In *From Traditional Phonology to Modern Speech Processing: Festschrift for Professor Wu Zongji's 95th Birthday.* G. Fant, H. Fujisaki, J. Cao and Y. Xu. (eds.) Beijing: Foreign Language Teaching and Research Press: 483-505.

Separation of Functional Components of Tone and Intonation from Observed F_0 Patterns

XU Yi

Abstract

To understand tone and intonation in speech, we need to identify their functional components. To identify these components from the acoustic signal of speech, it is critical to recognize that they are not equivalent to any directly observable surface patterns. This is because, as will be argued in this paper, there are multiple degrees of separation between functional components of tone and intonation and the surface acoustic patterns. Three degrees of separation will be identified: *articulatory implementation, target assignment* and *parallel encoding*. As the multiple degrees of separation are being recognized, the link between the surface F_0 patterns and the functional components of tone and intonation should become more transparent.

1. Introduction

One of the most important things that I learned from Professor Wu is the principle of "*ceteris paribus*". He explained the principle to me during one of the many study sessions I had with him while I was a student at the Institute of Linguistics. "*Ceteris paribus*" is a Latin expression that means "everything else being equal." The principle is about how to conduct scientific research, and it has been one of the most critical keys to the advancement of modern science. By its original spirit, the process of research must be divided into a series of independent experiments. In each experiment only one of the factors is controllably changed while the rest are kept constant. With the advancement of statistical methodology, several factors can be simultaneously controlled in a single experiment. Nevertheless, the spirit of the principle remains: to be sure that a given phenomenon is due to a particular factor, one has to be sure that all other potentially contributing factors are effectively controlled, *i.e.*, either kept constant or explicitly manipulated so that their effects are accounted for separately.

The importance of the principle of *ceteris paribus* lies in the fact that anything witnessed by an observer as a single phenomenon could be composed of multiple sub-phenomena, each with a unique underlying mechanism. Without *a priori* knowledge about their existence, it is often very hard to tell the sub-phenomena apart. As has been demonstrated by many scientific studies, this is also true of speech research, and the study of tone and intonation is no exception. More importantly, the application of the principle of *ceteris paribus* is critical not only for the individual experiments we conduct, but also for our overall understanding of how speech, and for that matter, tone and intonation, work in general.

In studying tone and intonation, an important goal is to identify the individual components and understand how they function in speech. Much of the research toward this goal is done by directly observing various aspects of the acoustic signals, including the fundamental frequency (F_0), amplitude, duration, voice quality, and spectral characteristics. Of these by far the most researched

is F_0, which is also what the present paper is mainly concerned with. To identify tonal and intonational components from F_0, much effort has been devoted to figuring out how observed F_0 curves should be *divided* into basic individual tonal and intonational components. Various proposals have been made. Some suggest that the components are in the form of rising, falling, and more complex shapes (Pike, 1945, 1948; Bolinger, 1951, 1986; Crystal, 1969; Abramson, 1978; 't Hart, Collier, & Cohen, 1990). Others suggest that they should be in the form of pitch registers such as H (high) and L (low) (Woo, 1969; Gandour, 1974; Anderson, 1978; Leben, 1978; Pierrehumbert, 1980; Pierrehumbert & Beckman, 1988; Duanmu, 1994), and each of the registers should be directly associated with F_0 peaks and valleys (Pierrehumbert, 1980, 1981; Arvaniti, Ladd, & Mennen, 1998; Ladd *et al.*, 1999; Ladd, Mennen & Schepman, 2000). Despite the dispute among the different approaches, however, they seem to share an implicit common assumption, namely, tonal and intonational components exist *overtly* in the acoustic signal, and hence can be directly observed from the F_0 contours.

In the present paper, I would like to argue, in the spirit of *ceteris paribus*, that observed F_0 contours rarely resemble the underlying forms of tonal or intonational components. Rather, F_0 is only a *reflection* of tone and intonation, and a grossly indirect one as such. This is because, as I will demonstrate, surface acoustic patterns of speech are detached from the functional components of tone and intonation by multiple degrees of separation. At least three degrees of separation can be identified: a) *articulatory implementation*, b) *target assignment* and c) *parallel encoding*.

2. Articulatory implementation

To generate speech melody related to tone and intonation is to produce temporally varying tonal patterns. This is done with the human larynx, the organ that produces the fundamental frequency of the voice (F_0). As a physical system, the larynx has many mechanical characteristics, which are inevitably reflected in the surface F_0 patterns. This section will examine those characteristics of the larynx that generate the most robust effects on F_0.

2.1 Dynamic threshold of pitch change

Unlike the piano, which can be played by pressing one key at precisely the moment when another key is released, or the pressing of the two keys can even overlap in time, the larynx can produce only one note at a time and it can shift to a new note only after the previous one is over. This means that how quickly two adjacent notes can be achieved in F_0 is dependent on how quickly the tension of the vocal folds can be changed. We may refer to the maximum speed at which speakers can voluntarily change pitch as the *dynamic threshold of pitch change*. Several attempts have been made to assess this threshold (Ohala and Ewan, 1973; Sundberg, 1979; Fujisaki, 1983; Xu & Sun, 2002). One consistent finding of these studies is that the speed of pitch change increases as the size of the change becomes bigger. At the same time, however, the time of a pitch change also increases with the size of the change (with the possible exception of lowering F_0 by professional singers (Sundberg, 1979)). As found in Xu and Sun (2002), the maximum speed of pitch change, measured in terms of peak velocity, namely, the maximum instantaneous velocity during a particular pitch change, is virtually linearly related to the size of pitch change. The following linear equations were obtained for the average speed of pitch change and time of pitch change averaged across 36 native speakers of American English and Mandarin Chinese (Xu & Sun, 2002). With these equations, given the magnitude of a particular pitch change, we can calculate both the mean maximum speed of the pitch change, and the average minimum time of the pitch change.

$s = 10.8 + 5.6\,d$ *(raising)* (1)

$s = 8.9 + 6.2\,d$ *(lowering)* (2)

$t = 89.6 + 8.7\,d$ *(raising)* (3)

$t = 100.4 + 5.8\,d$ *(lowering)* (4)

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 4 semitones. Below 4 semitones, a pitch rise is faster than a fall. Xu and Sun (2002) speculated that this may 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 which are activated only when the magnitude of F_0 lowering is large (Erickson, 1976; Erickson *et al.*, 1995; Hallé, 1994; Fujisaki, 2003).

2.2 Multiple muscles are involved in F_0 production

F_0 is produced by the vibration of the vocal folds, which is part of the larynx, a rather complex system. Producing F_0 requires the vibration of the vocal folds. According to the widely accepted *myo-elastic-aerodynamic theory* (van den Berg, 1958), the vocal folds are set into vibration by closing the glottis with the right amount of medial pressure, forcing the air from the lungs through the glottis, and creating just the right balance for the Bernoulli effect and the elasticity of the vocal folds to jointly make the glottis open and close rapidly and repeatedly. To create such a condition, a number of laryngeal muscles are involved. The posterior cricoarytenoids (PCA) need to contract to hold the arytenoids cartilages in place; the lateral cricoarytenoids (LCA) need to contract to press the vocal folds together; the transverse and oblique interarytenoids (IA) need to contract to pull the arytenoids together (depending on how non-breathy the speaker wants the voice to be (Hanson, 1997)); the cricothyroids (CT) and the thyroarytenoids (TA) both need to contract to create the right amount of tension of the vocal folds; and the internal and external intercostals and the diaphragm need to contract to control the right amount of air pressure across the glottis.

Once the vocal folds are set into vibration, the frequency and mode of the vibration is determined by the tension and the vibrating mass of the vocal folds (Titze & Talkin, 1979; Fujisaki, 1983), which are, in turn, determined by many factors, including the properties of the mucus membrane, connective tissues, the muscle tissues, the boundaries of the vocal folds, and the length of the vocal folds (Titze & Talkin, 1979). These factors are controlled by the contractions of a number of laryngeal muscles (Zemlin, 1988). Of particular importance is the fact that the vibrating vocal folds are a tissue-muscle complex, much of which is made up of TA. Because of this structural characteristic, TA can work as either a tensor or a relaxer of the vocal folds. When not opposed by other muscles, it relaxes the vocal folds and may assist in closing the glottis by drawing muscular processes of the arytenoids forward. When opposed by other muscles, it tenses the vocal folds. CT forms a agonist-antagonist pair with TA, and as such, they probably contribute to most of the precision control of F_0 (Kempster, Larson & Kistler, 1988; Zemlin, 1988). CT increases the distance between the thyroid angle and the vocal processes of the arytenoids. This increased distance in turn lengthens the vocal folds. However, increasing the length of the vocal folds alone does not necessarily lead to their increased tension. In fact, as shown by Hollien and colleagues in the early 1960s, the vocal folds are the longest at rest, and their length at various pitches never exceeds, and in fact seldom approaches the length of the vocal folds in their abducted position (Hollien, 1960; Hollien & Moore, 1960). Additionally, decreasing vocal fold tension to an extreme extent during phonation usually involves extrinsic laryngeal muscles, including the sternohyoids, omohyoids and sternothyroids (SH, OH, and ST) (Erickson, 1976; Erickson *et al.*, 1995; Hallé, 1994; Fujisaki, 2003).

To further complicate the matter, subglottal pressure is also known to be related to F_0 (Ohala, 1978; Zemlin, 1988), although it has been demonstrated that its changes do not correspond well with rapid

local pitch changes (Ohala, 1978). In recent years, there is converging evidence that subglottal pressure and related F_0 changes have much to do with control of intensity in speech. Brungart *et al.* (2002) found that F_0 changes alone can alter listeners' perceptual judgment about how far away the talker is from the listener. Watson, Ciccia and Weismer (2003) asked speakers to initiate speech from low, typical, and high lung volume levels. They found that with increased lung volume initiation levels, average sound pressure level, average F_0, and declination rate of F_0 all increased. Alku, Vintturi and Vilkman (2002) found that in producing loud voice, speakers use F_0 to increase the number of glottal closures per unit time, which raises vocal intensity. They reported that the average increase of SPL due to this active use of F_0 was approximately 4 dB in loud speech for both female and male speakers. These findings suggest an independent mechanism contributing to the surface F_0. As I will discuss later, this mechanism is probably related to the linguistic function of "new topic".

To summarize, controlling pitch takes a combination of different muscle activities. This rules out the possibility that any single muscle is solely responsible for controlling F_0. It also rules out the likelihood that passive forces such as vocal fold elasticity is employed as an effective means to lower F_0, because a decrease of vocal fold tension can be achieved more promptly and more effectively by relaxer muscles such as TA and LCA along with the reduced or halted contraction of tensor muscles such as CT.

2.3 Synchronization of laryngeal and supralaryngeal movement

In speech, laryngeal movement for producing F_0 patterns and supralaryngeal movements that generate spectral patterns have to be separately controlled. Such separation of control makes the concomitant production of lexical tones and pitch accents with consonants and vowels 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, Carello & Turvey (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 human movement control, it should apply to speech as well. That is, the synchronization constraint may force the functional 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). The average speaking rate of a normal speaker is about 5-7 syllables per second. This means that the average syllable duration is about 143-200 ms. According to equations (3) and (4), at the fastest speed of pitch change of an average speaker, it takes at least 124 ms to complete a 4-st rise or fall, and about 107 ms to complete a 2-st rise and 112 ms a 2-st fall. This means that, as far as pitch movement is concerned, things are going almost as fast as possible. This would make it very difficult for a speaker to maintain any phase relation between pitch movement and the syllable other than full synchrony. Indeed, in the case of tone production, there has been accumulating evidence that the syllable is the unit that tonal targets are aligned to.

Figure 1 displays the Mandarin F (Falling) tone when preceded by four different tones: H (High), R (Rising), L (Low), and F. As can be seen in Figure 1a, transitions toward F always start at the onset of the F-bearing syllable regardless of the distance to be covered. This is despite the fact that there is a strong articulatory constraint on the dynamic threshold of pitch change as discussed earlier. As a consequence, the location of the high F_0 turning point varies depending on the ending F_0 of the preceding tone: the lower the value, the later the turning point.

Even the transition toward an exaggerated F_0 value due to focus does not start earlier, as shown in Figure 1b. This suggests that there must be some kind of alignment constraint that is quite strong. Figure 1 also shows that, regardless of the preceding tones, the falling contour of F is always best approximated near the end of the syllable. This is further evidence that the implementation of a lexical tone in Mandarin starts at the onset of the host syllable and ends at the offset of the syllable.

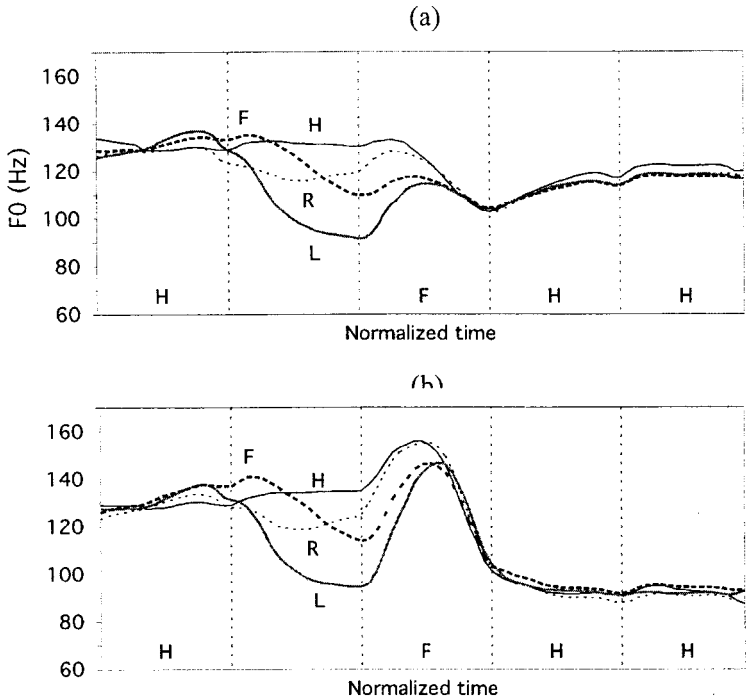

Figure 1: Mandarin tone F following four different tones. (a): no narrow focus in the sentence; (b): focus on the F-carrying syllable. Each curve is an average of 20 tokens produced by four male speakers (five repetitions per speaker). (Data from Xu, 1999).

Further evidence for tone-syllable synchronization in Mandarin comes from two sets of findings. First, Xu (1998) found that the basic F_0 pattern of a tone was more consistent when it was aligned with the entire syllable than when aligned with only the vocalic portion of a syllable with nasal coda. This remained true even when the proportional duration of the vowel and the nasal coda varied extensively due to differences in vowel height. Similar patterns have been reported for Thai (Ohala & Roengpitya, 2002). Second, Xu, Xu and Sun (in press) found that an F_0 movement toward a tonal target in Mandarin starts from the onset of a syllable even if the initial consonant is voiceless.

2.4 A pitch target model of local F_0 contour formation

The foregoing discussion has established the following:

(1) F_0 changes in any direction are under active muscle control; there is no apparent need for passive forces such as elasticity to be used as a major control mechanism.

(2) Because of inertia and possibly some other passive forces, pitch shift cannot be made instantaneously.

(3) Intended pitch movements related to lexical tones have to be synchronized with the syllable.

Based on these facts, Xu and Wang (2001) proposed the *pitch target approximation model* of tone realization. At the core of the model is the assumption that phonological tone categories are *not mapped* onto surface phonetic patterns. Rather, it is assumed that associated with each tone is an articulatorily operable unit called *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 applying a combination of muscle forces to approach the height and shape of the target. Figure 2 shows a schematic illustration of the model. The vertical lines in the figure indicate the onset and offset of two adjacent syllables. The dashed lines represent two adjacent pitch targets: a dynamic [rise] and a static [low]. , These targets are assumed to be associated with R and L carried by two adjacent syllables in Mandarin. 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 dynamic threshold of pitch change. Due to the combined pressure to realize them both as rapidly and as accurately as possible, these targets are approached asymptotically, as indicated by the shape of the solid curve corresponding to either syllable 1 or syllable 2.

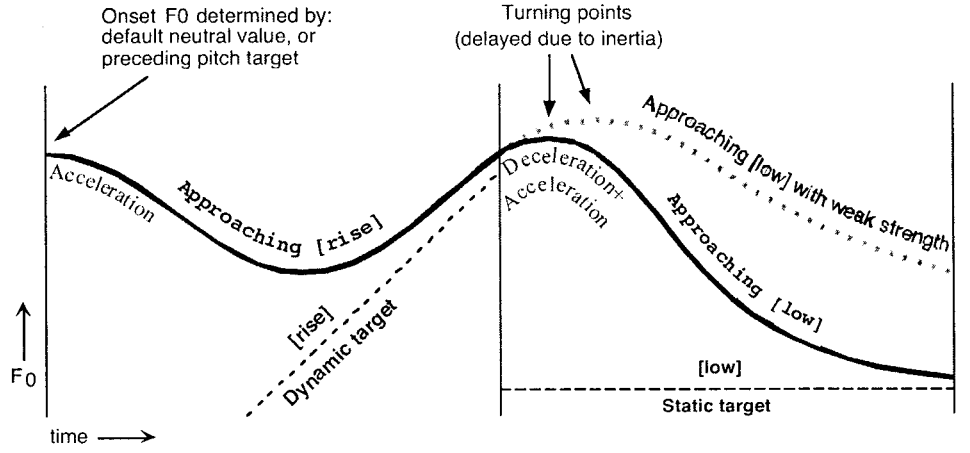

Figure 2: Dynamic and static pitch targets and their implementation. The vertical lines represent syllable boundaries. The dashed lines represent the underlying targets. The thick curve represents the F_0 contour that results from asymptotic approximation of the targets.

Note that in this model, unlike in the Fujisaki model (1983), there is no mechanism that *automatically* returns F_0 to a neutral value. This is because, first, as discussed in 2.2., articulatorily, it is unlikely that elasticity of the vocal folds constitutes an effective force in F_0 control. Secondly, acoustic data from Gandour *et al.* (1994), Xu (1997, 1998, 1999) and Li and Lee (2002) indicate that the most appropriate F_0 contour of a tone is best approximated in the final portion of a syllable, and that the subsequent F_0 contour in the following syllable always goes toward the next tonal target, as evident in Figure 1, rather than toward a common neutral value.

The only assumed passive force in the model is inertia. Due to inertia, for example, the F_0 drop from the initial relatively high F_0 in Figure 2 due to the preceding target takes some time to accelerate to full speed, resulting in a convex-up shape in the initial portion of the F_0 transition. Also in Figure 2,

the approximation of [rise] results in a fast F_0 rise at the end of the first syllable. To implement the [low] in the second syllable, however, F_0 needs to drop quickly. Deceleration of the rising movement and acceleration toward a low F_0 both take time, resulting in an F_0 peak in the initial portion of the second syllable. The effect of inertia may become more prominent when less muscle force is applied to implement a pitch target, as is illustrated by the dotted curve in the second syllable in Figure 2: when following a [rise], less articulatory force for implementing a [low] leads to both greater peak delay and slower F_0 drop toward the [low].

The pitch target approximation model was proposed to simulate some of the basic articulatory mechanisms of tone production in speech. As such, it may help to differentiate surface F_0 variations that are attributable to articulatory constraints from those that are not, as I will show in Section 3.

2.5 The total pitch range

It has been widely assumed that there is a normal pitch range within which all lexical tones in a language are produced. Based on this understanding, it could be argued that tone languages are not able to use F_0 for conveying other communicative functions. As pointed out by Chao long ago, however, there is a distinction between the "normal intonation" and "emotional intonation"(Chao, 1932). Thus the five-point scale (Chao, 1930), which adequately represents all lexical tones, probably uses only part of the total pitch range of the speaker, which should cover not only the normal scale, but also additional scales used for various pragmatic and affective functions. Indeed, according to Fairbanks (1959), a speaker's conversational pitch range can span as much as two octaves. Data from Xu (1999) indicate that at any particular sentence position, the total pitch range across the four Mandarin tones that are not under focus does not exceed one octave. This should leave a full octave of the speakers' total pitch range available for other uses.

Also worth noting is that there is a gross asymmetry in the use of one's pitch range in speech. According to Zemlin (1988), speakers regularly use only the lower part of their voice. In fact, the lower limit of one's voice pitch is constantly reached during speech, beyond which creaky voice results. This happens in the low pitched tone in many tone languages, such as the L tone in Mandarin, and in the low pitch accents of non-tone languages, such as English. In contrast, the upper limit of the pitch range is probably seldom reached in normal, non-excited speech. It is approached only when the speaker is highly excited. The pitch range beyond the "normal" tonal range thus may be employed in conveying various simultaneously encoded intonation components, as will be discussed later.

2.6 Vowel intrinsic F_0 and F_0 perturbation by consonants

Ever since Lehiste and Peterson's (1961) classic study, it has been known that consonants and vowels both introduce variations into surface F_0 contours. Since then, there has been more research on both effects. Regarding *Vowel intrinsic pitch*, it has been well established that, other things being equal, different vowels are produced with different F_0. This has been verified in virtually any languages where the effect has been looked at (Whalen & Levitt 1995). Such variations are mostly related to vowel height, with high vowels having higher intrinsic pitch than low vowels. Intrinsic pitch variations have also been found in Mandarin, *i.e.*, the same tones are produced with different F_0 under equal conditions (Shi & Zhang 1987). Although quite consistent, however, F_0 variations due to vowel intrinsic pitch may not be very large in magnitude. On average, it is in the range of 15.3 Hz or 1.65 semitones (Whalen & Levitt, 1995: 356). Also, at least one study has found that this effect is reduced in connected speech (Ladd & Silverman 1984).

Consonants may affect the F_0 contours of both preceding and following vowels. It is well known that voiceless consonants may raise the F_0 of the following vowel, and certain voiced consonant may lower the F_0 of the following vowel, but both effects are temporally quite local (Lehiste &

Peterson 1961; Howie 1974; Hombert 1978; Rose 1988). On the other hand, consonantal effect on the F_0 of the preceding vowels is not well understood, although there have been a report that there is no effect at all (van Santen & Hirschberg 1994). But informal observations have revealed certain very local lowering effects of stop consonants on the F_0 of the preceding vowel in Mandarin (Xu 1996). Further research on this effect is needed.

Since F_0 is often interrupted during the production of a voiceless consonant, a question naturally arises as to what effect such interruption may have on tonal alignment. As discussed in 2.3, there is evidence that tonal alignment remains synchronized with the syllable regardless of whether voicing is interrupted (Xu, Xu & Sun, in press).

2.7 Summary: separation due to articulatory constraints

The discussion in this section examines the nature of the first degree of separation between surface F_0 and the underlying components of tone and intonation. We have seen that the dynamic threshold of pitch change, synchronization of laryngeal and supralaryngeal movements, and consonantal perturbation all impose extensive effects on the height and shape of F_0 contours in speech. Furthermore, vowel intrinsic pitch and total pitch range introduce additional constraints on the pitch range in speech.

3. Target assignment

The foregoing discussion indicates that, regardless of what the functional components of tone and intonation are, their underlying forms are obscured by various articulatory constraints involved in the production of tone and intonation. In the following discussion I will show that the underlying functional components are further masked by the process of target assignment. That is, the assignment of targets is not only conventional and largely arbitrary, but also often *non-unique*. Such arbitrary and non-unique assignment imposes a further degree of separation between surface F_0 and the functional components of tone and intonation.

3.1 Non-unique assignment of pitch targets in Mandarin

The arbitrary and non-unique assignment of pitch targets is most vividly seen in tone sandhi, *i.e.*, the phenomenon that the realization of a tone varies with its adjacent tones. Chao (1948, 1968) documented an extensive set of sandhi patterns in Mandarin. The existence of these patterns has been largely confirmed by instrumental studies, (*e.g.*, Lin *et al.*, 1980; Lin & Yan, 1991; Shen, 1990, 1992; Shih, 1988; Wu, 1982, 1984; Xu, 1997, 1999). Chen (2000) devoted an entire volume to the documentation of various sandhi phenomena in a large variety of languages and dialects. The mechanisms behind these phenomena, however, remain mostly unclear, as pointed out by Chen (2000: 25f). The target approximation model described in 2.4, since it is based on explicit assumptions about the mechanisms of F_0 generation, may shed some light on this complicated issue. As I will show in the following discussion, it may at least help distinguish between types of tonal variations: those due to *target alternation* and those due to articulatory implementation (*implementational variation*). Target alternation occurs in cases where 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 the interaction between articulatory constraints and implementational strength. In the following, I will focus mostly on tone sandhi patterns in Mandarin, for which extensive systematic acoustic data are available.

3.1.1 L variations

Three types of variations of the tone L in Mandarin can be seen in Figure 3. The figure shows average F_0 curves of the four Mandarin tones said in isolation (data from Xu, 1997). Note that L in

this graph has a final rise, which largely agrees with Chao's (1968) description. Figure 3b displays disyllabic sequences produced in a carrier. Here, syllable 2 carries L while syllable 1 carries four different tones. We can see that L in syllable 2 has no trace of the final rise seen in the left graph. It could be argued that this lack of final rise is due to some kind of articulatory constraint, because with a carrier, the duration of each syllable should be much reduced. It is true that the duration of the L-carrying syllable in Figure 3b is shorter than that in Figure 3a (177 *vs.* 349 ms, based on data from Xu 1997). At the same time, however, 177 ms is still longer than the minimum time needed to lower pitch by the amount shown in Figure 3b. The amount of lowering in syllable 1 in the F L sequence, for example, is about 6 St. According to equation (4), an average speaker needs only about 135 ms to complete this much lowering at the maximum speed. This means that, had the pitch target being implemented for L in this situation been [low]+[high] or [low]+[mid], there would have been time for the final rise to be at least partially realized. In fact, making two movements within one syllable is not only possible, but also frequently done, as in the case of the dynamic tones such as R and F. Thus it is rather unlikely that the lack of final rise in L is due to the dynamic threshold of pitch change. It is more likely, instead, that the pitch target implemented for L in a non-final position 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 before the actual articulatory implementation begins.

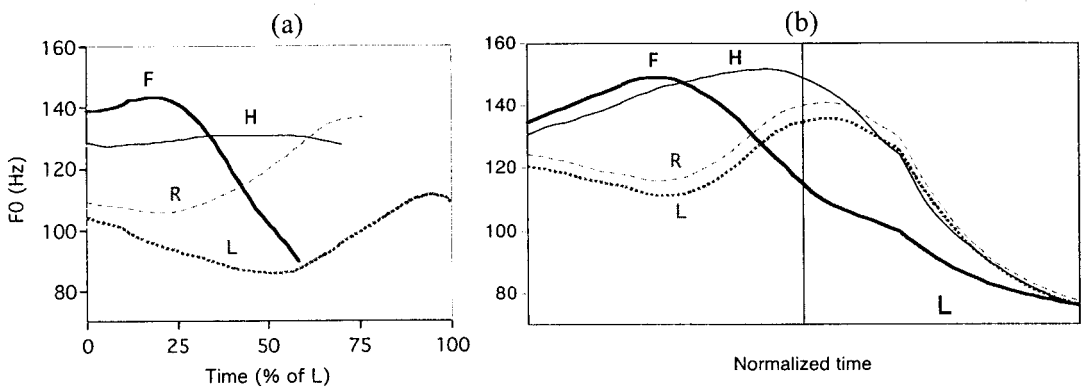

Figure 3: (a): Four Mandarin tones produced in isolation. (b): Mandarin L tone after four different tones, produced in carrier phrases. Adapted from Xu (1997).

The second type of L variation can be also seen in Figure 3b. 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. Although subsequent 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 (Xu, 1993, 1997; Zee, 1980) as is also apparent in Figure 3b, 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 pitch target approximation model. This is because the model provides no mechanism for generating a falling-rising contour by asymptotically approaching a [low] target. Would it be possible, however, that the R-like F_0 contour in the L L sequence results from implementing a complex target that is similar to that associated with the citation form of L as seen in Figure 3a? It is not totally inconceivable. For one thing, since the syllable duration for L in isolation is vastly different from that before another L: 349 *vs.* 177 ms. It is possible that the only viable way to squeeze a complex target such as [low+high], [low+mid] or [low+rise] into a shortened syllable is to sacrifice the F_0 minimum in favor of maintaining the whole contour shape, thus resulting in an F_0 contour not very different from that of R. One difficulty with this account is

that it has to provide a mechanism that raises not only the F_0 minimum from 85 Hz in Figure 3a to 110 Hz in Figure 3b, but also the maximum F_0 from 110 Hz in the former to 130 Hz in the latter. Although anticipatory raising reported in Xu (1997) may be a potential candidate, its average magnitude is only about 10 Hz for R. Thus it remains unclear whether the pitch target implemented for the first L in L L is the same as in that in R or that in isolated L. Nevertheless, it is at least very unlikely to be the same as that in L in other non-final positions which is presumably a static [low].

The third type of L variation can be seen in syllable 2 in Figure 3b. 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 values, are likely directly related to inertia, as assumed in the pitch target approximation model. They should therefore be considered as cases of implementational variation.

3.1.2 R variations

Xu (1994) made two findings about the R tone in Mandarin: (a) R produced on the second syllable of a tri-syllabic words is severely flattened if the tone of the first syllable is H or R and the tone of the third syllable is R or L; and (b) despite the flattening, R is still correctly identified about 88% of the time when listeners hear it in the original tonal context (with semantic information removed). Shih and Sproat (1992) report that R produced in the {H, R} __ T (where T represents any tone) context on the second syllable of a tri-syllabic word, though much distorted, is still distinct from H in the same position. They further find that the amount of distortion of R in different tonal contexts and rhythmic structures is related to the strength of the R-carrying syllable. Thus there is even greater distortion of R if it is on the second syllable of a quadrasyllabic word than if it is on the second syllable of a tri-syllabic word, because syllable strength is presumably even weaker in the former than in the latter. It is therefore likely that the variant form of R as described by Chao (1968) is a case of *implementational variation* with no change in the underlying pitch target.

3.1.3 F variations

Figure 4a shows mean F_0 curves of Mandarin F produced before four different tones in disyllabic sequences. Although there are small variations around the boundary between the two syllables, F_0 in the first syllable never approaches the bottom of the pitch range as indicated by the final point of L in the second syllable in Figure 4a or that of L in the first syllable in Figure 4b. Rather, it *always* reaches about half way toward the lowest point. This is in sharp contrast to the final F_0 reached in F produced in isolation as shown in Figure 3a. In effect, therefore, the F_0 contour of F is a "half fall" when followed by *any* tone in a disyllabic sequence. The Half F rule thus should be more appropriately stated as 51 → 53 / __ T.

As for why F becomes a "half fall", it has been suggested that this is because all Mandarin tones start either from 5 or 3, which then becomes a limiting factor for how low the previous tone can go (Shih, 1988). One problem with this account is that it is not true that a tone can never start from the bottom. As can be seen in Figure 4b, H, R, and F all start from the bottom. This despite the fact that the ideal starting pitch for theses tones are very different: high for H and F, and low for R. But this should be exactly the case according to the target approximation model (2.4.), because the onset F_0 of any tone is determined by the offset F_0 of the preceding tone. So, that F_0 does not fall to the bottom of the pitch range before another tone cannot be attributed to the characteristics of the following tone. Rather, it is more attributable to the fact that there *is* a following tone. Would it be possible, then, that having a following tone reduces the duration of the F-carrying syllable, thus leaving insufficient time for F_0 to fall to the bottom? In Figure 4a, however, the amount of F_0 drop within the first syllable is about 5 semitones, which, according to equation (4), should take 129 ms to complete at the maximum speed of pitch change. The fact that the mean syllable duration for F in this case is 178 ms (data from Xu, 1997) indicates that speakers were far away from their dynamic

threshold of pitch lowering. In fact, even if we include the entire fall from the top of F in syllable 1 to the bottom of L in syllable 2, the drop is about 12.5 semitones, and the minimum time needed for that is only 173 ms according to equation (4). This tells us that F_0 of F falls only to the mid pitch range not because of some articulatory constraint, but because it is probably the targeted height when the tone is not utterance final.

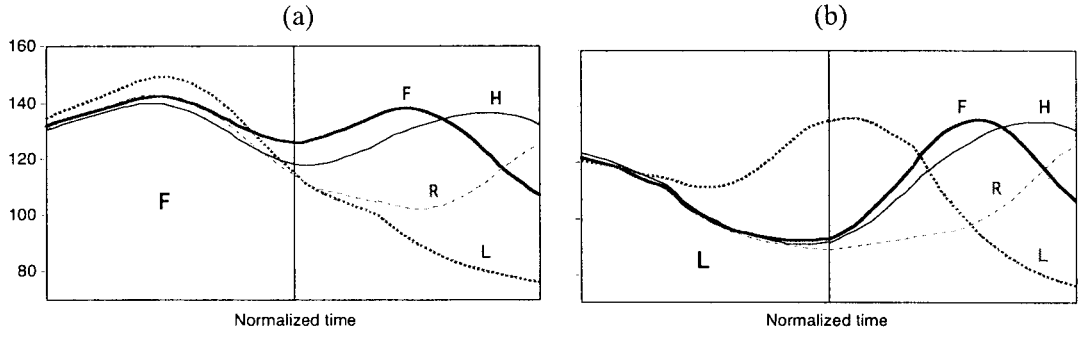

Figure 4: Effects of the following tone on the F_0 contour of the preceding tone in Mandarin syllable sequence /mama/ with different tonal combinations. In each panel, the tone of the first syllable is held constant: F in (a) and L in (b), while the tone of the second syllable is either H, R, L or F. The vertical lines indicate the onset of the initial nasal in syllable 2. Adapted from Xu (1997).

3.2 Non-unique assignment of pitch targets in Yoruba and English

Yoruba is known to have two tone spreading rules: (a) H → R / L __; (b) L → F / H __. Since they are assimilatory in nature, I once suggested that they were similar in nature as the Mandarin R variation as discussed in 3.1.2 (Xu, 2002). However, my recent informal observation told me that the rules would apply even at very low speaking rate or even across pauses. This was confirmed by personal communication with Ian Madison. Further research with control of speaking rate may help verify this observation.

Arbitrary and non-unique assignment of pitch target is not restricted to tone languages only. Xu and Xu (forthcoming) found evidence suggesting that in English declarative sentences, non-focused, non-final accents carry a static [high] target, whereas word-final accents under focus and sentence-final accents carry a dynamic [fall].

3.3 Summary: separation due to target assignment

To summarize, the discussion in this section demonstrates that target assignment is a separate process from articulatory implementation of the tonal targets. The often non-unique assignment of the target assignment thus further separates the functional components of tone and intonation from the surface F_0.

4. Parallel encoding

The discussion so far has established two degrees of separation between surface F_0 and the functional components of tone and intonation. As I will demonstrate next, a further degree of separation is imposed by *parallel encoding*. Probably because there are a large numbers of them, many communicative functions have to be transmitted simultaneously. For parallel encoding to be effective when using the same parameter, namely, F_0, there should be distinct manners with which the parameter is used. That is, there should be *encoding schemes* that are distinct from one another and readily decodable. In the following, I will first consider what are the likely and unlikely melodic primitives, i.e., the basic elements used by the encoding schemes. I will then discuss the

communicative functions that need to be encoded in parallel and how the encoding is possibly done.

4.1 Possible melodic primitives

4.1.1 Likely primitives

4.1.1.1 Local pitch target

Local pitch targets refer to the smallest articulatorily operable pitch units, as assumed in the pitch target approximation model (2.4). Local pitch targets can be either static, such as [high], [mid] and [low], or dynamic, such as [rise] and [fall]. They are assigned to a segmental host, usually a syllable. It is possible that the host can be smaller, such as a mora, or larger, such as a multi-syllabic prosodic word. But there has been no clear evidence for either of them. The most definitive experimental evidence so far seems to point to the syllable rather than any other unit as the host for local pitch target (Xu, 1998, 2001; Xu & Wang, 2001; Xu & Xu, forthcoming). It is also possible that in any language, every syllable needs to be assigned a pitch target, even if it is just a default neutral target when no distinctive target is specified (Xu & Xu, forthcoming; Chen & Xu, 2002).

Although they constitute just one of the melodic primitives, local pitch targets are probably the most basic melodic elements, because other primitives, at least those controlling F_0 directly, as will be shown next, function by modifying the manner with which local pitch targets are implemented.

4.1.1.2 Pitch range

Pitch range specifies the pitch interval within which local pitch target is implemented. It can be defined by two parameters: height and span (Ladd, 1996), which may be manipulated independently. As discussed in Section 2.5., for an average speaker, the entire exploitable pitch range is quite large, about 2 octave (Fairbanks, 1959). Lexical tone, however, takes up only one octave of the pitch range (Xu, 1999). This leaves much room for various other functions. Existing literature shows that focus and new topic both use extra pitch ranges, but in different ways, as will be discussed later. Certain affective functions may also employ extra pitch ranges. Nevertheless, so far, pitch range remains a relatively under-researched aspect of intonation.

4.1.1.3 Articulatory strength

To implement local pitch targets, actual physical effort needs to be exerted. In the pitch target approximation model (2.4), the amount of physical effort determines how effectively a pitch target is implemented during articulation. Other things being equal, a greater strength enables a pitch target to be approached sooner than a weaker strength. Recent research on Mandarin Neutral tone (Chen & Xu, 2002) and English intonation (Xu & Xu, forthcoming) shows evidence that articulatory strength can be used as an effective melodic primitive. Similar evidence has been seen in research on prosodic structure and intonation modeling (Shih & Sproat, 1992; Shih, 1993; Kochanski & Shih, 2003). Due to its potential implications, the role of articulatory strength as a linguistic parameter needs to be substantiated through further research.

4.1.1.4 Duration, intensity and voice quality

These parameters are less directly related to F_0, and so they will be discussed only briefly. First, duration seems to be an important encoding element for stress and rhythm. For example, it has been established that the Neutral tone in Mandarin is much shorter than other tones (Lin, 1985). Such durational variation should affect F_0 contours, according to the target approximation model. Everything else being equal, shorter duration makes a pitch target less likely to be reached by the end of the syllable. Thus shortened duration may be used in conjunction with reduced articulatory strength for encoding weak elements such as the Neutral tone in Mandarin (Chen & Xu, 2002) and weak stress in English (Xu & Xu, forthcoming).

Both intensity and voice quality may be used in conjunction with F_0 to encode various communicative functions, such as lexical register (Ladefoged & Madison, 1996), stress, focus, and emotion. However, their relation with F_0 is still largely unclear.

4.1.2 Unlikely melodic primitives

4.1.2.1 F_0 turning points

Although they have been extensively investigated in recent studies, and highly consistent alignment patterns have been reported for various languages (Arvaniti, Ladd, & Mennen, 1998; Ladd *et al.*, 1999; Ladd, Mennen & Schepman, 2000; Xu, 1999, 2001), F_0 turning points, especially their alignment with consonants and vowels, are unlikely to be one of the melodic primitives for intonation components. This is for several reasons. Firstly, turning points are not dependable, because sometimes they simply do not occur. This is true when two adjacent tones share a similar F_0 across the boundary, *e.g.* H H, or the same F_0 movement continues across the syllable boundary, *e.g.* R H and F L. Secondly, alignment of turning point often does not remain consistent even when the underlying tone stays the same, as is the case with the F_0 peaks in Figure 1. Finally, both the consistent and variable alignment of F_0 turning points can be predicted by the pitch target approximation model in which simple linear pitch targets are implemented in synchrony with their hosts, as argued in Xu and Wang (2001) and Xu (2002).

4.1.2.2 Downstep and declination

Both downstep and declination have been reported for many languages and are often considered to be universal. Downstep refers to the phenomenon that in a HLH sequence, the second H is lower in F_0 than the first, which has been found in both tone languages (Stewart 1965; Meeussen 1970; Hyman & Schuh 1974; Clements & Ford 1979; Shih 1988; Manfredi 1993) and non-tone languages (*e.g.* Pierrehumbert 1980; Pierrehumbert & Beckman 1988; Prieto, Shih, & Nibert 1996). Declination refers to the tendency for F_0 to gradually decline over the course of an utterance (Cohen & 't Hart 1965; Cohen, Collier & 't Hart 1982), which has been reported also for both tone languages (Shih 2000) and non-tone languages (Pike 1945; Cohen & 't Hart 1965; Maeda 1976; Cohen, Collier, & 't Hart 1982; Cooper & Sorensen 1981; Ladd 1984).

On the other hand, there has been accumulating evidence that both downstep and declination are byproducts of a mixture of more basic mechanisms. Downstep, at least in its classical form, is likely to be the byproduct of two rather independent effects: anticipatory raising and carryover lowering. That is, a L (and probably any tone that has a non-high component) raises the F_0 of the preceding syllable (Laniran, 1992; Laniran & Clements, 2003; Gandour *et al.*, 1992; Gandour, Potisuk & Dechongkit, 1994; Xu, 1993, 1997, 1999) and lowers the F_0 of the following syllable (Gandour *et al.*, 1994; Xu, 1997, 1999). The combination of the two effects therefore makes the first H much higher than the second. The automatic nature of these two mechanisms makes it unlikely that this type of downstep is used as an effective melodic primitive. Downstep is sometimes also used to refer to stepwise deliberate lowering of pitch, such as in the calling intonation in English (Pierrehumbert, 1980; Liberman & Pierrehumbert, 1984). Such downstep, however, does not seem to be very different from two successive pitch targets differing in height, *e.g.* [high] + [mid]. Thus it should not be confused with the automatic downstep just described.

For declination, there is evidence that it results also from a combination of independent effects, including downstep, focus and new topic (Pierrehumbert, 1980; Liberman and Pierrehumbert, 1984; Prieto *et al.*, 1996; Umeda 1982; Xu, 1999). When all of these effects are individually accounted for, the residual overall downtrend becomes rather small (Laniran, 1992; Xu, 1999; Laniran & Clements, 2003; Wang, 2003).

4.2 Encoding communicative functions in parallel

Due to space limit and lack of sufficient experimental data, I will discuss only a few functions for which there are relatively clear data. Lexical tones undoubtedly constitute an important communicative function, sine they serve to distinguish words just as consonants and vowels do. But since tones have already been discussed in several places, I will focus only on higher level functions in the following discussion.

4.2.1 Focus

Focus has been increasingly recognized in recent years as an independent linguistic function with robust acoustic manifestation. For example, if the sentence "Mary gave John the book" is said in response to the question "Who did Mary give the book to?", the word "John" is naturally emphasized, hence, "focused." A narrow definition of focus is therefore a discourse/pragmatics motivated emphasis. This definition makes focus distinct from lexical stress and accent. A finer distinction is sometimes made between emphatic focus and contrastive focus (or between even more kinds of focus, cf. Gussenhoven, in press). The aforementioned sentence would be an example of the emphatic focus. An example of the contrastive focus would be "Mary gave *John* the book, not Bill" in response to the statement "Mary gave Bill the book." Note that in both cases the speaker's choice of whether and where to use focus is based on the assessment of the information flow in the discourse rather than on any other concerns (Bolinger, 1972, 1989; van Heuven 1994). In other words, the location of focus is largely *independent* of lexical tone, lexical stress, syntax, and prosodic structure of the sentence. Furthermore, it has been demonstrated that units as small as a single segment can be put under focus if that is the only thing that needs to be emphasized (van Heuven 1994). Thus, neither the location nor 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, would have a default focus on the wh-component whether or not the syntax of the language requires it to move its position in sentence (Ishihara, in press).

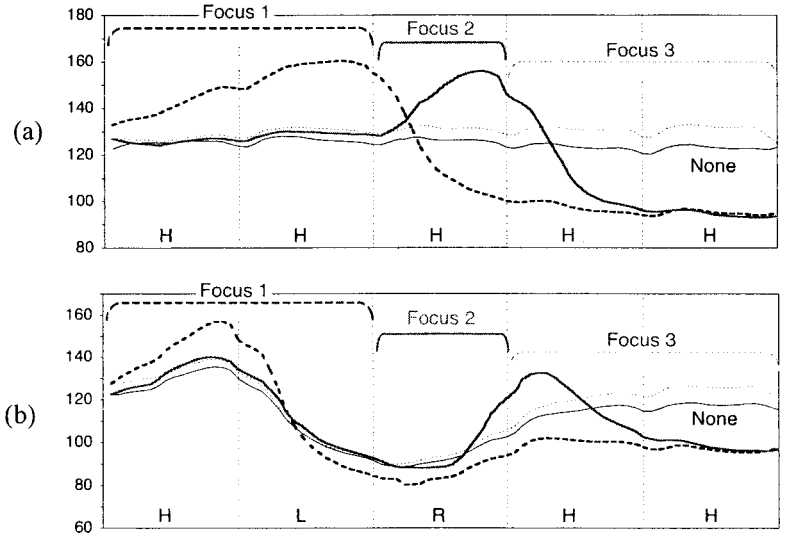

Figure 5: Pitch range variation due to focus in different positions. The curves in the figure are time-normalized F_0 averaged across 24 repetitions by 4 speakers (data from Xu, 1999).

There has been extensive evidence that focus is encoded in F_0 through manipulation of several pitch ranges around focus: expanding the pitch range of focused component, compressing that of the post-

focus components, and leave that of the pre-focus components largely neutral (Bruce 1977; Bruce & Touati, 1992; Cooper, Eady & Mueller, 1985; Gårding, 1987; Jin, 1996; Rump & Collier, 1996; Xu, 1999; Xu & Xu, forthcoming; Xu, Xu & Sun, in press). These pitch range manipulations can be clearly seen in Figure 5 for Mandarin.

Based on previous findings as well as new data from experiments on "prosodic restoration" and "imitation via prosodic restoration," Xu, Xu and Sun (in press) propose that focus is conveyed through a tri-zone pitch range control. According to this understanding, pitch range variations in all three regions are components of the encoding scheme of focus. This implies that the same kind of pitch range variations are no longer available for encoding other functions, unless *additional* pitch range variations are imposed.

4.2.2 New topic

As found in several studies (Lehiste 1975; Nakajima & Allen 1993; Umeda 1982), the F_0 of the first accented word in the first sentence of a paragraph is often much higher than in the later portion of the paragraph. Umeda (1982) suggests that an exceedingly high F_0 peak at the onset of the first sentence of a paragraph, which is different from the peaks that occur in stressed syllables in later words, is probably used as a beginning signal for a new topic. Nakajima & Allen (1993) reported data on high F_0 values related to topic-initiation (referred to as topic shift in their paper). Further evidence comes from investigation of F_0 reset between adjacent sentences as a discourse function (*e.g.* Swert, 1997). More recent research findings suggest that the function of new topic is to raise listener's attention rather than to contrast between different linguistic categories. Support for this understanding can be found in findings by Alku, Vintturi and Vilkman (2002) and by Brungart *et al.* (2002), as was discussed in 2.2. These findings seem to point to a non-contrastive attention raising function that is achieved through increasing both intensity and F_0. Needless to say, this hypothesis needs to be verified through properly designed experiments.

4.2.3 Syntactic structure: sentence type

Although it has been often assumed that syntax has a direct role in intonation, few syntactic functions have been demonstrated definitively to play such a role (Shattuck-hufnagel & Turk 1996). Even the well-known question/statement dichotomy has been found not to be the ultimate determinant of final rise/fall in intonation (Bolinger 1989). Also as argued by Shih (1986), many conditional tonal variations are determined by the prosodic rather than syntactic structure of an utterance. Nonetheless, some syntactic functions may still be correlated with certain F_0 variations (Shattuck-hufnagel & Turk 1996). It is possible that the correlation of syntax with F_0 is mostly through the mediation of semantics, pragmatics and discourse. That is, certain syntactic structures are more likely related to certain pragmatic functions than others. A good example is the question intonation. As Bolinger (1989) convincingly argued, it is mostly meaning rather than syntactic structure that determines the intonation. Regarding questions, whether with question syntax or not, different kinds of meanings may be conveyed through F_0 patterns: incredulity, inquiry, politeness, uncertainty, confirmation, or even order. It is still an open question as to how different languages encode these shades of meaning differently in question intonation. For English, there is evidence that under simple experimental conditions, speakers produce questions not only by raising F_0 at the end of the sentence, but also by using a low-rise pitch for the focused word (Eady & Cooper, 1986). Thus question intonation in English involves at least two encoding schemes: assigning low-rise target to focused word, and raising sentence final F_0, probably by assigning a high boundary tone (Pierrehumbert and Hirschberg, 1990), which could be a special kind of local pitch target. In a tone language such as Mandarin the question intonation has to be encoded in parallel with the lexical tones as well as other intonational functions such as focus. While there has been much research on how question intonation can coexist with lexical tones (Chao, 1968; Ho, 1977; Wu, 1982, 1984;

Yuan, Shih & Kochanski, 2002), its relation with focus is still rather unclear. Further research is therefore needed.

4.2.4 Lexical stress, Neutral tone, and accent

In languages like English, in addition to focus, there are also other prominence related factors, *e.g.* lexical stress and accent. Lexical stress is known to serve to distinguish between certain words. For English, it has been demonstrated that the most effective acoustic cue for stress is F_0 (Fry, 1958). Unlike lexical tone in a tone language, however, lexical stress in English may have different pitch targets depending on other linguistic functions. There has been evidence that the stressed syllable has very different pitch targets when the sentence is a statement and when it is a question (Eady *et al.*, 1986). The stress function is encoded not only in the F_0 of the stressed syllable, but also in that of the unstressed syllable. A recent study of English intonation found that the F_0 contours of unstressed syllables in English are likely produced with much weaker articulatory strength than stressed syllables (Xu & Xu, forthcoming). A recent study of Mandarin Neutral tone also found similar weak articulatory strength in Neutral tone syllables (Chen & Xu, 2002). These findings demonstrate the effectiveness of articulatory strength as a linguistic parameter. For intonation, it provides a mechanism for generating F_0 contours in unstressed or Neutral tone syllables previously believed to be generated through *interpolation* between stressed or full tone syllables.

Accent has been used to refer to various things in speech prosody. Its definition often overlaps with other prosodic functions. The term *nuclear accent*, for example, largely overlap with *focus*. With the increased use of the notion of focus whose definition is more restricted, it is probably beneficial to exclude focus from accent. What would be left in the term "accent," however, may still be a mixed bag. It could include prominence variations due to newness of information, part of speech or rhythm and prosodic grouping. Thus there is a clear need to separate both focus and lexical stress from accent. What remains in "accent," then, can only be elucidated by further research.

4.2.5 Other pragmatic functions and emotion

The foregoing discussion covers only a small portion of the communicative functions that are conveyed through F_0. There are still a vast amount of other pragmatic, attitudinal as well as emotional functions that may be conveyed mainly or partially through F_0. The importance of understanding the functions discussed above is that after their establishment, it should become easier to recognize the other functions, because they are more likely to be conveyed in parallel with the known functions by further modifying pitch range and articulatory strength, and/or by introducing additional pitch target assignment rules.

4.3 Summary: separation due to parallel encoding

With parallel encoding, different communicative functions can be transmitted simultaneously through F_0. From the perspective of the researchers, however, such parallel encoding imposes an additional degree of separation between surface F_0 and the functional components of tone and intonation. Thus during any time interval, and indeed at any particular moment, various aspects of F_0, such as height, contour, velocity, turning points, *etc.*, carry information about different tonal and intonational functions simultaneously.

5. General summary and conclusion

I would like to wrap up the discussion with an illustration. Figure 6a displays F_0 tracing of a single repetition of the Mandarin sentence "Māomǐ mō māomǐ." What can be observed in the curve are two peaks, a valley and an overall downtrend. These F_0 events, while clearly discernable, and having been the focus of many studies, do not seem to directly tell us much about the forms of the underlying tonal and intonational components. Figure 6b shows averaged F_0 curves of the same

sentence said with and without initial focus (thick/thin solid curves), together with an all-H sentence as reference (dashed curve). Also displayed in the figure are lexical tones and segmental boundaries of each syllable. As is evident from Figure 6b, the effects of all three degrees of separation should be taken into consideration before the F_0 events shown in Figure 6a can be comprehended. First, pitch target assignment designates an underlying tonal target to each syllable, as represented by the short horizontal lines. Second, articulatory implementation makes F_0 in each syllable approach the assigned target asymptotically, giving rise to the extensive transitions during syllables 1-3, as indicated by the thin arrows. It also produces the peaks in syllables 1 and 3 (the latter only when with initial focus), and the valley in syllable 3. Also can be seen in the figure are the mechanical effects of downstep brought about by L, which raises F_0 of the preceding H and lowers the F_0 of the following H. Third, the extra expansion of the pitch range of the first word, as indicated by the two block arrows, and the lowered pitch range of words 2 and 3 in the thick curve, as indicated by the filled block arrow, are directly related to the initial focus on word 1. Finally, even in the all-H sentence, there is a slight drop of F_0 from the first H to the last. This could be related to new topic. But because potential new topic effects were not manipulated in that study, this cannot be known for certain.

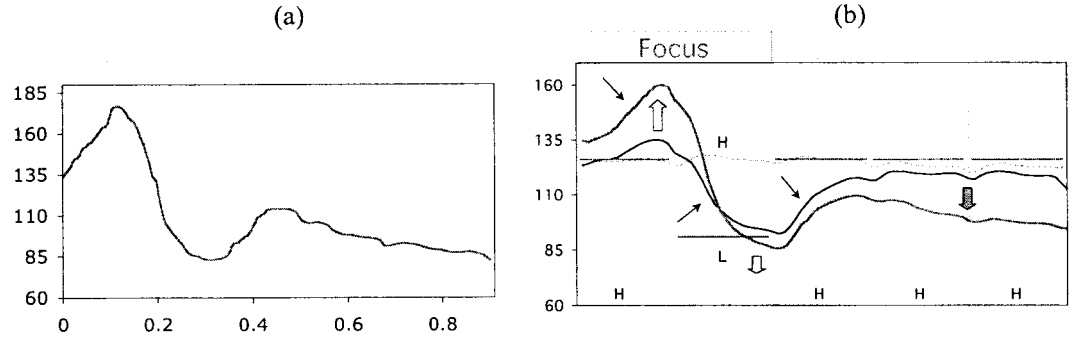

Figure 6: (a): F_0 of a single repetition of the sentence "Māomǐ mō māomī" [Cat-rice strokes Kitty], with focus on the first word "Māomǐ." X-axis: time in second. Y-axis: F_0 in Hz. (b): Averages F_0 of 20 repetitions of HLHHH and HHHHH sequences by 4 male speakers. Thick solid curve: focus on "Māomǐ."; thin solid curve: no focus; dashed curve: HHHHH. 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.

In conclusion, the principle of *ceteris paribus* says that we can be certain about the mechanism of a phenomenon only when everything else that may have introduced variations along the same dimensions of the phenomenon is held constant. In the spirit of this principle, the basic form of each tonal or intonational component of speech is only the part that is independent of everything else that may have produced influences along the dimensions of the component. Application of this understanding has lead to the recognition of at least three degrees of separation between the underlying tonal and intonational components and the observed surface acoustic patterns: *articulatory implementation*, *target assignment* and *parallel encoding*. Articulatory implementation introduces mechanical characteristics of the articulators as well as characteristics of the motor control system. Target assignment introduces stipulative alternation of underlying tonal targets. Parallel encoding introduces characteristics of different tonal and intonational functions that are simultaneously transmitted through F_0. With the recognition of multiple degrees of separation, the link between the surface F_0 patterns and the functional components of tone and intonation should

become more transparent. Note that our understanding of the three degrees of separation is still rather preliminary. In particular we still cannot clearly identify all the detailed relations among the individual components. And we still know very little about the coding schemes of many more pragmatic and affective functions. So, eventually we may discover even more degrees of separation between functional components of tone and intonation and surface F_0 contours.

Acknowledgment

This study was supported in part by NIH grant DC03902.

References

Abramson, A. (1978). The phonetic plausibility of the segmentation of tones in Thai phonology. *Proceedings of The twelfth International Congress of Linguistics*, Vienna, 760-763.

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 Communication 38*, 321–334.

Arvaniti, A., Ladd, D. R. & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics,* 36, 3-25.

Bolinger, D.L. (1951). Intonation: levels versus configuration. *Word*, 7, 199-210.

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

Bolinger, D. (1986). *Intonation and its parts: melody in spoken English*, Palo Alto: Stanford University Press.

Bolinger, D. (1989). *Intonation and Its Uses -- Melody in Grammar and Discourse*, Stanford, California: Stanford University Press.

Bruce, G. (1977). Swedish word accents in sentence perspective. In B. Malmberg and K. Hadding (eds.). *Travaux de L'institut de Linguistique De Lund,* Xii, Lund: Gleerup.

Bruce, G. & Touati, P. (1992). On the analysis of prosody in spontaneous speech with exemplification from Swedish and French. *Speech Communication*, 11, 453-458.

Brungart, D.S., Kordik, A.J., Das, K. & Shawy, A.K. (2002). The Effects of F_0 Manipulation on the Perceived Distance of Speech. *Proceedings of 7th International Conference On Spoken Language Processing*, Denver, Colorado, 1641-1644.

Chao, Y.R. (1930). A system of "tone letters". *Le Maître Phonétique*, 45, 24-27.

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*, 105-156.

Chao, Y.R. (1948). *Mandarin Primer*, Cambridge: Harvard University Press.

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

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

Chen, Y. & Xu, Y. (2002). Pitch Target of Mandarin Neutral Tone. *Presented at LabPhon 8*, New Haven, CT.

Clements, G.N. & Ford, K. (1979). Kikuyu tone shift and its synchronic consequences. *Linguistic Inquir,* 10, 179-210.

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. *Proceedings of the Fifth International Congress on Acoustics*. D.E.Commins, Liège, A. 16.

Cooper, W.E., Eady, S.J. & Mueller, P.R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America,* 77, 2142-2156.

Cooper, W.E. & Sorenson, J.M. (1981). *Fundamental frequency in sentence production.* New York: Springer-Verlag.

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

Duanmu, S. (1994). Against contour tone units. *Linguistic Inquiry* **25**: 555-608.

Eady, S.J. & Cooper, W.E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America,* 80, 402-416.

Erickson, D., Honda, K., Hirai, H. & Beckman, M.E. (1995). The production of low tones in English intonation. *Journal of Phonetics,* 23, 179-188.

Erickson, D.M. (1976). *A Physiological Analysis of the Tones of Thai.* Dissertation: The University of Connecticut.

Fairbanks, G. (1959). *Voice and Articulation Drillbook*, New York: Harper & Row.

Fry, D.B. (1958). Experiments in the perception of stress. *Language and 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. *The Production of Speech.* P. F. MacNeilage, New York: Springer-Verlag, 39-55.

Fujisaki, H. (2003). Prosody, Information, and Modeling — with Emphasis on Tonal Features of Speech. *Proceedings of Workshop on Spoken Language Processing*, 5-14.

Gandour, J. (1974). On the representation of tone in Siamese. *UCLA Working Papers in Phonetics,* 27, 118-146.

Gandour, J., Potisuk, S. & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics,* 22, 477-492.

Gandour, J., Potisuk, S., Dechongkit, S. & Ponglorpisit, S. (1992). Anticipatory tonal coarticulation in Thai noun compounds. *Linguistics of the Tibeto-Burman Area,* 15, 111-124.

Gårding, E. (1987). Speech act and tonal pattern in Standard Chinese. *Phonetica,* 44, 13-29.

Gussenhoven, C., (ed.). *Types of focus in English.* Topic and Focus: Intonation and Meaning. Theoretical and Crosslinguistic Perspectives. Dordrecht: Kluwer. (in press).

Hallé, P.A. (1994). Evidence for tone-specific activity of the sternohyoid muscle in Modern Standard Chinese. *Language and Speech,* 37, 103-123.

Hanson, H.M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America,* 101, 466-481.

Ho, A.T. (1977). Intonation variation in a Mandarin sentence for three expressions: interrogative, exclamatory and declarative. *Phonetica,* 34, 446-457.

Hollien, H. (1960). Vocal pitch variation related to changes in vocal fold length. *Journal of Speech & Hearing Research,* 3, 150-156.

Hollien, H. & Moore, G. P. (1960). Measurements of the vocal folds during changes in pitch. *Journal of Speech & Hearing Research,* 3, 157-165.

Hombert, J.-M. (1978). Consonant types, vowel quality, and tone. In V. A. Fromkin (ed). *Tone: A linguistic survey.* New York: Academic Press, 77-111.

Howie, J.M., (1974). On the domain of tone in Mandarin. *Phonetica,* 30, 129-148.

Hyman, L. & Schuh, R. (1974). Universals of tone rules. *Linguistic Inquiry,* 5, 81-115.

Ishihara, S. Syntax-Phonology Interface of Wh-Constructions in Japanese. *Proceedings of TCP2002,* (in press).

Jin, S. (1996). *An Acoustic Study of Sentence Stress in Mandarin Chinese.* Dissertation, The Ohio State University.

Kelso, J.A.S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulatory, Intergrative and Comparative,* 246, R1000-R1004.

Kempster, G.B., Larson, C.R. & Kistler, M.K. (1988). Effects of electrical stimulation of cricothyroid and thyroarytenoid muscles on voice fundamental frequency. *Journal of Voice,* 2, 221-229.

Kochanski, G. & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication,* 39, 311–352.

Krakow, R.A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics,* 27, 23-54.

Ladd, D.R. (1984). Declination: A review and some hypothesis. *Phonology Yearbook,* 1, 53-74.

Ladd, D.R. (1996). *Intonational phonology.* Cambridge: Cambridge University Press.

Ladd, D.R., Faulkner, D., Faulkner, H. & Schepman, A. (1999). Constant "segmental anchoring" of F_0 movements under changes in speech rate. *Journal of the Acoustical Society of America,* 106, 1543-1554.

Ladd, D.R., Mennen, I. & Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America,* 107, 2685-2696.

Ladd, D.R. & Silverman, K.E.A. (1984). Vowel intrinsic pitch in connected speech. *Phonetica,* 41, 31-40.

Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages*, Oxford, UK: Blackwell.

Laniran, Y. O. & Clements, G. N. (2003). Downstep and high raising: interacting factors in Yoruba tone production. *Journal of Phonetics, 31*, 203–250.

Laniran, Y. (1992). Intonation in Tone Languages: The phonetic Implementation of Tones in Yorùbá. Dissertation: Cornell University.

Leben, W.R. (1978). The representation of tone. In V. A. Fromkin (ed). *Tone: A linguistic survey.* New York: Academic Press. 177-219.

Lehiste, I. & Peterson, G.E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America,* 33, 419-425.

Lehiste, I. (1975). The phonetic structure of paragraphs. In A. Cohen and S.*E.G.* Nooteboom (eds.). *Structure and process in speech perception.*, New York: Springer-Verlag, 195-206.

Liberman, M. & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle (eds.). *Language Sound Structure.* Cambridge, Massachusetts: M.I.T. Press, 157-233.

Li, Y.J. & Lee, T. (2002). Acoustical F_0 analysis of continuous Cantonese speech. *Proceedings of International Symposium on Chinese Spoken Language Processing 2002*, Taipei, Taiwan. 127-130.

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],* 74-79.

Lin, M. & Yan, J. (1991). Tonal coarticulation patterns in quadrisyllabic words and phrases of Mandarin. *Proceedings of The 12th International Congress of Phonetic Sciences*, Aix-en-Provence, France, 242-245.

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

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

Manfredi, V. (1993). Spreading and downstep: Prosodic government in tone languages. In H. v. d. Hulst and K. Snider (eds.). *The Phonology of Tone.* New York: Mouton de Gruyter. 133-184.

Meeussen, A.E. (1970). Tone typologies for West African Languages. *African Language Studies,* 11, 266-71.

Nakajima, S. & Allen, J.F. (1993). "A study on prosody and discourse structure in cooperative dialogues." *Phonetica,* 50, 197-210.

Ohala, J.J. & Ewan, W.G. (1973). Speed of pitch change. *Journal of the Acoustical Society of America,* 53, 345(A).

Ohala, J.J. (1978). Production of tone. In V. A. Fromkin (ed). *Tone: A linguistic survey.* New York: Academic Press, 5-39.

Ohala, J.J. & Roengpitya, R. (2002). Duration related phase realignment of Thai tones. *Proceedings of 7th International Conference On Spoken Language Processing*, Denver, Colorado, 2285-2288.

Peng, S.-h. (2000). Lexical versus 'phonological' representations of Mandarin Sandhi tones. In M. B. Broe and J. B. Pierrehumbert (eds.). *Papers in Laboratory Phonology V: Acquisition and the Lexicon.* Cambridge: Cambridge University Press.

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

Pike, K.L. (1948). *Tone Languages*, Ann Arbor: University of Michigan Press.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* Dissertation, MIT, Cambridge, MA.

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America,* 70, 985-995.

Pierrehumbert, J. & Beckman, M. (1988). *Japanese Tone Structure*, Cambridge, MA: The MIT Press.

Pierrehumbert, J. & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan and M. E. Pollack (eds.). *Intentions in Communication.* Cambridge, Massachusetts: MIT Press, 271-311.

Prieto, P., Shih, C. & Nibert, H. (1996). Pitch downtrend in Spanish. *Journal of Phonetics,* 24, 445-473.

Rose, P.J. (1988). On the non-equivalence of fundamental frequency and pitch in tonal description. In D. Bradley, E.J.A. Henderson and M. Mazaudon (eds.). *Prosodic Analysis and Asian Linguistics: To Honour R. K. Sprigg.* Canberra: Pacific Linguistics, 55-82.

Rump, H.H. & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech,* 39, 1-17.

Schmidt, R.C., Carello, C. & Turvey, M.T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance,* 16, 227-247.

Shattuck-Hufnagel, S. & Turk, A.E. (1996). A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of psycholinguistic research,* 25, 193.

Shen, X.S. (1990). *The Prosody of Mandarin Chinese.* Berkeley: University of California Press, 1990.

Shen, X.S., (1992). On tone sandhi and tonal coarticulation. *Acta Linguistica Hafniensia,* 24, 131-152.

Shi, B. & Zhang, J. (1987). Vowel intrinsic pitch in Standard Chinese. *Proceedings of The 11th International Congress of Phonetic Sciences*, Tallinn, Estonia, 142-145.

Shih, C. (1986). *The Prosodic Domain of Tone Sandhi in Chinese.* Dissertation. University of California, San Diego.

Shih, C. (1988). Tone and intonation in Mandarin. *Working Papers, Cornell Phonetics Laboratory* 83-109.

Shih, C. (1993). Relative prominence of tonal targets. *Proceedings of The 5th North American Conference on Chinese Lingusitics*, Newark, Delaware: University of Delaware, 36.

Shih, C. (2000). A declination model of Mandarin Chinese. In A. Botinis (ed.). *Intonation: Analysis, Modelling and Technology.* Kluwer Academic Publishers, 243-268.

Shih, C. & Sproat, R. (1992). Variations of the Mandarin rising tone. *Proceedings of The IRCS Workshop on Prosody in Natural Speech No. 92-37*, Philadelphia: The Institute for Research in Cognitive Science, University of Pennsylvania, 193-200.

Stewart, J.M. (1965). *The typology of the Twi tone system*, Legon, Ghana: Institute of African Studies, University of Ghana.

Sundberg, J. (1979). Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics,* 7, 71-79.

Swerts, M. (1997). Prosodic features at discourse boundaries of different length. *Journal of the Acoustical Society of America,* 101, 514-521.

't Hart, J., Collier, R. and Cohen, A. (1990). *A perceptual Study of Intonation — An experimental-phonetic approach to speech melody.* Cambridge, Cambridge University Press.

Titze, I.R. & Talkin, D. (1979). A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *Journal of the Acoustical Society of America,* 66, 60-74.

Umeda, N., (1982). " F_0 declination" is situation dependent. *Journal of Phonetics,* 10, 279-290.

van den Berg, J. (1958). Myo-elastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research,* 1, 227-244.

van Heuven, V.J. (1994). What is the smallest prosodic domain? In P.A. Keating (ed). *Papers in Laboratory Phonology.* 3, Cambridge: CUP, 76-98.

van Santen, J.P.H. & Hirschberg, J. (1994). Segmental effects on timing and height of pitch contours. *Proceedings of The International Conference on Spoken Language Processing*, 719-722.

Wang, A. (2003). *Research on the pitch downtrent of intonation in Putonghua.* Dissertation. Beijing University.

Wang, W.S.-Y. & Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech & Hearing research,* 10, 629-636.

Watson, P.J., Ciccia, A.H. & Weismer, G. (2003). The relation of lung volume initiation to selected acoustic properties of speech. *Journal of the Acoustical Society of America,* 113, 2812-2819.

Whalen, D.H. & Levitt, A.G. (1995). The universality of intrinsic F_0 of vowels. *Journal of Phonetics,* 23, 349-366.

Woo, N. (1969). *Prosody and phonology.* Dissertation, Massachusetts Institute of Technology.

Wu, Z. (1982). Putonghua yuju zhong de shengdiao bianhua [Tonal variations in Mandarin sentences]. *Zhongguo Yuwen [Chinese Linguistics],* 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. Effects of Consonant Aspiration on Mandarin Tones. *Journal of the International Phonetic Association.* (in press).

Xu, Y. (1993). *Contextual Tonal Variation in Mandarin Chinese.* Dissertation. The University of Connecticut.

Xu, Y. (1996). Factors affecting the surface tonal contours of Mandarin. *Proceedings of The 3rd National Conference on Chinese Phonetics*, Beijing, 35-36.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics,* 25, 61-83.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica,* 55, 179-203.

Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica,* 58, 26-52.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F_0 contours. *Journal of Phonetics,* 27, 55-105.

Xu, Y. (2002). Articulatory constraints and tonal alignment. *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France, 91-100.

Xu, Y. & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America,* 111, 1399-1413.

Xu, Y. & Wang, Q.E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication,* 33, 319-337.

Xu, Y. & Xu, C.X. Intonation components in short English statements. (forthcoming):

Xu, Y. & Xu, C.X. (2001). Exploring underlying pitch targets in English statements. *Journal of the Acoustical Society of America,* 110, Pt. 2, 2736-2737.

Xu, Y., Xu, C.X. & Sun, X. (in press). On the Temporal Domain of Focus. To appear in *Proceedings of The 2ndt International Conference on Speech Prosody*, Nara, Japan, March, 2004.

Yuan, J., Shih, C. & Kochanski, G. P. (2002). Comparison of declarative and interrogative intonation in Chinese. *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France, 711-714.

Zee, E. (1980). A spectrographic investigation of Mandarin tone sandhi. *UCLA Working Papers in Phonetics,* 49, 98-116.

Zemlin, W.R. (1988). *Speech and Hearing Science — Anatomy and Physiology*, Englewood Cliffs, New Jersey: Prentice Hall.