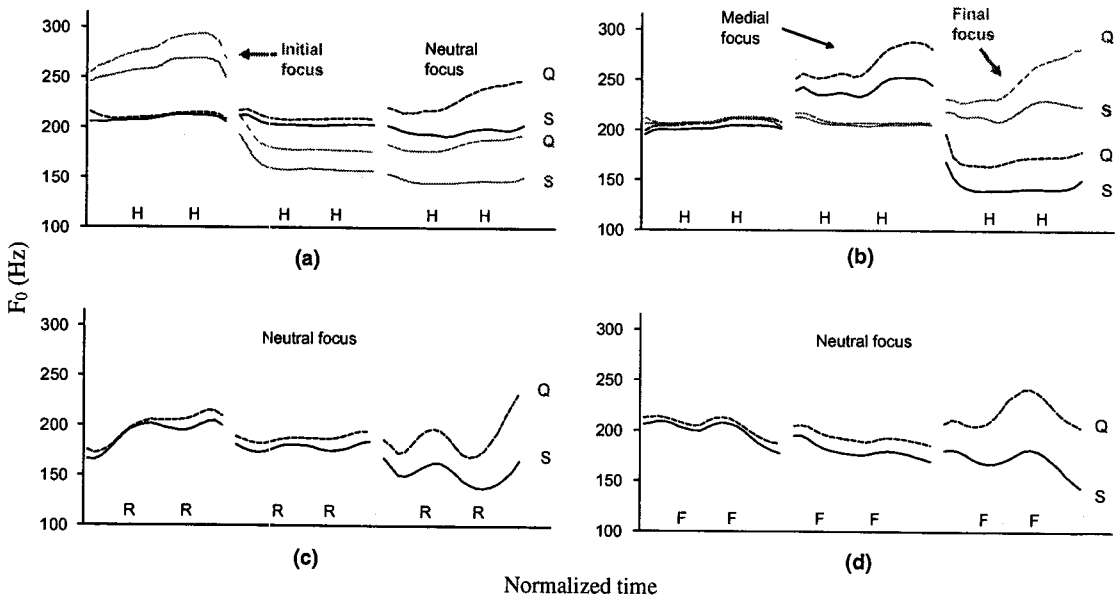

Fig. 8. Mean F_0 contours (40 repetitions by eight subjects) of Mandarin question versus statement. Sentences in (a) and (b) consist of only H tone, but differ in focus conditions. In (a) either focus is on the sentence initial word, or there is no narrow focus. In (b) Focus is either sentence medial or sentence final. Sentences in (c) and (d) consist of only R tone or F tone. Data from Liu and Xu (2004).

The focus-conditioned F_0 raising in yes–no questions indicates that the final rise, though often the most obvious characteristic, does not constitute the entire encoding scheme of question intonation. Myers (2004) attributes the immediate pre-boundary raising to coarticulatory assimilation to the boundary tone and the longer-range global raising to perceptual assimilation. The premise of this hypothesis is that only the boundary tone is the direct melodic correlate of questions, while the other variations are linked to questions only through boundary tone. As has been seen in Fig. 1 (and will be seen again in Fig. 9), a truly local pitch target is not articulatorily anticipated during the preceding syllables. It is thus unlikely that the pitch raising before the final word in questions is due to coarticulation. Rather, the F_0 patterns found in tone languages demonstrate that question intonation as an encoding scheme is independent of lexical tone, as can be seen in Fig. 8c and d in which are overlaid F_0 contours of Mandarin statements and yes–no questions consisting of either all R tone (c) or all F tone (d) (Liu and Xu, 2004). While the overall heights of the two F_0 tracks in each plot become increasingly different over time, the local contours related to lexical tones remain largely the same. The encoding scheme for sentence type thus seems to consist mainly of pitch range adjustments (or phrase curve adjustment suggested by Yuan et al., 2002) rather than the assignment of a single tonal element at the end of the sentence. Such pitch range adjustment is described as a linear function by Thorsen (1980). But data from Liu and Xu (2004) suggest that the function may be nonlinear, with greater increase toward the end of the sentence. The nonlinearity may explain why the final raising is typically much more prominent than in earlier regions. Note that linear or nonlinear, the pitch range adjustments are, articulatorily, part of the execution of the question function itself rather than anticipation of a single pitch target associated with only the right edge of the sentence.

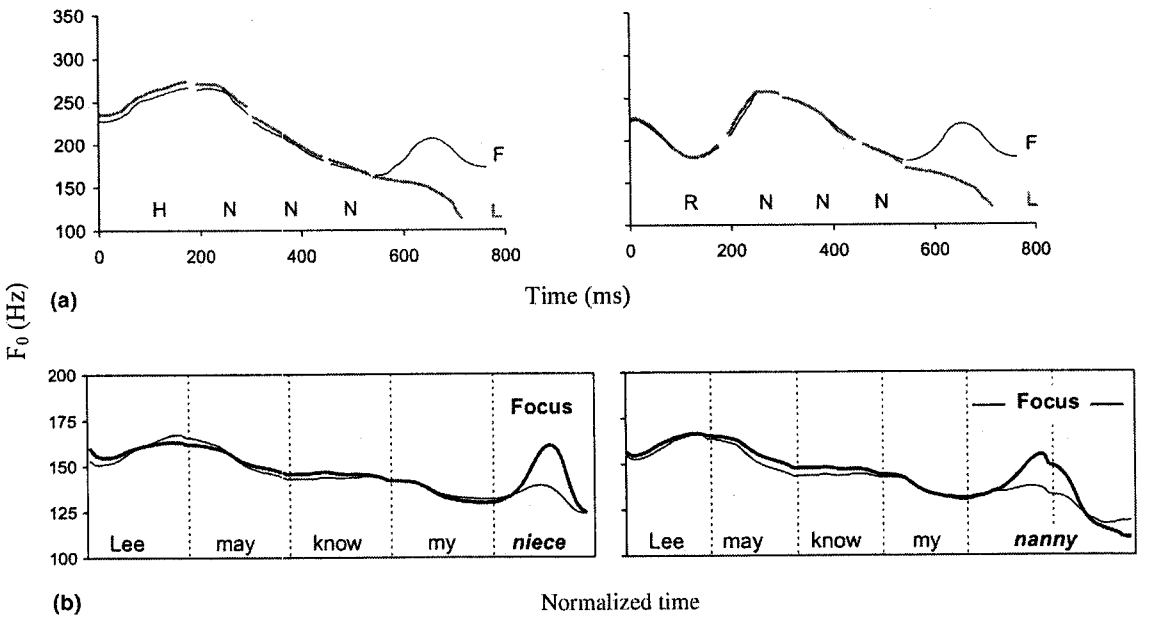

Fig. 9. (a) Mandarin sentences containing three consecutive neutral tone (N) syllables. They are preceded by H (left graph) or R (right graph) and followed by F or L (in both graphs) (data from Chen and Xu, forthcoming). (b) Mean F_0 curves of the American English sentence "Lee may know my niece" (left) and "Lee may know my nanny" (right) (adapted from Xu and Xu, in press). The sentences either have no narrow focus (thin line) or a narrow focus on the last word.

The focus-controlled, nonlinear pitch range modification may not be the sole component of the encoding scheme for question intonation. In English, for example, there is evidence that speakers produce questions not only by raising F_0 at the end of the sentence, but also by using a low-rise pitch pattern for the focused word (Eady and Cooper, 1986). Thus the encoding scheme of questions in English may involve, in addition to the focus-conditioned nonlinear pitch range raising, [low] or [rise] pitch targets assigned to the stressed syllables in the sentence. In Greek, Hungarian, Romanian and Neapolitan Italian, as reported by Grice et al. (2000) and D'Imperio (2001, 2002), F_0 always drops rather than rises at the end of a yes–no question, but the alignment of the final F_0 peak may differ between question and statement. Furthermore, there is evidence that at least in Neapolitan Italian there are also pitch range differences related to sentence type (D'Imperio, 2001). It is thus possible that the encoding schemes for sentence type in the other languages in this group also include similar nonlinear pitch range manipulation. Further investigation into these languages in this regard is needed.

Similar to nonlinear F_0 raising in questions, there is evidence that the gradual F_0 lowering in statements is also nonlinear, both in regard to the effects often referred to as downstep and declination ('t Hart et al., 1990; Ladd, 1984), and in regard to the "final lowering" found by Liberman and Pierrehumbert (1984). Also similar to the case of question intonation, both the more gradual lowering toward the end of the sentence and the much greater lowering at the end of the sentence should belong to a single encoding scheme that contrasts statements with questions.

Finally, as suggested by Bolinger (1989) and confirmed by Shattuck-Hufnagel and Turk (1996), only about half of the sentences with question syntax are said with a "rising intonation" during conversation. At the same time, many sentences without question syntax are said with question-like rising intonation. It is possible, as Bolinger (1989) has argued, it is mostly pragmatic meanings rather than syntactic structures that determine the employment of question intonation. Furthermore, it is possible that additional pitch range manipulations can further differentiate different kinds of meanings related to questions: incredulity, ignorance, inquiry, politeness, uncertainty, confirmation, demand, etc. It is still an open question as to how different languages encode these shades of meanings differently in question intonation.

4.2.3. New topic/turn taking

The previous section shows that part of the overall F_0 decline of a sentence, which has been found in many studies (Cohen et al., 1982; Cohen and 't Hart, 1965; Cooper and Sorenson, 1981; Pike, 1945; Ladd, 1984; Maeda, 1976; Shih, 2000), is attributable to the encoding scheme of contrasting statements with questions (and possibly also with continuation). But this contrast may not explain the full amount of declination, especially considering that the lowering in a statement is nonlinearly increasing toward the end of the sentence, as just discussed. This *increasing* decline is different from the generally reported "declination line," which is more nonlinear near the beginning of the sentence, with faster F_0 descent followed by slower descent, as is modeled by the exponential downstep function proposed by Pierrehumbert (1981) for English, and the exponential declination function proposed by Shih (2000) for Mandarin. As found by several studies (Lehiste, 1975; Nakajima and Allen, 1993; Umeda, 1982), the extreme cases of very high initial F_0 and very fast descent afterwards typically occur in the first sentence of a paragraph. Umeda (1982) suggests that such high initial F_0 is used as a beginning signal for a new topic. Nakajima and Allen (1993) reported data on high F_0 values related to topic shift. Further evidence comes from investigation of F_0 reset between adjacent sentences as a discourse function (e.g. Swerts, 1997). More recent research findings by Alku et al. (2002) and by Brungart et al. (2002) indicate a close link between high F_0 and high intensity. This seems to suggest that F_0 raising at the start of a new topic or conversational turn is for the sake of increasing amplitude in order to draw listeners' attention. Regardless of its precise nature, the initial F_0 raising, which is apparently also nonlinear, seems to be belong to an encoding scheme for new topic/turn taking,

thus is independent of the encoding scheme for sentence type as discussed in Section 4.2.2.

4.2.4. The effects of pitch range manipulation combined

With pitch range adjustment as their major means of coding, the three proposed encoding schemes just described can be effectively manifested in parallel with encoding schemes that mainly involve local pitch targets as discussed in Section 4.1. Also importantly, the pitch range adjustments by these encoding schemes, when combined with each other and with the effects of local pitch targets, can generate robust variations in the overall F_0 profile of an utterance. Such combined effects may account for a large portion, if not all, of declination. As we have seen so far, several functions contribute to the overall downtrend in an utterance, each in its own unique way. Downstep is brought by any pitch target with a low component, which raises the F_0 of the preceding syllable and lowers that of the following syllable (Gandour et al., 1994; Laniran and Clements, 2003; Xu, 1999). Focus introduces an asymmetrical pitch range adjustment that suppresses post-focus F_0 much more than pre-focus F_0 (Cooper et al., 1985; Xu, 1999). New topic/turn taking raises sentence initial F_0 (Lehiste, 1975; Nakajima and Allen, 1993; Swerts, 1997; Umeda, 1982). And, statement intonation lowers F_0 toward the end of the sentence (Liberman and Pierrehumbert, 1984; Liu and Xu, 2004). Note that based on this understanding, depending on which of these encoding schemes are in effect, the steepness of the declination line can vary significantly across sentences. This may explain the vastly different declination rates one may find in the literature.

4.3. Encoding schemes involving strength

To implement pitch targets, physical effort needs to be exerted. Such effort involves muscle forces whose strength determines how quickly a pitch target, together with its pitch range adjustment, is approached during articulation. Other things being equal, therefore, greater strength would enable a pitch target to be approached sooner than weaker strength would. Though probably not intuitively as obvious as pitch target and pitch range, strength has potential impact on surface F_0 contours, as has been illustrated in Fig. 5. As a directly controllable parameter with predicable acoustic consequence, strength is thus a likely encoding element for speech melody. The functional use of strength has not been studied as extensively as local F_0 shapes and overall pitch ranges. Nevertheless, there is already increasing evidence for it. Interestingly, the effectiveness of strength is more easily seen when it is weaker rather than when it is stronger. Such is the case with the Mandarin neutral tone, English unstressed syllables, and rhythmic structure in Mandarin.

4.3.1. Strength as lexical information

For various languages, certain syllables are considered to have no pitch units of their own, and their F_0 is assumed to come from interpolation between adjacent tones. In Mandarin, the neutral tone, which is a weak tone typically much shorter than the four full lexical tones (Lin, 1985; Lin and Yan, 1980), is often considered to be toneless or targetless (as reviewed by Yip, 2002). For English, vocalic segments between the accents are considered to be targetless and their F_0 assumed to be derived from interpolation. For some African tone languages, certain tones, e.g., L in Chichewa, are assumed to be unspecified for tonal values because their F_0 varies with surrounding tones and they do not participate in various phonological processes (Myers, 1999b). Several recent studies, however, show evidence that the F_0 of weak syllables cannot have come from interpolation between the preceding and following syllables. Fig. 9a shows Mandarin sentences containing three consecutive neutral tone (N) syllables (Chen and Xu, forthcoming). They are preceded by H (left graph) or R (right graph) and followed by F or L (in both graphs). We can see that the full tone of the last syllable (F or L) has no influence on the preceding neutral tones. Fig. 9b shows mean F_0 curves of the American English sentence "Lee may know my niece" (left) and "Lee may know my nanny" (right) (Xu and Xu, in press). The sentences either have a narrow focus on the last word (thick curves) or no narrow focus (thin curves). Just as in the Mandarin case, the extensive F_0 difference in the last word

due to focus has no effect on the F_0 of the preceding weak syllable "my." Meanwhile, as found by Chen and Xu (forthcoming), in Mandarin, the F_0 of a full-tone syllable is dependent on that of the preceding neutral-tone syllable. Ladd and Schepman (2003) also showed that when the number of weak syllables between two accented syllables changed in English, it is the F_0 of the following rather than the preceding accented syllable that is affected. Such asymmetry in contextual F_0 variation makes it unlikely that the F_0 of a weak syllable in Mandarin or English is derived from interpolation between the surrounding strong syllables, for the interpolation line would have to be drawn toward an endpoint whose F_0 is dependent on F_0 in the past—a process that is obviously circular.

Given that the F_0 of the weak syllable is extensively influenced by the *preceding* syllable, the mechanism of F_0 production in the weak syllable is probably not mechanistically different from that of the strong syllables as shown in Fig. 1 and summarized in the TA model (Fig. 4). But there is a critical difference nonetheless, mainly in terms of the speed of target approximation, which determines how fully a target is approached within the allocated time interval, and syllable duration, which determines how much time is allocated to the target approximation process. The weak syllables also differ from the strong ones, of course, in terms of their pitch target values. For that, there is evidence that the target value of weak syllables in Mandarin (Chen and Xu, forthcoming) and English (Xu and Xu, in press) is half way between the maximum and minimum F_0 of the upcoming full tones or stressed syllables, indicating a mid-level static target in both languages. Due to the weak strength, the approximation of the static target associated with a weak syllable can be very slow. As found by Chen and Xu (forthcoming), the F_0 trajectories of consecutive neutral tone syllable in Mandarin are not fully converged even by the end of the third neutral-tone syllable after different full tones. This is in sharp contrast to those of the full tones, which are virtually converged by the end of the first syllable after different tones, as can be seen in Fig. 1. The slow target approximation thus gives rise to the interpolation-like F_0 transitions across the weak syllables.

As to how weak the strength can be, it has been suggested that sometimes the level of articulatory effort is zero, i.e., with no active muscle forces present. The assumption is that when muscle forces are absent, F_0 would automatically relax toward a rest position due to the elasticity of the vocal folds. This has been suggested as one of the possible ways how the neutral tone in Mandarin is phonetically implemented (cf. discussion in (Yip, 2002)). Relaxation toward a rest position is a core component of the command-response model (Fujisaki, 1983), in which F_0 automatically returns to a base line when there are no active muscle commands. Note that such automatic return can happen only under a limited number of scenarios, (1) if the vocal folds are stretched beyond their resting lengths during phonation, and (2) when an antagonist muscle is in constant contraction, thus acting like a stretched rubber band. Scenario 1 is unlikely according to Hollien (1960) and Hollien and Moore (1960), who have shown that the vocal folds are typically shorter during phonation than during rest. Scenario 2 requires that at least one of the F_0 controlling laryngeal muscles, e.g., the vocalis, be in constant isometric contraction during phonation, which is also unlikely, although future evidence could show otherwise. Note that the lack of automatic return to a neutral state does not mean the absence of a neutral target. It is highly plausible that the [mid] target reported in (Chen and Xu, forthcoming) for Mandarin and in (Xu and Xu, in press) for English is related to a neutral pitch register near the level of the habitual pitch (Zemlin, 1988). However, this pitch is likely to be a real target and hence is implemented with active muscle force, although the strength of the force may be very weak at times. Note also that following this understanding, while the presence of a target is obligatory, the type of target and the amount of strength applied to the target implementation is by choice, and hence information bearing.

The target-plus-weak-strength hypothesis for the weak syllables has implications for the issue of whether pitch accents in languages like English have only word-size whole contour specifications, as mentioned in Section 4.1.2. While it is likely that at a certain abstract level an accent functions

as a whole entity, at the level of articulatory implementation, the accent shape has to be realized though accent-internal pitch target assignment. That is, each and every weak syllable would be assigned a default [mid] target as well as weak articulatory strength. This would make it possible for the TA process to generate F_0 contours during the weak syllables. In addition to the theoretical considerations presented here and acoustic data reported elsewhere (Chen and Xu, forthcoming; Xu and Xu, in press), there has also been preliminary support from synthetic simulations. Sun (2002) compared three algorithms for synthesizing English intonation based on speech in a news-reading corpus. He showed that the two algorithms that use the syllable as the unit of F_0 contour parameterization—the Three-target (per syllable) model (Black and Hunt, 1996) and a simplified TA model, both performed better than an event-based algorithm that uses whole pitch accent as the unit of F_0 contour parameterization (Taylor, 2000).

4.3.2. Strength as grouping, chunking, demarcation information

In addition to encoding lexical information, strength may also serve to group syllables into chunks, and to group smaller chunks into larger ones. Such grouping is closely related to a hierarchical prosodic structure whose function is not yet quite clear. It has been argued that prosodic structures are governed by syntactic structures (Selkirk, 2002). But there is also evidence that the link between the two is not always close. Shih (1986) and Speer et al. (1989) demonstrate that the occurrence of the L-tone sandhi in Mandarin (as discussed in Section 4.1.1), is more directly governed by prosodic rather than by syntactic structure of the utterance. What is determined by prosodic structure is the grouping relation among the linear sequence of syllables and words. Shih and Sproat (1992) and Shih (1993) further demonstrate that a syllable's ability to resist tonal coarticulation from adjacent tones is related to its prosodic strength. Tones on prosodically weak, but not necessarily lexically weak, syllables tend to have less extreme tone shapes, or, in other words, greater undershoot.

Xu and Wang (2005) investigated how prosodic grouping is manifested through strength by comparing, among other things, degrees of undershoot of tonal targets as related to the number of syllables in a group. Fig. 10 displays mean F_0 contours of R and F sequences with varying numbers of syllables said in the middle of a carrier. The all-R and all-F sequences impose great pressure on tone production (Xu, 1994) and hence would best reveal the strength differences among the syllables. As can be seen in Fig. 10, as the number of syllables increases, the magnitude of F_0 excursion decreases, which appears to be related to the insufficient increase in syllable duration as the group size becomes larger. Meanwhile, the magnitude reduction differs in various locations. But it seems that in each case it is the first and last syllables in the group that are reduced the least. There thus appears to be an "edge marking" effect. These patterns are similar to the prosodic strength patterns obtained by Kochanski et al. (2003) for Mandarin, although the details are different. Whereas they reported consistent stronger strength in the first syllable than in the second syllable, in our data, as shown in Fig. 10, the second syllable is weaker than the first only in tri-syllabic and quadra-syllabic groups. Our data are still preliminary and it awaits further analysis to confirm what the F_0 plots appear to show.

The three cases just discussed, namely, Mandarin neutral tone, English weak stress, and prosodic grouping in Mandarin, demonstrate the potential effectiveness of strength as an encoding element. But this aspect of speech melody is just beginning to be investigated. Much further research is needed to determine its specific usages and their effectiveness in encoding certain functions in various languages.

4.4. Encoding schemes involving syllable duration

Intuitively, duration is not part of speech melody. It nevertheless is an intricate aspect of speech prosody, for much of the prosodic structure and variation may have to do with duration. Such durational variation should affect F_0 contours, according to the TA model. That is, other things being equal, shorter duration makes a pitch target

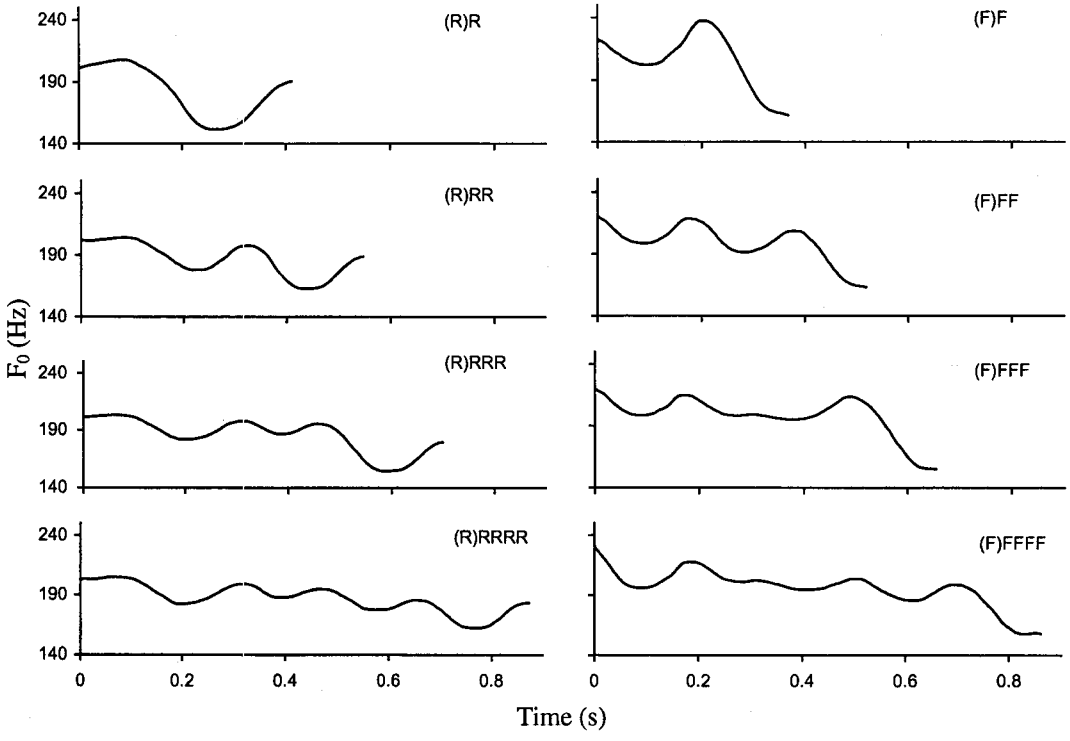

Fig. 10. Mean F_0 contours of consecutive R and F sequences with varying numbers of syllables said in middle of a carrier (Xu and Wang, 2005).

less likely to be reached by the end of the syllable, as we have seen in Fig. 5. Thus shortened duration may be used in conjunction with reduced articulatory strength for encoding certain functions. The grouping function discussed in Section 4.3.2, for example, is likely at least partially implemented with manipulation of syllable duration.

The importance of syllable duration can be also seen in the distributional patterns of contour tones among tone languages. Recent typological studies have found a close relation between the occurrence of dynamic tones and average vowel length (hence syllable length) in a language (Gordon, 1999; Zhang, 2001). Xu (2004a) argues that this may have to do with the fact that when syllable duration becomes very short, it is articulatorily impossible to produce a dynamic tone in many tonal contexts. This is because the implementation of a dynamic tone like R or F requires two F_0 movements within a single syllable when it is preceded by a tone that generates an offset F_0 very different from the initial pitch of the R or F. When syllable duration is about 150 ms or shorter, which frequently happens in languages with "weak syllable weight," e.g., Shanghai Chinese (Duanmu, 1994), there is virtually no way for an average speaker to produce two movements in a syllable according to Eqs. (3) and (4) presented in Section 2.1.

5. Modeling

5.1. The parallel encoding and target approximation (PENTA) model

The discussion thus far has demonstrated the following understanding of speech melody. First, maximum speed of pitch change and synchronization of pitch movements with syllables impose strong articulatory constraints on the generation of surface F_0 contours. Second, the basic articulatory mechanism of F_0 generation is likely to be

syllable-synchronized sequential target approximation. Third, the assignment of pitch targets by various functions is language specific and often non-unique. Fourth, the melodic components of speech are likely defined and organized in terms of function rather than form. Finally, multiple communicative functions can be transmitted in parallel, likely through distinctive encoding schemes, each specifying the values of one or more of the four melodic primitives—pitch target, pitch range, strength and duration.

This new level of understanding allows us to expand the target approximation (TA) model (Xu et al., 1999; Xu and Wang, 2001) into a more comprehensive model of speech melody. The parallel encoding and target approximation (PENTA) model was thus proposed (Xu, 2004b). A diagram of the PENTA model is shown in Fig. 11. The stacked boxes on the far left represent individual communicative functions. These functions control F_0 through distinctive *encoding schemes* (the second stack of boxes) that specify the values of the melodic primitives (middle block), which include *local pitch target*, *pitch range*, *articulatory strength* and *duration*. The values of the melodic primitives as stipulated by different encoding schemes can be specified both symbolically and numerically. Table 1 shows possible symbolic values of the melodic primitives.

As can be seen in Table 1, a local pitch target can be either static, e.g., [high], [low] or [mid], or dynamic, e.g., [rise] or [fall]. When a target is static, its relative pitch height is the only intended goal. Pitch range determines the frequency scope across which local pitch targets are implemented. It has two kinds of specifications: height and span. Height specifies the relative height of the pitch range, e.g., high, low or mid. Span specifies the width of the pitch range, e.g., wide or narrow. Articulatory strength specifies the speed at which a local pitch target is approached. When the strength is **strong**, the target is approached faster than when it is **weak**. Duration specifies the length of the time interval (typically that of syllable) during which a target is approximated. Note that although symbolic values are suggested for all the melodic primitives in Table 1, their specifications do not have to be always categorical. The categoricalness of the melodic specifications depends on the function being encoded. It is conceivable, for example, that the pitch range specification by new topic/turn taking and strength specification by grouping are gradient, as the functions themselves are conceivably rather gradient.

Table 1
Possible symbolic values of the melodic primitives: *pitch target*, *pitch range*, *strength* and *duration*, which may be notationally distinguished from one another by [ ], underline, **boldface** and *italic*, respectively

Local target		[high], [low], [rise], [fall], [mid]
Pitch range	Height	high, low, mid
	Span	wide, narrow, normal
Strength		**strong, weak, normal**
Duration		*long, short, normal*

An example of encoding schemes for lexical tone and focus is presented in Fig. 12, which shows

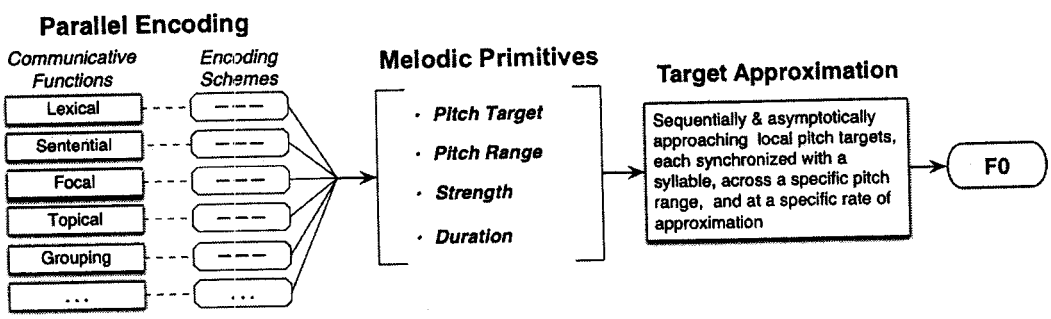

Fig. 11. A schematic sketch of the PENTA model. See text for explanations. The unnamed block at the bottom left indicates communicative functions yet to be identified.

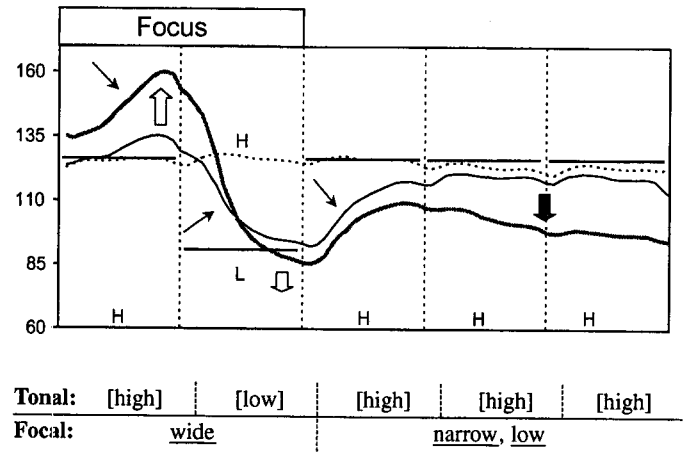

Fig. 12. Mean F_0 of 20 repetitions of HLHHH sequence by four Mandarin male speakers. Thick curve: focus on "Māomǐ."; thin curve, no focus. Vertical grids indicate locations of nasal murmur onset (data from Xu, 1999). Short horizontal lines indicate hypothetical pitch targets [high] and [low]. Thin arrows point to F_0 variations due to inertia. Unfilled block arrows indicate on-focus pitch range expansion. Filled block arrow indicates post-focus pitch range narrowing and lowering.

an actual case of simultaneous transmission of lexical tones and focus in Mandarin, with decomposition of the functional components in the framework of the PENTA model. Displayed in the graphic part of Fig. 12 are the mean F_0 curves of the Mandarin sentence "Māomǐ mō māomī" [Cat-rice strokes kitty] (tone sequence: HLHHH) said with and without initial focus (thick-solid/-thin curves), together with the average F_0 curve of an all-H sentence without narrow focus as reference (dotted line). The lexical tones are associated, via tonal rules specific to Mandarin, with the local targets [high] and [low], respectively, as shown in the *Tonal* tier below the F_0 plot. When the first (disyllabic) word of the sentence is focused, the encoding scheme of focus assigns a wide pitch range to the focused syllables, and a narrow + low pitch range to the post-focus words, as shown in the *Focal* tier below the F_0 plot.

The symbolic representations of the local targets and pitch ranges also correspond to specific numerical values. The specific height and shape of the local pitch targets corresponding to the lexical tones are depicted by the short horizontal lines in the figure, indicating that they can be represented numerically by simple linear functions. The pitch range adjustments by focus are indicated by the block arrows. The two unfilled block arrows on the left indicate a wide pitch range as compared to a normal range not explicitly depicted in the graph. The filled block arrow on the right indicates a narrow + low pitch range (though narrow is not obvious because all the local targets in these post-focus words are static).

In the PENTA model, the melodic primitives are, *at the same time*, control parameters for the TA model that simulates articulatory implementation of the pitch targets. Through the TA model, the control parameters corresponding to the melodic primitives are turned into continuous F_0 contours through asymptotic approximation of pitch targets, which are synchronized with their associated syllables. Under the constraint of maximum speed of pitch change (Xu and Sun, 2002), such asymptotic approximation often results in a long transition in the early part of a syllable, as we have seen in Fig. 1. The asymptotic transitions are also apparent in syllables 1–3 in Fig. 12, as indicated by the line arrows. In addition, the target approximation process also produces the peaks in syllables 1 and 3 (the latter only when with initial focus), and the valley in syllable 3. Thus through target assignment by the encoding schemes and target approximation by articulatory implementation, the functional components of intonation are eventually turned into continuous F_0 contours.

Initial effort to quantify the target approximation part of the PENTA model was made in (Xu

et al., 1999). The general notion of syllable-synchronized sequential target approximation has been tested in an attempt to synthesize English intonation through target parameter estimation based on real speech and re-synthesis using the target parameters (Sun, 2002). The results were encouraging, as has been discussed in Section 4.3.1. However, the quantization of the TA model in (Xu et al., 1999) is not fully adequate in that it is not able to capture the effect of final velocity of the preceding syllable on the F_0 of the following syllable. Such influence has since been found to be quite substantial in both Mandarin and English when the strength of the second syllable is weak (Chen and Xu, forthcoming; Xu and Xu, in press). An improved quantitative implementation of the PENTA mode should therefore fully capture this aspect of the articulatory mechanism.

5.2. Comparing PENTA with other models

To compare PENTA with other models, it may help to first divide the models into two broad groups. The first may be referred to as the *direct F_0 specification models*, and the second *articulatorily oriented models*. Direct F_0 specification models try to capture surface forms directly, and to map them onto either communicative functions or phonological categories. Other than F_0 itself, parameters of these models usually bear no articulatory connotations or assumptions. The basic, often implicit, assumption is that all the major details of F_0 contours are deliberately produced to satisfy perception or linguistic specifications of the language in question. The main examples of this group of models are the IPO model ('t Hart et al., 1990), the tilt model (Taylor, 2000), the quadratic spline model (Hirst and Espesser, 1993), and the Pierrehumbert model (Pierrehumbert, 1981). As demonstrated in Section 1.1 and 2, many of the details in pitch trajectories are directly attributable to properties of the articulatory system that generates F_0 contours. As such, they can be neither attributed entirely to perception or linguistic specifications, nor modeled adequately without taking the basic articulatory mechanisms into consideration.

Articulatorily oriented models are those that try to simulate the articulatory process underlying the generation of surface F_0 contours. In these models, parameters often have some articulatory connotations. And it is generally assumed in these models that perception and linguistic specifications play only a partial role in determining the F_0 trajectories. Among these are the Öhman model (Öhman, 1967), the command-response model (Fujisaki, 1983, 2003) and the soft template model (or Stem-ML) model (Kochanski and Shih, 2003). The PENTA model described in Section 4.1 belongs to this group.

The Öhman model and the command-response model by Fujisaki are somewhat similar in nature. In the Fujisaki model (Fujisaki, 1983, 2003), F_0 contours arise from responses of a second order linear system to two types of *muscular* commands, accent commands and phrase commands, which have idealized stepwise waveforms and idealized impulse waveforms, respectively. The system responds to these commands by generating critically damped oscillations of F_0, which rises or falls exponentially in the direction of the commands and then *returns to the baseline* after the termination of the commands. The soft template model (Kochanski and Shih, 2003) describes F_0 contours as resulting from implementing underlying tonal templates with different amounts of muscle forces under the physical constraint of *smoothness*. The smoothness constraint makes the connection between adjacent templates seamless, and the varying muscle forces determine the degree to which the shape of each template is preserved in the surface F_0. An important characteristic of this model is that the smoothness constraint is *bidirectional*, i.e., it exerts both carryover and anticipatory assimilatory influences.

Though sharing the common assumption that articulatory mechanisms need to be incorporated into the modeling process, the PENTA model differs from the other articulatorily oriented models in a number of nontrivial ways. First, it assumes *syllable-synchronized sequential target approximation* as the basic mechanism of F_0 contour generation. The synchrony assumption is absent in the command-response (Fujisaki, 1983, 2003), the Öhman model, and is only loosely held in the soft template model, in which the precise alignment of tonal templates with the syllable is flexible.

The sequential assumption, which allows only left-to-right assimilatory influences, differs from the assumption of the soft template model which assumes *bidirectional* smoothing as the major shaping mechanism of surface F_0 contours. Second, for reasons discussed in Section 4.3.1, the PENTA model assumes no or negligible returning force in the direction of a "baseline" as assumed in the command-response model. Third, similar to the soft template model but different from the command-response models, the most basic information bearing components in the PENTA model are ideal pitch patterns rather than muscle commands. The rationale behind this assumption is that, as seen in Fig. 1 and the discussion in Section 1.1, while the ideal pitch pattern may remain constant, the actual F_0 trajectories approaching it, and hence the muscle commands generating the trajectories may differ significantly in various tonal and intonational contexts. A further support for this assumption comes from the recent finding that on-line auditory feedback plays a role in the production of tone sequences in Mandarin (Xu et al., 2004a), which implies that muscle forces are constantly adjusted during articulation and hence cannot be stored as invariant patterns corresponding directly to phonological tones or even phonetic tonal targets.

At the local level, therefore, the PENTA model is built on rather stringent assumptions. These assumptions may turn out to be wrong, but their explicitness should make them easily testable.

With regard to the communicative aspects of speech melody, both the Öhman model (Öhman, 1967) and the command-response model (Fujisaki, 1983, 2003) have considered only two possible types of neural commands: local accents and phrasal or global intonation commands. The soft template model has considered tonal templates, prosodic strength and phrase curve as three controllable parameters. In the PENTA model, the parallel transmission of *multiple* communicative functions is part of the basic assumption and such transmission is assumed to be enabled through distinct and sometimes complex encoding schemes that are either language specific or universal. Note that, however, the PENTA model by itself does not stipulate the properties of the encoding schemes. It only provides a mechanistic framework for the encoding schemes to be implementable. The detailed properties of the encoding schemes, whether universal or language specific, can be discovered only through empirical investigations in which potential contributors to surface F_0 contours are systematically controlled.

6. General summary and conclusion

There might be a natural tendency for those mainly interested in the linguistic and communicative aspects of speech to want to stay away from the details of speech production mechanisms, and for those mainly interested in the phonetic aspects of speech to want to avoid directly dealing with too many complex linguistic or communicative functions. The discussion in the present paper has demonstrated that the articulatory mechanisms of F_0 production and the rich communicative functions conveyed through F_0 are so intricately related that the understanding of even a single aspect of speech melody requires knowledge about how the system works as a whole. The present paper is thus an attempt to consider articulatory mechanisms and communicative functions together. It is recognized that speech melody is used to convey a variety of communicative functions. But such conveyance is done through an articulatory system with various properties. The nature of these properties are such that the function transmission has to be done by controlling a limited number of articulatorily operable parameters, which I have referred to as the melodic primitives. Four such primitives have been considered: local pitch target, pitch range, strength and duration. The mapping of these primitives to communicative meanings, however, is not direct, but through a number of encoding schemes, each corresponding to a specific function. The encoding schemes are distinct from each other in their manner of controlling the melodic primitives. This allows multiple communicative functions to be conveyed in parallel through speech melody. The communicative functions are ultimately converted to continuous, detailed surface acoustic patterns through an articulatory process of *syllable-synchronized*

sequential target approximation, which takes as its control parameters the melodic primitives specified by the encoding schemes. This view of speech melody is summarized into a comprehensive model of tone and intonation, namely, the parallel encoding and target approximation (PENTA) model. Compared to the existing models of speech melody, the PENTA model is very stringent in terms of its articulatory assumptions, but at the same time it specifies mechanisms for multiple communicative functions to be transmitted in parallel.

Admittedly, many important tonal and intonational phenomena are not covered in the present paper. An important reason for this inadequacy, other than the apparent space limit, is the lack of systematically obtained data. For example, whether question intonation in certain Eastern European languages involves pitch range manipulation in addition to the local F_0 contours discussed in (Grice et al., 2000) cannot be known until F_0 height is compared between statement and question (although there is already some initial evidence for pitch range raising, e.g., D'Imperio, 2001 and Gussenhoven, 2002). For another example, we cannot know for certain how many of the reportedly tonal phenomena in many African tone languages are purely tonal or rather at least partially intonational until data have been collected in investigations that systematically control intonational variables as well as tonal variables.

Also, the present paper has made no attempt to link intonational meanings directly to the "biological codes" proposed by Ohala (1983, 1984) and elaborated by Gussenhoven (2002). Instead, the proposed encoding schemes corresponding to the communicative functions considered here are assumed to be largely codified through long-term use in each language, or in other words, "grammaticalized" in (Gussenhoven's, 2002) term. On the other hand, there may be communicative functions that are not as codified, notably those related to attitudes and emotions (e.g., Cowie and Cornelius, 2003; Mozziconacci, 2002; Scherer, 2003; Wichmann, 2000, among others). As has been argued, a better understanding of emotion and attitude is dependent on improved understanding of 'normative' use of prosody (Wichmann, 2002), and that emotion and attitude are best studied when pitch variation is represented in the theoretical framework of a model of intonation (Mozziconacci, 2002). It is hoped that the discussion in the present paper will help to establish such a theoretical framework.

Finally, the principles and mechanisms discussed in the present paper are probably relevant also to the segmental aspect of speech, especially in regard to the variability of consonants and vowels related to nonlexical factors. In fact, the main benefit of considering the melodic aspect of speech separately from the segmental aspect is to have a more focused discussion. While such narrow focus may be beneficial in the short run, in the long run, it is necessary to develop a unified theory encompassing multiple aspects of speech based on simultaneous consideration of both articulatory constraints and communicative functions.

Acknowledgments

I thank two anonymous reviewers for their critical comments and suggestions on an earlier version of this paper. I am also grateful for the comments and suggestions by Carlos Gussenhoven, Andrew Wallace, Jill House and Maolin Wang. This work is supported in part by NIH Grant R01DC03902 and NIH grant 1R01DC006243.

References

Alku, P., Vintturi, J., Vilkman, E., 2002. Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. Speech Commun. 38, 321–334.

Arvaniti, A., Ladd, D.R., Mennen, I., 1998. Stability of tonal alignment: The case of Greek prenuclear accents. J. Phonetics 36, 3–25.

Atterer, M., Ladd, D.R., 2004. On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German. J. Phonetics 32, 177–197.

Black, A., Hunt, A., 1996. Generating F0 contours from ToBI labels using linear regression. In: Proceedings of International Conference on Spoken Language Processing, Philadelphia.

Blevins, J., 1993. A tonal analysis of Lithuanian nominal accent. Language 69, 237–273.

Bolinger, D., 1964. Intonation: Around the edge of language. Harvard Educ. Rev. 34, 282–296.

Bolinger, D., 1972. Accent is predictable (if you're a mind reader). Language 48, 633–644.

Bolinger, D., 1989. Intonation and Its Uses—Melody in Grammar and Discourse. Stanford University Press, Stanford, California.

Botinis, A., Bannert, R., Tatham, M., 2000. Contrastive tonal analysis of focus perception in Greek and Swedish. In: Botinis, A. (Ed.), Intonation: Analysis, Modelling and Technology. Kluwer Academic Publishers, Boston, pp. 97–116.

Bruce, G., 1977. Swedish word accents in sentence perspective. Travaux de L'institute de Linguistique De Lund Xii. B. Malmberg and K. Hadding, Lund: Gleerup.

Brungart, D.S., Kordik, A.J., Das, K., Shawy, A.K., 2002. The effects of F0 manipulation on the perceived distance of speech. In: Proceedings of 7th International Conference on Spoken Language Processing, Denver, Colorado, pp. 1641–1644.

Carton, F., Hirst, D., Marchal, A., Séguinot, A., 1976. L'accent D'insistance—Emphatic Stress. Didier, Montréal, Paris, Bruxelles.

Chao, Y.R., 1932. A preliminary study of English intonation (with American variants) and its Chinese equivalents. In: Shiyusuo Jikan [A Collection by Shiyusuo]: Special issue—A Festschrift to honor Mr. Cai Yuanpei, pp. 105–156.

Chao, Y.R., 1968. A Grammar of Spoken Chinese. University of California Press, Berkeley, CA.

Chen, M.Y., 2000. Tone Sandhi: Patterns Across Chinese Dialects. Cambridge University Press, Cambridge, UK.

Chen, Y., Xu, Y., forthcoming. Production of weak elements in speech – evidence from F0 patterns of neutral tone in Standard Chinese.

Cohen, A., Collier, R., 't Hart, J., 1982. Declination: Construct or intrinsic feature of speech pitch. Phonetica 39, 254–273.

Cohen, A., 't Hart, J., 1965. Perceptual analysis of intonation patterns. In: Commins, D.E. (Ed.), Proceedings of the Fifth International Congress on Acoustics, Liège, A. 16.

Cooper, W.E., Sorenson, J.M., 1981. Fundamental Frequency in Sentence Production. Springer-Verlag, New York.

Cooper, W.E., Eady, S.J., Mueller, P.R., 1985. Acoustical aspects of contrastive stress in question–answer contexts. J. Acoust. Soc. Amer. 77, 2142–2156.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Commun. 40, 5–32.

Cruttenden, A., 1997. Intonation, second ed. Cambridge University Press.

D'Imperio, M., 2001. Focus and tonal structure in Neapolitan Italian. Speech Commun. 33, 339–356.

D'Imperio, M., 2002. Italian intonation: An overview and some questions. Probus 14, 37–69.

Duanmu, S., 1994. Syllabic weight and syllable durations: A correlation between phonology and phonetics. Phonology 11, 1–24.

Eady, S.J., Cooper, W.E., 1986. Speech intonation and focus location in matched statements and questions. J. Acoust. Soc. Amer. 80, 402–416.

Fairbanks, G., 1959. Voice and Articulation Drillbook. Harper & Row, New York.

Fry, D.B., 1958. Experiments in the perception of stress. Language Speech 1, 126–152.

Fujimura, O., 2000. The C/D model and prosodic control of articulatory behavior. Phonetica 57, 128–138.

Fujisaki, H., 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In: MacNeilage, P.F. (Ed.), The Production of Speech. Springer-Verlag, New York, pp. 39–55.

Fujisaki, H., 2003. Prosody, information, and modeling—with emphasis on tonal features of speech. Proc. Workshop Spoken Language Process., 5–14.

Gandour, J., Potisuk, S., Dechongkit, S., 1994. Tonal coarticulation in Thai. J. Phonetics 22, 477–492.

Gårding, E., 1982. Swedish prosody—Summary of a project. Phonetica 39, 288–301.

Gordon, M., 1999. Syllable Weight: Phonetics, Phonology, and Typology. Ph.D. dissertation, UCLA.

Grice, M., Ladd, D.R., Arvaniti, A., 2000. On the place of phrase accents in intonational phonology. Phonology 17, 143–185.

Gussenhoven, C., 2002. Intonation and interpretation: Phonetics and phonology. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France, pp. 47–57.

Gussenhoven, C., in press. Types of focus in English. In: Bring, D., Gordon, M., Lee, C. (Eds.), Topic and Focus: Intonation and Meaning. Theoretical and Crosslinguistic Perspectives. Kluwer, Dordrecht.

Hasegawa, Y., Hata, K., 1992. Fundamental frequency as an acoustic cue to accent perception. Language Speech 35, 87–98.

Hirst, D., Espesser, R., 1993. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix 15, 75–85.

Hollien, H., 1960. Vocal pitch variation related to changes in vocal fold length. J. Speech Hearing Res. 3, 150–156.

Hollien, H., Moore, G.P., 1960. Measurements of the vocal folds during changes in pitch. J. Speech Hearing Res. 3, 157–165.

Honorof, D.N., Whalen, D.H., in press. Perception of pitch location within a speaker's F0 range. Journal of the Acoustical Society of America.

Ishihara, S., 2002. Syntax–phonology interface of Wh-constructions in Japanese. In: Proceedings of TCP 2002, Tokyo, pp. 165–189.

Jiao, W., 2001. Zunhua fangyan kouyu zhong shangsheng de bianhua [Variations of the third tone in conversational speech of Zunhua dialect]. In: Proceedings of 5th National Conference on Modern Phonetics, Beijing, China, pp. 328–342.

Jin, S., 1996. An Acoustic Study of Sentence Stress in Mandarin Chinese. Ph.D. dissertation, The Ohio State University.

Kelso, J.A.S., 1984. Phase transitions and critical behavior in human bimanual coordination. Amer. J. Physiol.: Regulatory Integrative Comp. 246, R1000–R1004.

Kochanski, G., Shih, C., 2003. Prosody modeling with soft templates. Speech Commun. 39, 311–352.

Kochanski, G., Shih, C., Jing, H., 2003. Quantitative measurement of prosodic strength in Mandarin. Speech Commun. 41, 625–645.

Krahmer, E., Swerts, M., 2001. On the alleged existence of contrastive accents. Speech Commun. 34, 391–405.

Krakow, R.A., 1999. Physiological organization of syllables: A review. J. Phonetics 27, 23–54.

Ladd, D.R., 1984. Declination: A review and some hypotheses. Phonol. Yearbook 1, 53–74.

Ladd, D.R., 1996. Intonational Phonology. Cambridge University Press, Cambridge.

Ladd, D.R., Schepman, A., 2003. "Sagging transitions" between high pitch accents in English: Experimental evidence. J. Phonetics 31, 81–112.

Ladd, D.R., Faulkner, D., Faulkner, H., Schepman, A., 1999. Constant "segmental anchoring" of F0 movements under changes in speech rate. J. Acoust. Soc. Amer. 106, 1543–1554.

Ladd, D.R., Mennen, I., Schepman, A., 2000. Phonological conditioning of peak alignment in rising pitch accents in Dutch. J. Acoust. Soc. Amer. 107, 2685–2696.

Laniran, Y.O., Clements, G.N., 2003. Downstep and high raising: Interacting factors in Yoruba tone production. J. Phonetics 31, 203–250.

Lehiste, I., 1975. The phonetic structure of paragraphs. In: Cohen, A., Nooteboom, S.E.G. (Eds.), Structure and Process in Speech Perception. Springer-Verlag, New York, pp. 195–206.

Li, Y.J., Lee, T., 2002. Acoustical F0 analysis of continuous Cantonese speech. In: Proceedings of International Symposium on Chinese Spoken Language Processing 2002, Taipei, Taiwan, pp. 127–130.

Liberman, M., Pierrehumbert, J., 1984. Intonational invariance under changes in pitch range and length. In: Aronoff, M., Oehrle, R. (Eds.), Language Sound Structure. MIT Press, Cambridge Massachusetts, pp. 157–233.

Lin, M., Yan, J., 1980. Beijinghua qingsheng de shengxue xingzhi [Acoustic properties of Mandarin neutral tone]. Dialect (3), 166–178.

Lin, M., Yan, J., 1991. Tonal coarticulation patterns in quadrisyllabic words and phrases of Mandarin. In: Proceedings of the 12th International Congress of Phonetic Sciences, Aix-en-Provence, France, pp. 242–245.

Lin, M., Lin, L., Xia, G., Cao, Y., 1980. Putonghua erzici biandiao de shiyan yanjiu [An experimental study of tonal variation in disyllabic words in Standard Chinese]. Zhongguo Yuwen [Chinese Linguistics] (1), 74–79.

Lin, T., 1985. Preliminary experiments on the nature of Mandarin neutral tone [in Chinese]. In: Lin, T., Wang, L. (Eds.), Working Papers in Experimental Phonetics. Beijing University Press, Beijing, pp. 1–26.

Liu, F., Xu, Y., 2004. Asking questions with focus. J. Acoust. Soc. Amer. 115 (Pt. 2), 2397.

MacNeilage, P.F., 1998. The frame/content theory of evolution of speech production. Behav. Brain Sci. 21, 499–546.

Maeda, S., 1976. A Characterization of American English Intonation. MIT Press, Cambridge, MA.

Man, V.C.H., 2002. Focus effects on Cantonese tones: An acoustic study. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France, pp. 467–470.

Mixdorff, H., 2004. Quantitative tone and intonation modeling across languages. In: Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, pp. 137–142.

Mozziconacci, S., 2002. Prosody and emotions. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France, pp. 1–9.

Myers, S., 1999a. Downdrift and pitch range in chichewa intonation. In: Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, pp. 1981–1984.

Myers, S., 1999b. Surface underspecification of tone in Chichewa. Phonology 15, 367–392.

Myers, S., 2004. The effects of boundary tones on the f0 scaling of lexical tones. In: Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, pp. 147–150.

Nakajima, S., Allen, J.F., 1993. A study on prosody and discourse structure in cooperative dialogues. Phonetica 50, 197–210.

Nelson, W.L., 1983. Physical principles for economies of skilled movements. Biol. Cybernet. 46, 135–147.

Ohala, J.J., 1981. The listener as a source of sound change. In: Masek, C.S., Hendrick, R.A., Miller, M.F. (Eds.), Papers from the Parasession on Language and Behavior. Chicago Linguistic Society, Chicago, pp. 178–203.

Ohala, J.J., 1983. Cross-language uses of pitch. Phonetica 40, 1–18.

Ohala, J.J., 1984. An ethological perspective on common cross-language utilization of F0 of voice. Phonetica 41, 1–16.

Ohala, J.J., 1992. What's cognitive, what's not, in sound change. In: Kellermann, G., Morrissey, M.D. (Eds.), Diachrony within Synchrony: Language History and Cognition. Peter Lang, Berlin, pp. 308–355.

Ohala, J.J., Ewan, W.G., 1973. Speed of pitch change. J. Acoust. Soc. Amer. 53, 345(A).

Ohala, J.J., Roengpitya, R., 2002. Duration related phase realignment of Thai tones. In: Proceedings of 7th International Conference on Spoken Language Processing, Denver, Colorado, pp. 2285–2288.

Öhman, S.E.G., 1967. Word and sentence intonation, a quantitative model. Quarterly Progress and Status Report, 2, Department of Speech Communication, Royal Institute of Technology (KTH), pp. 20–54.

Peng, S.-h., 2000. Lexical versus 'phonological' representations of Mandarin Sandhi tones. In: Broe, M.B., Pierrehumbert, J.B. (Eds.), Papers in Laboratory Phonology V: Acquisition

and the Lexicon. Cambridge University Press, Cambridge, pp. 152–167.
Pike, K.L., 1945. The Intonation of American English. University of Michigan Press, Ann Arbor.
Pierrehumbert, J., 1980. The Phonology and Phonetics of English Intonation. Ph.D. dissertation, MIT, Cambridge, MA.
Pierrehumbert, J., 1981. Synthesizing intonation. J. Acoust. Soc. Amer. 70, 985–995.
Pierrehumbert, J., 2000. Tonal elements and their alignment. In: Horne, M. (Ed.), Prosody: Theory and Experiment—Studies Presented to Gösta Bruce. Kluwer Academic Publishers, London, pp. 11–36.
Pierrehumbert, J., Beckman, M., 1988. Japanese Tone Structure. The MIT Press, Cambridge, MA.
Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P.R., Morgan, J., Pollack, M.E. (Eds.), Intentions in Communication. MIT Press, Cambridge, Massachusetts, pp. 271–311.
Rose, P.J., 1988. On the non-equivalence of fundamental frequency and pitch in tonal description. In: Bradley, D., Henderson, E.J.A., Mazaudon, M. (Eds.), Prosodic Analysis and Asian Linguistics: To Honour R.K. Sprigg. Pacific Linguistics, Canberra, pp. 55–82.
Rump, H.H., Collier, R., 1996. Focus conditions and the prominence of pitch-accented syllables. Language Speech 39, 1–17.
Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigms. Speech Commun. 40, 227–256.
Schmidt, R.C., Carello, C., Turvey, M.T., 1990. Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. J. Exp. Psychol.: Human Percept. Perform. 16, 227–247.
Selkirk, E., 2002. Contrastive FOCUS vs. presentational focus: Prosodic evidence from right node raising in English. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France, pp. 643–646.
Selkirk, E., Shen, T., 1990. Prosodic domains in Shanghai Chinese. In: Inkelas, S., Zec, D. (Eds.), The Phonology–syntax Connection. University of Chicago Press, Chicago, pp. 313–337.
Shattuck-Hufnagel, S., Turk, A.E., 1996. A prosody tutorial for investigators of auditory sentence processing. J. Psycholinguistic Res. 25 (2), 193.
Shen, X.S., 1990. The Prosody of Mandarin Chinese. University of California Press, Berkeley.
Shen, X.S., 1992. On tone sandhi and tonal coarticulation. Acta Linguistica Hafniensia 24, 131–152.
Shih, C., 1988. Tone and intonation in Mandarin. Working Papers, Cornell Phonetics Laboratory No. 3, pp. 83–109.
Shih, C., 1986. The Prosodic Domain of Tone Sandhi in Chinese. Ph.D. dissertation, University of California, San Diego.
Shih, C., 1993. Relative prominence of tonal targets. In: Proceedings of the 5th North American Conference on Chinese Linguistics, Newark, Delaware. University of Delaware, World, p. 36.
Shih, C., 2000. A declination model of Mandarin Chinese. In: Botinis, A. (Ed.), Intonation: Analysis, Modelling and Technology. Kluwer Academic Publishers, pp. 243–268.
Shih, C., Sproat, R., 1992. Variations of the Mandarin rising tone. In: Proceedings of the IRCS Workshop on Prosody in Natural Speech No. 92-37, Philadelphia. The Institute for Research in Cognitive Science, University of Pennsylvania, pp. 193–200.
Silverman, K.E.A., Pierrehumbert, J.B., 1990. The timing of prenuclear high accents in English. In: Kingston, J., Beckman, M.E. (Eds.), Papers in Laboratory Phonology 1—Between the Grammar and Physics of Speech. Cambridge University Press, Cambridge, pp. 72–106.
Speer, S.R., Shih, C., Slowiaczek, M.L., 1989. Prosodic structure in language understanding: Evidence from tone Sandhi in Mandarin. Language Speech 32 (4), 337–354.
Steele, S.A., 1986. Nuclear accent F0 peak location: Effects of rate, vowel, and number of following syllables. J. Acoust. Soc. Amer. 80, S51.
Sun, X., 2002. The Determination, Analysis, and Synthesis of Fundamental Frequency. Ph.D. dissertation, Northwestern University, 2002.
Sundberg, J., 1979. Maximum speed of pitch changes in singers and untrained subjects. J. Phonetics 7, 71–79.
Swerts, M., 1997. Prosodic features at discourse boundaries of different length. J. Acoust. Soc. Amer. 101, 514–521.
Taylor, P., 2000. Analysis and synthesis of intonation using the Tilt model. J. Acoust. Soc. Amer. 107, 1697–1714.
't Hart, J., Collier, R., Cohen, A., 1990. A Perceptual Study of Intonation—an Experimental-phonetic Approach to Speech Melody. Cambridge University Press, Cambridge.
Titze, I.R., Talkin, D., 1979. A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. J. Acoust. Soc. Amer. 66, 60–74.
Thorsen, N.G., 1980. A study of the perception of sentence intonation—Evidence from Danish. J. Acoust. Soc. Amer. 67, 1014–1030.
Umeda, N., 1982. F0 declination is situation dependent. J. Phonetics 10, 279–290.
van Heuven, V.J., 1994. What is the smallest prosodic domain? In: Keating, P.A. (Ed.), Papers in Laboratory Phonology, 3. Cambridge University Press, pp. 76–98.
Wang, W.S.-Y., Li, K.-P., 1967. Tone 3 in Pekinese. J. Speech Hearing Res. 10, 629–636.
Wichmann, A., 2000. The attitudinal effects of prosody, and how they relate to emotion. In: Proceedings of ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast.
Wichmann, A., 2002. Attitudinal intonation and the inferential process. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France.
Wu, Z., 1982. Putonghua yuju zhong de shengdiao bianhua [Tonal variations in Mandarin sentences]. Zhongguo Yuwen [Chinese Linguistics] (6), 439–450.

Wu, Z., 1984. Putonghua sanzizu biandiao guilü [Rules of tone sandhi in trisyllabic words in Standard Chinese]. Zhongguo Yuyan Xuebao [Bulletin of Chinese Linguistics] 2, 70–92.

Xu, C.X., Xu, Y., Luo, L.-S., 1999. A pitch target approximation model for F0 contours in Mandarin. In: Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, pp. 2359–2362.

Xu, C.X., Xu, Y., Sun, X., 2003. Effects of consonant aspiration on Mandarin tones. J. Internat. Phonetic Assoc. 33, 165–181.

Xu, Y., 1993. Contextual Tonal Variation in Mandarin Chinese. Ph.D. dissertation, The University of Connecticut.

Xu, Y., 1994. Production and perception of coarticulated tones. J. Acoust. Soc. Amer. 95, 2240–2253.

Xu, Y., 1997. Contextual tonal variations in Mandarin. J. 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. J. Phonetics 27, 55–105.

Xu, Y., 2001. Fundamental frequency peak delay in Mandarin. Phonetica 58, 26–52.

Xu, Y., 2002. Articulatory constraints and tonal alignment. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France, pp. 91–100.

Xu, Y., 2003. Three levels of tune-text relations. In: Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, pp. 257–260.

Xu, Y., 2004a. Understanding tone from the perspective of production and perception. Language Linguistics 5, 757–797.

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, Beijing, pp. 215–220.

Xu, Y., Sun, X., 2002. Maximum speed of pitch change and how it may relate to speech. J. Acoust. Soc. Amer. 111, 1399–1413.

Xu, Y., Wallace, A., 2004. Multiple effects of consonant manner of articulation and intonation type on F0 in English. J. Acoust. Soc. Amer. 115 (Pt. 2), 2397.

Xu, Y., Wang, Q.E., 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. Speech Commun. 33, 319–337.

Xu, Y., Wang, M., 2005. Tonal and durational variations as phonetic coding for syllable grouping. J. Acoust. Soc. Amer. 117, 2573.

Xu, Y., Xu, C.X., in press. Phonetic realization of focus in English declarative intonation (to appear).

Xu, Y., Larson, C.R., Bauer, J.J., Hain, T.C., 2004a. Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. J. Acoust. Soc. Amer. 116, 1168–1178.

Xu, Y., Xu, C.X., Sun, X., 2004b. On the Temporal Domain of Focus. In: Proceedings of International Conference on Speech Prosody 2004, Nara, Japan, pp. 81–84.

Yip, M., 2002. Tone. Cambridge University Press, Cambridge.

Yuan, J., Shih, C., Kochanski, G.P., 2002. Comparison of declarative and interrogative intonation in Chinese. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France, pp. 711–714.

Zee, E., 1980. A spectrographic investigation of Mandarin tone sandhi. UCLA Working Papers in Phonetics 49, pp. 98–116.

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

Zhang, J., 2001. The Effects of Duration and Sonority on Contour Tone Distribution—Typological Survey and Formal Analysis. Ph.D. dissertation, UCLA.