LANGUAGE AND LINGUISTICS 5.4:757-797, 2004
2004-0-005-004-000094-1

Understanding Tone from the Perspective of Production and Perception[*]

Yi Xu
Haskins Laboratories

Our understanding of tone can be significantly improved if we take the constraints of speech production and perception seriously. In particular, the maximum speed of pitch change and the coordination of laryngeal and supralaryngeal movements impose certain impassable limits on the way lexical tones are produced. At the same time, although the human perceptual system is highly proficient in processing fast-changing acoustic events as well as resolving distortions due to articulatory constraints, there are limits as to how much undershoot can be perceptually reversed. The understanding of these constraints has led to the Target Approximation model of tone production. The model simulates the generation of F_0 contours as a process of asymptotically approximating underlying pitch targets that are associated with individual tones via language-specific rules. The application of the Target Approximation model helps the understanding of various tone-related issues by providing explicit criteria for distinguishing between tonal variations due to articulatory implementation and those due to alternation of the tonal targets. In light of the Target Approximation model, new insights are offered regarding the nature and distribution of contour tones, distinction between different kinds of sandhi-like tone phenomena, and the target and manner of implementation of the neutral tone.

Key words: tone, tone sandhi, pitch target, target approximation, neutral tone

1. Introduction

While few would doubt that there are physical limits to our speech production system, the act of speaking usually feels so effortless that it is often tempting to believe that in most cases we are safely away from those limits when speaking, and for that matter when producing tones. Assuming that this intuition actually reflects reality, various phonological processes related to tones should then have little to do with articulatory limitations and are probably related more to perception than to production. Over the years, however, a number of phonetic studies have shown that tone production

[*] I would like to thank Carlos Gussenhoven and an anonymous reviewer for their insightful comments. This study was supported in part by NIH grant DC03902.

is subject to certain articulatory constraints (Abramson 1979, Gandour, Potisuk, & Dechongkit 1994, Lin & Yan 1991, Shih & Sproat 1992, Xu 1994). Nevertheless, the role of articulatory constraints is in general still considered to be rather limited as far as tone is concerned. And in tonal phonology there does not seem to be a strong need for treating articulatory constraints as part of the phonological process. In this paper I would like to show that recent advances in experimental phonetics are starting to compel us to take articulatory constraints more seriously than before. In particular, I shall show that the maximum speed of pitch change is slower than has been thought before, and that in tone production speakers often have to get very close to this limit (Xu & Sun 2002). I shall also show that an additional—probably equally strong—constraint on tone production comes from the coordination of laryngeal and supralaryngeal movements (Xu & Wang 2001). Furthermore, I shall show that human perception is actually quite good at handling very fast acoustic events in the speech signals, probably better than previously thought (Janse 2003, Lee 2001). Based on new empirical data, I shall argue that many observed tonal patterns are more closely related to articulatory limits than we had thought before. I shall further demonstrate how the new findings can be incorporated into the Target Approximation model of F_0 contour generation (Xu & Wang 2001). When applied to a number of tonal phenomena, this model may help us gain new insights into their underlying mechanisms.

2. Articulatory constraints

When considering articulatory constraints, what first come to mind are likely static limits such as the highest and lowest pitch values, the highest and lowest jaw positions, etc. Indeed, some of these limits are probably seldom approached in speech, such as the lowest jaw position and the highest pitch. Other static limits, on the other hand, are probably quite often approached, such as the highest jaw position and the lowest pitch. Beside these static limits, however, there are also dynamic limits inherent to the articulatory system that are probably just as important, but have not yet received much attention. In the following I shall consider recent findings about some of the dynamic limits. In particular, I shall discuss the maximum speed of pitch change and the coordination of laryngeal and supralaryngeal movements. I shall also very briefly discuss the maximum speed of other articulatory movements for which new data are just being collected.

2.1 Maximum speed of pitch change

The maximum speed of pitch change has been investigated before (Ohala & Ewan 1973, Sundberg 1979). However, data reported in those studies have often been misunderstood, presumably because of the special form of the data obtained. In order to assess the maximum speed of pitch change in a form that is more directly usable for speech research, we recently did a study revisiting this issue (Xu & Sun 2002). In that study we asked 36 speakers of Mandarin and English to imitate very fast sequences of resynthesized model pitch alternation patterns such as HLHLH or LHLHL, where the H and L differ in pitch by 4, 7, or 12 semitones (1 semitone = 1/12 octave), and the duration of each HL or LH cycle is either 250 or 166.7 ms. Figure 1 shows the waveform and pitch tracking of one of the model pitch alternation patterns used in the experiment. Note that the pitch shifts are nearly instantaneous; although, due to the window size used in implementing the pitch shifts and the smoothing algorithm in F_0 tracking in Praat (www.praat.org), each shift appears to be completed in about 10-20 ms.

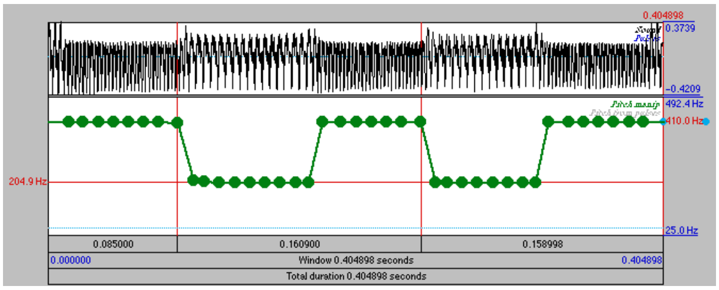

Figure 1: One of the model pitch alternation patterns used in Xu & Sun (2002). The pattern is HLHLH with a pitch range of 12 semitones (1 octave) and a HL cycle duration of 166.7 ms (6 cycles per second).

Figure 2 shows the F_0 track of an actual pitch alternation pattern produced by one of the subjects in Xu & Sun (2002), and an illustration of the measurements used for assessing the maximum speed of pitch change. An immediately apparent characteristic of this pitch pattern is that there are no static high and low regions as in the model pitch alternation patterns. Instead, the transitions all seem gradual and continuous and only peaks and valleys can be seen where H and L are supposed to be. To assess the maximum speed of pitch change, we measured the time interval between adjacent upper and lower turning points as well as the pitch difference between the adjacent turning points. In addition, we also computed the speed of pitch rises and falls by dividing the magnitude of each pitch shift by the time interval between the adjacent turning points.

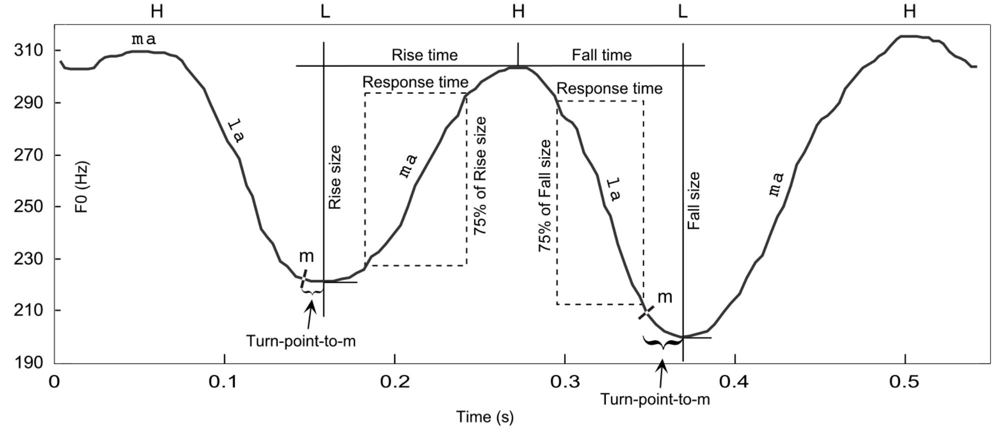

Figure 2: Illustration of measurement of rise and fall excursion time, rise and fall "response time", and turn-point-to-m in a HLHLH trial spoken with /malamalama/ in Xu & Sun (2002). See original paper for more detailed explanations.

Our analyses of the measurements revealed several interesting results. First, the maximum speed of pitch change varied quite linearly with the size of the pitch change: the larger the size, the faster the maximum speed. The relations of maximum speed of pitch rises and falls as a function of pitch change size are represented by the linear equations (1) and (2), respectively,

(1) $s = 10.8 + 5.6\,d$
(2) $s = 8.9 + 6.2\,d$

where s is the average maximum speed of pitch change in semitones per second (st/s), and d is the size of pitch shift in semitone. The importance of these relations is that when considering the speed of pitch change, it is critical to take the magnitude of the change into consideration.

The second result of interest in Xu & Sun (2002) is that the minimum time it takes to complete a pitch change is also related to the magnitude of the change, although the correlation is lower than that between the speed and size of the pitch change. The relations of minimum time of pitch rises and falls as functions of pitch change size are represented by the linear equations in (3) and (4), respectively,

(3) $t = 89.6 + 8.7\,d$
(4) $t = 100.4 + 5.8\,d$

where t is the amount of time (ms) it takes to complete the pitch shift, and d is the size of pitch shift in semitone. The slopes of 8.7 ms/st and 5.8 ms/st mean that the minimum amount of time needed for a pitch change increases rather moderately with the size of pitch change. This agrees with the findings of Ohala & Ewan (1973) and Sundberg (1979). Though not new, this fact has often being overlooked by researchers when considering possible contributions of the maximum speed of pitch change to observed F_0 patterns in speech (e.g., Caspers & van Heuvan 1993, 't Hart et al. 1990, as discussed in detail in Xu & Sun 2002).

The third result of Xu & Sun (2002) is that, compared to the maximum speed of pitch change obtained in the study, the speed of pitch change in real speech as reported in previous studies were similar in many cases. For example, with equations (1) and (2) it can be computed that when the pitch shift size is 6 st, the average maximum speed of pitch rise is 44.4 st/s and that of pitch fall is 46.1 st/s. This is comparable to the 50 st/s at 6 st as reported by 't Hart et al. (1990). Also, the fastest speed of pitch change reported by Caspers & van Heuven (1993) was comparable to the maximum speed of pitch change at similar pitch shift intervals as computed with equations (1) and (2). The maximum speed of pitch change reported by Xu & Sun (2002) also matched the speed of pitch change in the dynamic tones (R and F) in Mandarin as recorded in Xu (1999) (but not in the static tones, i.e., H and L). For English, 't Hart et al. (1990) report that full-size rises and falls can span an octave and the rate of change can reach 75 st/s. This is comparable to the maximum mean excursion speed of 78 st/s and 83 st/s for 12-st rises and falls computed with (1) and (2). These comparisons indicate that on many occasions, the fastest speed of pitch change is indeed approached in speech.

The fourth result of Xu & Sun (2002) relevant to this paper is that, in terms of the maximum speed of pitch change, there are no overall differences between native speakers of American English and native speakers of Mandarin Chinese. This was true with data from 16 English speakers and 20 Chinese speakers. It indicates that, as different as English and Chinese can be in terms of their linguistic uses of F_0 contours, being native speakers of either language did not result in significant physiological differences as far as the speed of pitch movement is concerned. So, unless there are new data showing clear evidence to indicate otherwise, we can assume that the maximum speed of pitch change obtained in Xu & Sun (2002) is applicable to languages in general. On the other hand, however, we did find differences across individuals in both Mandarin and English subjects. For example, the standard deviation for the maximum speed of raising pitch by 12 st is 14.2 st/s and that of lowering pitch by the same amount is 15.8 st/s (cf. Table V in Xu & Sun 2002). So, although the discussion in this paper will mostly refer to the average maximum speed of pitch change, one should keep in mind that individual speakers are either faster or slower than the averages.

Coming back to the example shown in Figure 2, it is now apparent why no pitch plateaux are formed when the speaker tries to imitate the model pitch alternation patterns consisting of static H and L. It is because those patterns are much faster than the maximum speed of pitch change the speaker can produce. Even the fastest subject in our experiment took 91 ms to complete a 4 st pitch rise (Table V, Xu & Sun 2002), longer than the 166.7/2 = 83 ms in the model pitch alternation pattern. What this means is that when a speaker tries to complete a pitch change in a time interval shorter or equal to the physically allowed minimum time, the F_0 contour is inevitably continuous and free of consistent steady states.

With data provided by equations (1-4), we can conveniently examine real speech data to see whether and when the maximum speed of pitch change is approached and how much of the observed pitch variations may be attributed to this constraint. For example, in recent studies on contextual tonal variations in Mandarin (Xu 1997, 1999), it was found that the F_0 contour of a tone varies closely with the offset F_0 of the preceding tone. H (High) in Mandarin, for instance, is found to be produced with an apparently rising contour when following L (Low), as shown in Figure 3. Likewise, L is produced with an apparently falling contour when preceded by H. Equations (3) and (4) provide a first-order account for the observed F_0 contours. According to (3) and (4), if the pitch range for a tone is 6 semitone, it would take an average speaker at least 142 ms to complete a pitch rise and 135 ms to complete a pitch fall. This means that in a syllable with an average duration of 180-186 ms (Xu 1997, 1999), the greater half of the F_0 contour would have to be used for completing the pitch movement from L to H or from H to L, even if the maximum speed of pitch change is used. The long transitions observed in H L and L H and other tone sequences in Xu (1997, 1999), as seen in Figure 3, therefore should mostly be attributed to the constraint of maximum speed of pitch change.

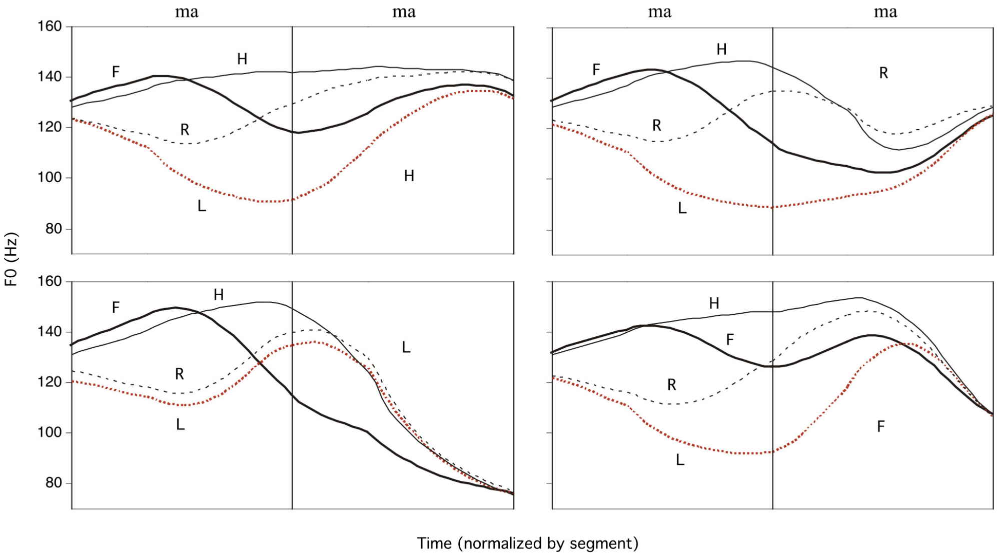

Figure 3: Effects of preceding tone on the F_0 contour of the following tone in Mandarin. In each panel, the tone of the second syllable is held constant, while the tone of the first syllable is either H, R, L, or F. The vertical lines indicate the onsets of syllable-initial nasals. Each curve is a (segment-by-segment) time-normalized average of 192 tokens produced by eight speakers. (Adapted from Xu 1997)

2.2 Coordination of laryngeal and supralaryngeal movements

As is widely known, in many tone languages of East Asia and Africa, there is lexically a one-to-one association between tone and syllable, i.e., each monosyllabic word or morpheme is associated with a tone. Despite this lexical association, however, conceptually the two do not have to be perfectly aligned with each other. E.g., for Mandarin some phonetic/phonological accounts have in fact suggested various micro-adjustment schemes for tone-syllable alignment (Howie 1974, Lin 1995, Rose 1988, Shih 1988). Furthermore, as the previous discussion of maximum speed of pitch change indicates, sometimes it is actually quite hard to implement a tone. In a tone sequence such as LH, HL, LF, and HR, it is imaginable that speakers would want to readjust the micro-alignment of a tone to make the transition between two adjacent tones easier. As found in Xu (1999), however, speakers do not seem to do that. Figure 4 shows the mean F_0 curves of the tone sequences H*x*FHH, where *x* stands for any of the four Mandarin tones (H, R, L, or F), averaged across five repetitions produced by four male speakers. The sentences in Figure 4(a) carry no narrow focus, while those in Figure 4(b) carry a narrow focus on the third syllable. One of the most striking things about the F_0 contours of F in Figure 4 is that, regardless of the tone of the second syllable, the F_0 contour corresponding to F in the third syllable always starts to climb sharply right after the syllable onset. The slope of the climb differs depending on the ending F_0 of the tone in syllable 2. It is the steepest after L and the shallowest after H. This difference in slope, however, does not seem to be enough to fully compensate for the differences among the four tones in syllable 2 in terms of their offset F_0. As a result, the peak F_0 of F is much lower after L than after H. In fact, in Figure 4(a) at least, the height of the peak F_0 of F seems proportional to the offset F_0 of syllable 2. The fact that F_0 goes up even after H (in syllable 2), which has the highest offset F_0, suggests that the initial pitch targeted for F is quite high. This means that the peak F_0 of F after L (and in fact after R and F as well) is a clear compromise, or undershoot. Assuming that such a compromise is not really desirable, for which I shall show some evidence later in 3.1, it is conceivable that speakers could have readjusted the tone-syllable alignment so that F would have more time to attain its target.[1] Such a readjustment would make sense especially when F is under focus and the preceding tone thus would be less important and its time slot vulnerable to encroachment by the surrounding tones. The fact that no such alignment readjustment seems to have occurred, as seen in Figure 4(b), suggests that there must be some kind of strong constraint that has prevented the readjustment from happening.[2]

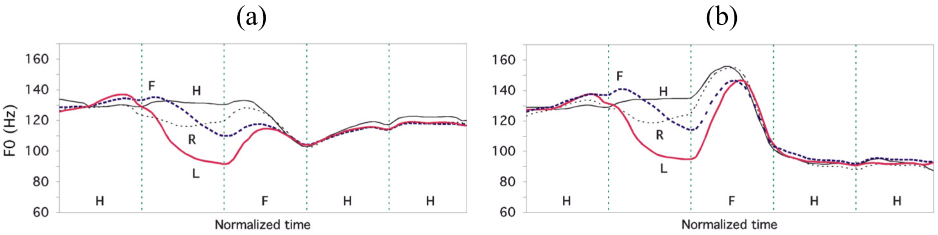

Figure 4: Mandarin tone F following four different tones. (a): no narrow focus in the sentence; (b): focus on the F-carrying syllable. Each curve is an average of 20 tokens produced by four male speakers (five repetitions per speaker). (Data from Xu 1999)

[1] Here and throughout the paper, a target refers to an intended goal rather than an actually realized value such as a peak or valley, as is typically used in the intonation literature; e.g., in Pierrehumbert 1980.

[2] It is reported by Atterer & Ladd (2004) that in English, Greek, Northern German, and Southern German, F_0 rises have variant onset times relative to the syllable onset. These variations, however, are not described by the authors as triggered by specific tonal contexts. So, while certainly worth further investigation, their data do not seem to constitute evidence that individual speakers can micro-adjust target-syllable alignment according to specific tonal contexts or in reaction to changed time pressure for laryngeal movements.

This constraint, I would like to suggest, probably stems from the fact that tone articulation and syllable articulation are concurrent movements controlled by a single central nervous system. To carry out any concurrent movements, as found by studies on limb movements, performers have very few choices in terms of the phase relation between the movements (Kelso et al. 1981, Kelso 1984, Schmidt, Carello & Turvey 1990). At relatively low speed, the phase angle between two movements has to be either 180º, i.e., starting one movement as the other is half way through its cycle, or 0º, i.e., starting and ending the two movements simultaneously. At high speed, however, only the 0º phase angle is possible. 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 124 ms to complete a 4-st rise or fall, and about 107 ms to complete a 2-st rise and 112 ms to complete 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 the speaker to maintain a 180º phase angle between pitch movement and the syllable, assuming that the syllable functions as a cyclic coordinate structure under which both supralaryngeal and laryngeal units are organized. Other odd phase angles would be even less likely based on the findings of Kelso et al. (1981), Kelso (1984), and Schmidt et al. (1990). Therefore, the only likely choice left to the speaker is to maintain a constant 0º phase angle between pitch movement and the syllable, i.e., keeping them fully synchronized.

2.3 Maximum speed of other articulatory movements

The recognition of articulatory limits on the maximum speed of pitch change naturally raises questions about whether there are limits on the maximum speed of movements for other articulators such as the lips, the tongue, the jaw, the velum, etc. We are currently conducting a study to investigate the speed of repetitive movements that are used in speech, involving the lips, the tongue, and the jaw (Xu et al. in progress). Subjects are asked to do two different tasks. In the non-speech task, similar to what we did in Xu & Sun (2002), we ask the subjects to imitate model syllable sequences such as /bababababa/, /mamamamama/, and /wawawawawa/ that are naturally produced but then resynthesized with rate increased to 9 syllables per second. In the speech condition, subjects were asked to read aloud sentences containing CVC strings such as /bab/, /mam/, and /waw/, both at the normal rate and as fast as possible but without slurring. Preliminary analyses of data from three subjects show that the minimum syllable duration is very similar in the speech and non-speech tasks: ranging from 110 to 140 ms.

A syllable cycle consists of two phases—onset and offset—corresponding to lip closing and opening or tongue raising and lowering. In other words, they correspond to C and V, respectively. If the minimum syllable duration is 120-130, each C or V phase may take about half of the syllable cycle, i.e., 65-70 ms. (We are also taking separate measurement of the onset and offset phases. But the analysis is still underway.) In pitch movement, as indicated by equations (3) and (4), each movement in one direction takes at least 124 ms if the magnitude of movement is 4 semitones. This is twice as long as a C phase or a V phase, but roughly as long as the shortest symmetrical syllable, e.g., bab, mam, etc. (Asymmetrical syllables could be even shorter, since there is less direct conflict in the same articulator). This observation, pending further confirmation when data collection is completed, provides part of the basis for the phase relation schematic to be discussed later in 4.1.

3. Perceptual processing of tonal variations due to articulatory constraints

There have been many studies on the perception of pitch, pitch glide, and tone. Studies that look into the human perceptual limit on pitch processing often tend to find the human perception system quite sensitive and accurate about pitch events (Klatt 1973). However, some studies reported lower sensitivity to pitch difference and pitch change (Greenberg & Zee 1979, Harris & Umeda 1987, 't Hart 1981). These studies, however, often use non-speech tasks such as judging whether the pitch of two stimulus sentences are the same, or whether the magnitude of pitch change rate is the same ('t Hart 1981), or judging the contouricity of a pitch pattern (Greenberg & Zee 1979). In regard to tone perception, what we should be concerned with is how effectively a tone can be identified. In particular, we want to know the limit beyond which the perception system is no longer able to factor out variations due to the articulatory constraints such as those discussed earlier. In the following, I shall discuss three studies that show that the human perceptual system is actually quite remarkable in its ability to handle constraint-caused variations. At the same time, however, there seem to be limits beyond which the variations can no longer be handled effectively by the perceptual system.

3.1 Perception of tones with undershoot

In tone languages, lexical tones function to distinguish words that are otherwise phonologically identical. Because of this, there is presumably pressure for them to remain distinct from each other as much as possible when being produced in speech. However, as discussed earlier, the two types of articulatory constraints—maximum speed of pitch change and coordination of laryngeal and supralaryngeal movements—

introduce substantial variability into the surface form of the tones. At times, the variations are so extensive that one tone may appear to resemble a different tone. When this happens, questions may arise as to whether the tone has actually changed its identity. A case in point is R in Mandarin. According to Chao (1968), in fluent speech, this tone may change into H when it is in a three-syllable word in which the tone of the first syllable is H or R. For example,

[ts^{h}ōŋ jóu pǐŋ] (H R L) → [ts^{h}ōŋ jōu pǐŋ] (H H L) ‘green onion pancake’

This description of the tonal variation implies that the identity of R is changed in this environment, which means that listeners should hear the tone of [jóu] as identical to H. To see if R has indeed become indistinct from H in this case, a study was done to investigate both the acoustics and perception of R in this kind of tonal environment (Xu 1994). In the study F_0 contours of R produced in different tonal contexts were first examined. It was found that in H R L: [¯ / _], referred to as a “conflicting” condition, the contour of R indeed became flattened: [¯ – _], as opposed to that in L R H: [_ / ¯] where the contour remained rising: [_ / ¯]. When presented to listeners for identification with the surrounding tones removed, R was heard most of the time as H if it had originally been in H R L, but as R when it had been originally in L R H. When the tonal context was kept intact, but the lexical identity of the original words were neutralized by replacing the initial consonant of the first and/or the last syllable, listeners were able to correctly identify R most of the time. This demonstrates that the perceptual system is largely able to factor out the effect of tonal context when identifying a tone that deviates from its canonical form due to tonal context.

The perceptual experiments in Xu (1994) also yielded another interesting result. When the tonal contexts were kept intact, although the identification rates were high for all tonal contexts, there was a significant difference between contexts which caused extensive distortion in the tone of the middle syllable, such as in H R L: [¯ / _], and those that did not cause much distortion, such as in L R H: [_ / ¯]. Tones that had undergone extensive distortions had significantly lower identification rate (88%) than tones with minimal distortion (96%). The fact that articulatory constraints did take a toll on the perception of R in the case of H R L and the like indicates that there is a perceptual limit as to how much of the contextual effect can be successfully factored out.

The recognition of the limit on perceptually resolving contextual variations is important. The existence of this limit means that there is an actual perceptual pressure for reducing contextual distortion of the tonal contours to a minimum. Relating this to the earlier discussion of articulatory constraints, we can see that tonal variations due to articulatory constraints such as maximum speed of pitch change and coordination of laryngeal and supralaryngeal movements are not really *perceptually desirable*, but rather *articulatorily unavoidable*. In other words, speakers probably did not intend to produce the variant forms of the affected tones: they simply could not avoid producing them. Similar evidence has been found in segmental production (cf. 3.3).

3.2 Perception of initial fragments of tones

The findings of Xu (1994) demonstrate the importance of the information provided by the tonal context for the recognition of tones that have deviated extensively from their canonical forms due to the limit of maximum speed of pitch change. The importance and the usefulness of this information is further demonstrated by a recent study on the online perceptual processing of tone. Lee (2001) did a gating experiment in which subjects listened to fragments (in 20 ms increments) of a target syllable in Mandarin, which is the last syllable embedded in a carrier sentence "zhège zì shì __" [This character is __]. What he found was that listeners could correctly recognize a Mandarin tone well before the entire F_0 contour of the tone was heard. Figure 5 is a schematic representation of the findings of this gating experiment. The time locations indicated by the arrows in Figure 5 are taken from Lee (2001), but the F_0 contours are adapted from Xu (1997). The first syllable in the graph carries F, which is similar to the tone of the syllable (shì) before the target syllable in Lee's experiment. Among the main findings of the experiment are (a) most subjects are able to correctly identify whether the tone of the target syllable bears a high- or low-onset tone (H, F vs. R, L) with 20-40 ms of F_0 input from the target syllable (lower right arrow); (b) in cases where tones have potentially similar onset pitch—H vs. F, R vs. L, subjects can correctly identify them about 70 ms after the voice onset, as indicated by the upper arrow; and (c) when the onset of the target is a sonorant (m, n, l), tones can be identified even before any portion of the vowel is heard (lower left arrow).

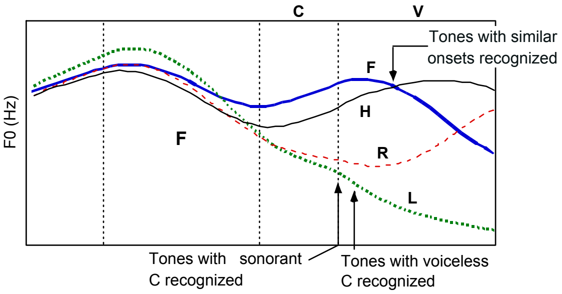

Figure 5: Schematic representation of the findings of Lee (2001). The tonal contours are adapted from Xu (1997). The arrows point to the moments when tones start to be correctly perceived.

Lee's (2001) findings would seem puzzling were we literally to believe that tones are carried only by vowel or syllable rime. It would also be puzzling were we to believe that the entire F_0 contour is needed for the perceptual recognition of a tone. However, Lee's findings would make sense if we assume that listeners are constantly looking for information that can eliminate competing candidates. As can be seen in Figure 5, by the time of the lower left arrow, the F_0 curves corresponding to all four tones become separate, or "unique," provided that the preceding tonal context is known. Of course, Lee's (2001) findings should not be interpreted as suggesting that the F_0 contour carried by the later portion of the syllable is irrelevant. Rather, assuming that the underlying pitch target associated with a tone is synchronously implemented with the syllable, what Lee's subjects heard before the vowel, or in the 20 ms worth of the vowel when the initial C was an obstruent, were transitions from the end of the previous tone to the pitch target of the present tone. In other words, the differences in contours corresponding to the later portion of the syllable are determined by the underlying tonal target, and these give rise to the differences in the initial transitions. Listeners seem to know this, because they make surprisingly effective use of this information as is clearly demonstrated by Lee's (2001) data.

3.3 Perception of human-produced fast speech and time-compressed resynthesized speech

What Xu (1994) and Lee (2001) have demonstrated is listeners' remarkable ability to make full use of the tonal information not only from F_0 contours that closely resemble the underlying tonal targets, but also from F_0 transitions toward those targets. If we contrast this ability with what is found by Janse (2003), we can see an even clearer picture of the characteristics of the human speech perception system. Janse (2003) reports a series of experiments in which she examined Dutch speakers' ability to speed up speech and Dutch listeners' ability to understand natural fast speech as well as computer-speeded up normal speech. She found that when speakers tried to speak as fast as possible, they speeded up from a normal rate of 6.7 syll./sec to only about 10 syll./sec. That is, they could not even double their normal articulation rate. At the same time, however, the intelligibility of such natural fast speech was much reduced. In contrast, when normal speech was linearly speeded up through computer resynthesis (using the PSOLA algorithm) to almost three times the original rate, intelligibility remained very high. These findings indicate that the human perceptual system is highly capable of processing very rapidly changing acoustic events, as in the computer-speeded-up speech. On the other hand, human perception is apparently less able to handle excessive undershoot presumably due to the maximum speed of various articulators, as in the "slurred" natural fast speech produced by the Dutch speakers in Janse (2003) as well as in the flattened tones produced (at normal speaking rate) by Mandarin speakers in Xu (1994).

4. The target approximation model of tonal contour formation

What the foregoing discussion has demonstrated is that there are different forces working from different directions and interacting with each other during tone production. Following Xu (2001a), these forces can be divided into two major categories—voluntary and involuntary. Voluntary forces originate from communicative demands, while involuntary forces originate from articulatory constraints. Communicative demands correspond to linguistic and paralinguistic information that needs to be conveyed during speech communication. In tone languages, tones serve to distinguish words or to indicate certain syntactic functions. The conveyance of their identities to the listener is thus part of the communicative demands. Tonal identities are presumably represented by their canonical/underlying forms. To realize these forms, speakers employ their articulators which, as a part of a physical system, have various inherent limitations. These limitations constitute involuntary forces that are orthogonal to the voluntary forces. The interaction between the voluntary and involuntary forces could bring about robust variations in the F_0 contours of tones, as discussed in the previous sections. It is of course not the case that no one is aware of this interaction. It is just that until the recent study on the maximum speed of pitch change (Xu & Sun 2002), the general consensus had been that speakers are usually so far away from any potential physical limits that there is no need to always keep them in mind when trying to understand observed tonal contours. The finding that it often takes almost the entire duration of a syllable to complete a pitch shift the size of only a few semitones compels us to take physical limits really seriously. The recognition of the interaction between voluntary and involuntary forces also makes it possible to model F_0 contour generation more naturally and more accurately. In the following discussion, I shall briefly sketch the Target Approximation model that is based on the new understanding of the interaction between voluntary and involuntary forces. The model was first outlined in Xu & Wang (1997, 2001). A quantitative implementation of it was attempted in Xu, Xu, & Luo (1999).

4.1 The model

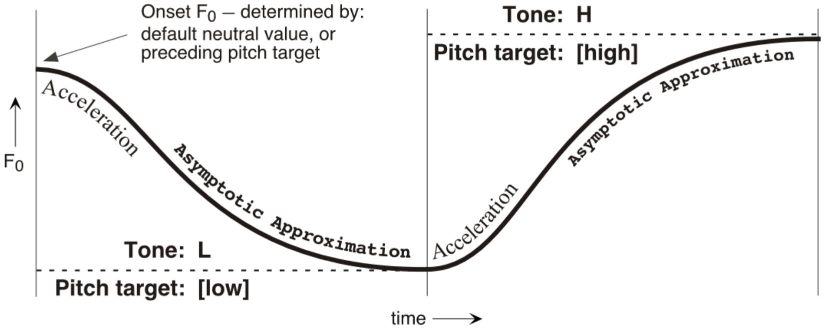

Figure 6: A schematic sketch of the Target Approximation model. The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from articulatory implementation of the pitch targets.

The basic operation of the model is schematically sketched in Figure 6. At the core of the model is the assumption that phonological tone categories are not directly mapped onto surface phonetic patterns; rather, each tone is associated with an ideal pitch target (or sometimes more than one, cf. discussion in 5.2) that is articulatorily operable. Each pitch target has a simple form such as [high], [low], [mid], [rise], or [fall]. The process of realizing each tone is to approximate articulatorily the shape of its associated pitch target. The implementation is done, however, under various articulatory constraints. The first is the coordination of laryngeal and supralaryngeal movements discussed in 2.2, which requires that the implementation of the tone and the syllable (which may consist of various C and V combinations) be fully in phase. Figure 7 shows schematics of this phase relation in a sequence of two CV syllables carrying a tone series of H L. Figure 7(a) shows the sequence said at a fast rate. As discussed in 2.3, articulatory movements associated with C and V are likely to be quite fast. It is therefore possible to fit the C and V cycles consecutively in a syllable cycle even at a fast rate (as indicated by the dotted curve on the bottom in Figure 7), which still satisfies the synchronization requirement. At a slower rate, V is often longer than C (as shown in Figure 7(b)). Nevertheless, the two are still implemented sequentially within the syllable cycle. Note that the depiction of the temporal relations between C and V in Figure 7 does not include their overlap with one another, which has been demonstrated in many studies (Fowler 1984, Fowler & Smith 1986, for example). There has been evidence, however, that if the adjacent CV involves the same articulator, they are actually executed sequentially with no overlap (Bell-Berti 1993, Xu & Liu 2002, Liu & Xu 2003). In cases where different articulators are involved, there has also been evidence that the overlap is “time-locked” and the sequential order is maintained (Bell-Berti & Harris 1982, Bell-Berti 1993).

The fitting of tones into this phase relation needs some explaining. It has been argued that only vowels or rhymes are tone-bearing units (Howie 1974, Lin 1995). However, when C is voiceless, during which the vocal folds are not vibrating, the laryngeal muscles that control pitch movements do not have to be deactivated. And according to Mandarin data from Xu (1997, 1998, 1999) and Xu, Xu, & Sun (2002), as will be discussed in 5.1.2 in greater detail, their activities continue during C whether or not there is voicing. In other words, the tonal target is synchronously implemented with the syllable with no special adjustment for any variation in the syllable-internal CV phase relations. The phase relation depicted in Figure 7 therefore shows that tone cycles simply coincide with syllable cycles.

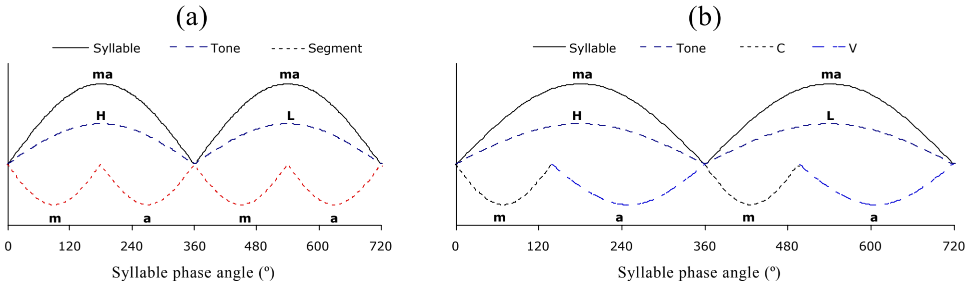

Figure 7: Schematics of synchronized phase relation among syllable, tone, consonant (C), and vowel (V): (a) fast rate, (b) slower rate.

A synchronized phase relation means that the approximation of a target does not start until its cycle begins. In other words, as suggested by Figure 7, the first segment of the syllable and the tone associated with the syllable both start their movement toward their respective targets at phase angle 0º. Synchronization also means that the implementation of the pitch target as well as that of the last segment of the syllable ends at phase angle 360º or 720º, where the syllable ends. Because every articulatory movement takes time, depending on the allotted time interval by the synchronized phase relation, each target may or may not be fully attained by the end of its cycle. As a result, various scenarios may occur:

1. There is plenty of time for a target, e.g., when a monophthong vowel is well over 100 ms long, or a static tone is well over 200 ms long—The target value would be attained before the end of its allotted time; and a pseudo steady state at that value might be achieved in the case of static tone or vowel. An illustration of this scenario is shown in Figure 8(a).

2. A target is given just enough time, e.g., when a monophthong vowel is 75 ms long, or a static tone is 150 ms long—Within the allotted time interval, the movement toward the target would proceed continuously, and the target value would not be achieved until the end of its allotted time interval. An illustration of this scenario is shown in Figure 8(b).
3. A target is given insufficient amount of time, e.g., a vowel is much less than 75 ms long, or a static tone is much less than 150 ms long—Within the allotted time interval, again the movement toward the target would proceed continuously; but the target would not be approached even by the end of its allotted time interval. By the time its cycle ends, however, the implementation of the target has to terminate and the implementation of the next target has to start, as is illustrated in Figure 8(c).

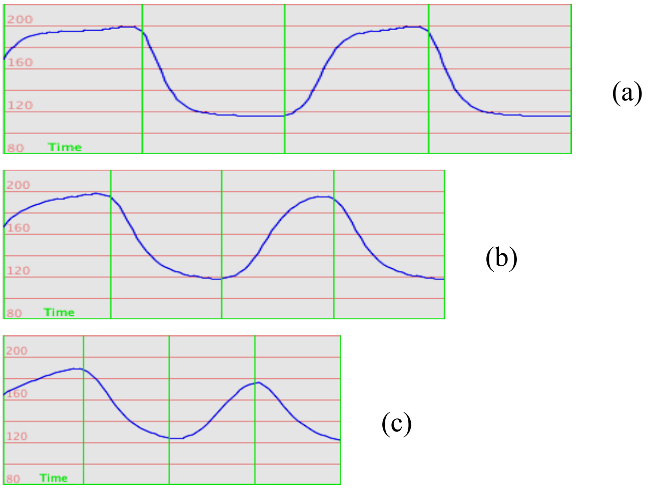

Figure 8: Degrees of attaining the targets in a [high] [low] [high] [low] sequence, given different amount of allotted time. (a) Excessive time: full target attainment with sustained pseudo steady state. (b) Just sufficient time: virtually full target attainment but without steady state. (c) Insufficient time: incomplete target attainment.

As will be discussed later, all these scenarios seem to occur quite frequently in speech in any given language. In addition to the coordination and speed constraints, Figure 8 also illustrates a more subtle constraint, i.e., due to inertia it takes time for an articulatory movement to accelerate to full speed. For example, the F_0 drop from the initial high F_0 in syllable 2 and the rise from the initial low F_0 in syllable 3 both take some time to accelerate to full speed, resulting in a convex-up shape in the early portion of syllable 2 and a concave-down shape in the early portion of syllable 3.

4.2 Types of pitch targets

There is no requirement in the Target Approximation model that all pitch targets be static and mono-valued, such as [high], [low], or [mid]. For some tones, we have noted that static targets cannot produce pitch contours similar to the observed ones. There also seem to be targets that are dynamic, such as [rise] and [fall] (Xu 1998, 2001b). For a dynamic target, the movement itself is the goal. A dynamic [rise] and its implementation are illustrated in Figure 9 together with a static [low]. The slanting dashed line in syllable 1 represents the target [rise] associated with, say, R in Mandarin. The level dashed line in syllable 2 represents the target [low] associated with, say, Mandarin L. The solid curve represents the F_0 contour resulting from implementing the pitch targets under articulatory constraints. Note that the slanting line is not drawn to be aligned with the entire syllable 1. This is because the target is only associated with the syllable and the synchronization demand I have been talking about is not for the alignment of the underlying targets themselves. What has to be synchronized is only the articulatory implementation of the target. Note also that the approximation of [rise] results in a fast rising movement at the end of syllable 1. Although the implementation of [low] in the second syllable starts at the syllable onset, the deceleration of the rising movement and acceleration toward the low F_0 both take some time. As a consequence, an F_0 peak occurs in the initial portion of syllable 2. In this model, therefore, this seeming delay of the F_0 peak often seen in connection with R (Xu 1997, 1998, 1999) results directly from implementing a [rise] when followed by a [low] or another target also with a low onset. No underlying delay of the tone relative to the syllable needs to be assumed.

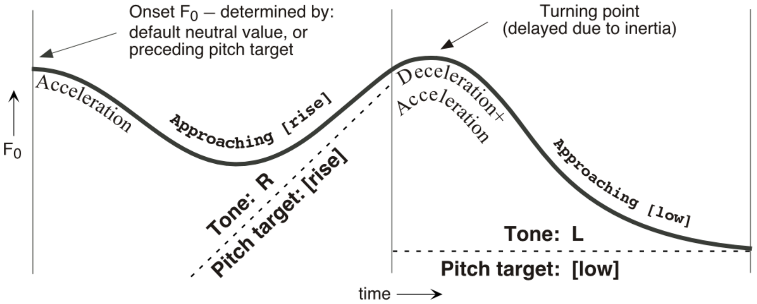

Figure 9: Dynamic and static targets and their implementation. The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from asymptotic approximation of the pitch targets.

Besides the simple pitch targets discussed so far there can also be other more complicated targets. And it is also possible to assign two pitch targets to a tone and even two tones to a syllable. But each target type needs to be independently justified, and it needs to be shown that its implementation is articulatorily possible. One possible candidate of a tone having two pitch targets is the Beijing Mandarin L uttered in isolation or in a pre-pausal position. An isolated or pre-pausal syllable is often long enough to allow for the implementation of two targets, and the F_0 contour of a long L seems to suggest that there is probably either a static [mid] or [mid-high] or a dynamic [rise] following the early [low]. No acoustic investigation that I am aware of has been designed to address this possibility, however. It thus awaits future research to determine the pitch target composition of the long L in Mandarin.

5. Implications

It could be argued that the Target Approximation model just sketched is too simplistic, because it seems to attempt to attribute almost everything to phonetics, ignoring the fact that many complicated tonal phenomena are language specific, and thus cannot be reduced to only a few physical factors. What I shall show in the rest of the paper is that to understand tones, language specific factors undoubtedly need to be considered. But they need to be considered along with the physical factors that I have been discussing. And, as I shall show, many tonal phenomena are better understood as resulting from the speaker's effort to implement simple pitch targets, whose association with lexical tones is language-specific, under various articulatory constraints that are universal.

5.1 The nature and distribution of contour tones

Many tone languages are known to have contours tones. But the nature of contour tones is far from clear. Both the underlying forms of contour tones and the conditions under which contour tones can occur are still being vigorously investigated. In the following I shall try to apply the Target Approximation model to the case of contour tones and see if it can shed some new light on this old conundrum.

5.1.1 Two consecutive static elements or a single dynamic element

There is a longstanding debate over whether a phonetically dynamic F_0 contour corresponding to an intonational component is composed of a single contour component, such as rising or falling, or successive level elements such as HL or LH. The more traditional views of intonation describe intonation patterns as consisting of both pitch levels and pitch contours such as rises and falls (Bolinger 1951, O'Connor & Arnold 1961, 't Hart & Cohen 1973, Ladd 1978). In contrast, many later intonation researchers argue that simple level elements such as H and L are the most basic components of intonation, and dynamic contours such as rise and fall are derived from concatenated static pitch targets, i.e., rise = LH, and fall = HL. The most fully developed of such theories are presented by Pierrehumbert (1980), Liberman & Pierrehumbert (1984), and Pierrehumbert & Beckman (1988). Similar debate has been going on concerning lexical tones. While some researchers (e.g., Pike 1948, Wang 1967, Abramson 1978) argue that contour tones found in languages such as Thai or Mandarin should be considered as single units, while others (e.g. Duanmu 1994b, Gandour 1974, Leben 1973, Woo 1969) treat contour tones as sequences of H and L.

As anyone who has ever worked on acoustic analysis of tones will have noticed, contours are almost ubiquitous in the F_0 tracings of a tone language, whether the basic form of the tones in question is considered to be phonologically level or dynamic. This seems to make it extremely hard to determine whether an observed contour is underlyingly also a contour or in fact consisting of level elements. The interaction between voluntary and involuntary forces that I have been discussing indicates that it is imperative that we take articulatory constraints into consideration when trying to understand any observed acoustic pattern in speech. To reexamine the issue of contour vs. level, I would therefore like to apply the Target Approximation model, because it incorporates the major articulatory constraints relevant for F_0 contour production.

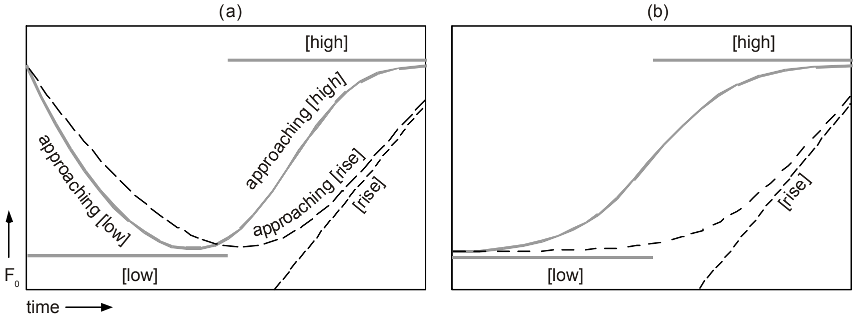

Figure 10: Hypothetical F_0 trajectories (curved lines) that asymptotically approach either two consecutive static targets (solid lines) or a single dynamic target (dashed lines). In (a) the previous tone ends high, while in (b) the previous tone ends low.

According to the model, if we assume that a particular tone consists of two elements, then each of them is a target in its own right. (Note that the in-phase requirement would not prevent two static pitch targets from being fitted into one syllable, as discussed earlier.) When there is sufficient time assigned to the consecutive targets, there should be a clear transition between the two elements. Sufficient time means the amount of time over and beyond the minimum time needed for making a complete pitch shift according to equations (3) and (4). So, when there is sufficient time, an LH combination should look like the solid curve in Figure 10(a), assuming that the previous tone happens to end high. If, however, the target itself is dynamic, like the dashed slanting line in Figure 10(a), even when there is sufficient time, the curve resulting from implementing this dynamic target should have the shape of the dashed curve. If the previous tone happens to end low, then the curves resulting from implementing two static targets or one dynamic target should look like the solid or dashed curve in Figure 10(b), respectively. As shown in Figure 11, in Mandarin, when the tone sequence of L H is produced at a slow rate, we can see that the curve corresponding to the L H portion (i.e., left of the short vertical line) indeed resembles the solid curve in Figure 10(a).

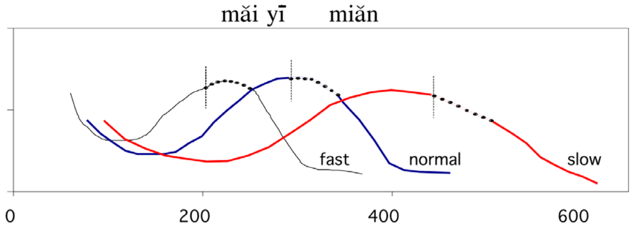

Figure 11: Mean F_0 contours a L H L sequence in Mandarin spoken at fast, normal and slow rate. The gap between time 0 and the beginning of each curve corresponds to the mean duration of the initial consonant in syllable 1. The dotted region in each curve corresponds to the initial nasal of syllable 3. The short vertical lines indicate the onset of the initial nasal of syllable 3. (Adapted from Xu 2001b)

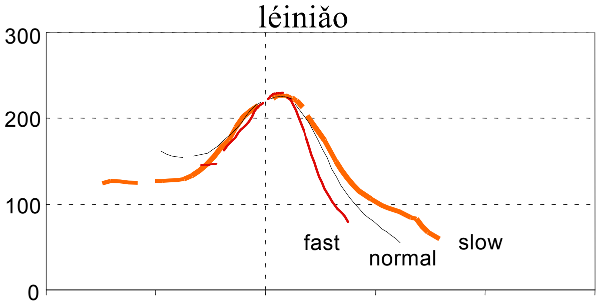

Figure 12: Mandarin tone sequence R L spoken at three different rates. The curves are aligned to the onset of [n] in the second syllable as marked by the vertical line. Each curve is an average of five repetitions. Data from one subject in Xu (1998)

In the case of R in Mandarin, in contrast, as found in Xu (1998), when spoken at a slow rate, the shape of its F_0 curve does not resemble the solid curve in either Figure 10(a) or Figure 10(b). Rather, it is typically like the thick curve in Figure 12, which is more like the dashed curve in Figure 10(b) (or 10(a) for some subjects, when they end the previous tone high). The regression analyses in Xu (1998) show that the most dynamic portion of the rising contour (i.e., velocity peak) in R moves increasingly into the later portion of the syllable as the duration of the syllable increases, as shown in Figure 13, and that the maximum velocity of the rise does not change systematically with variations of syllable duration. These results demonstrate that the rising contour as a whole shifts more and more into the later portion of the syllable without systematic changes in the slope of the rise as the syllable duration increases. This can be interpreted as evidence that a coherent rising contour is being implemented.

It thus seems that, whatever the underlying phonemic representation they may have, tones like R and F in Mandarin are associated with dynamic [rise] and [fall] as the articulatorily executable pitch targets. On the other hand, of course, existence of dynamic tonal targets in one language does not mean that contour-like tones in any language are underlyingly dynamic. It is entirely possible, as far as articulation is concerned, to have tones that are composed of consecutive static targets. Evidence for such tones can be obtained using similar experimental methods that have been used on Mandarin tones (Xu 1998, 2001b). Similar methods can be also used in determining the dynamic/static nature of pitch targets that are associated with various intonation components in non-tone languages such as English.

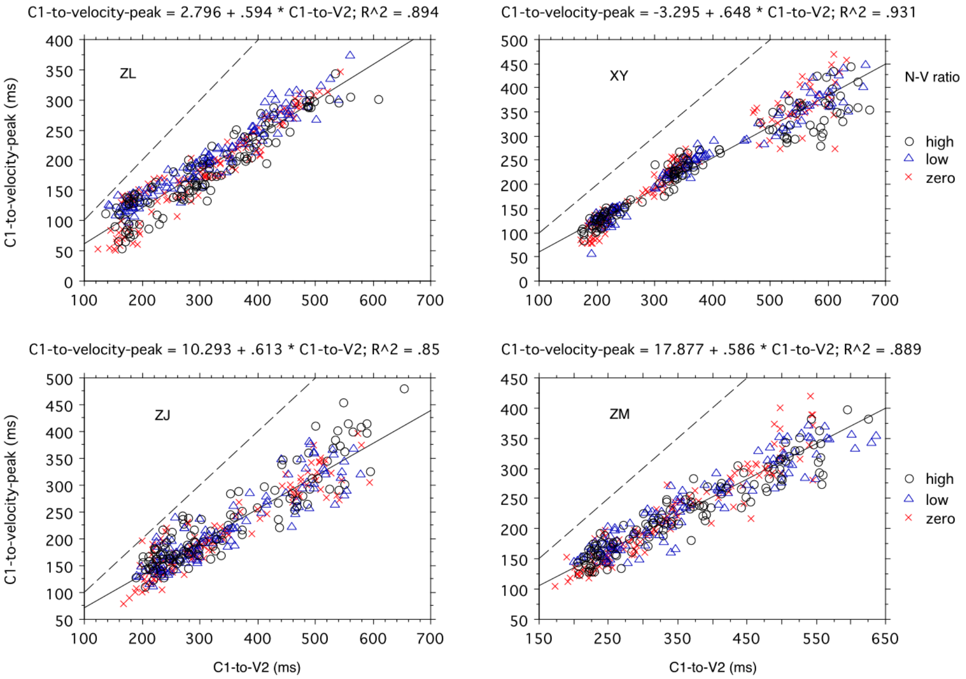

Figure 13: Location of peak velocity of F0 rise relative to the onset of syllable 1 (y-axes) plotted against the distance between the onset of syllable 1 and the Sonorant-Vowel boundary in syllable 2 (x-axes) for each subject. The plotting symbols represent different Nasal-Vowel ratios. The dashed line is a reference whose slope is 1 and intercept is 0. (Xu 1998:198)

5.1.2 The distribution of contour tones

In recent years, there have been a number of studies looking at the distribution of contour tones in different languages (Duanmu 1994a, Gordon 1999, 2001, Zhang 2001, to cite just a few). Through these studies, a consensus seems to be emerging. That is, the ability of a syllable to carry contour tones is directly related to the duration of its tone-bearing portion. Zhang (2001:331) reports that, for example, "syllable types which have longer sonorous duration of the rime, e.g., long-vowelled, sonorant-closed, stressed, final in a prosodic domain, and being in a shorter word, are more likely to carry contour tones." Gordan (2001: abstract) states that "long vowels are most likely to carry contour tones, followed by syllables containing a short vowel plus a sonorant coda, followed by syllables containing a short vowel plus an obstruent coda, followed by open syllables containing a short vowel." What is still not yet clear from these mostly typological studies is the exact phonetic mechanisms of such duration dependency. So far, most accounts are treating perception as the determinant factor. Our newly gained understanding about articulatory constraints as discussed in Section 2, however, suggests that we should also take a closer look at the articulation process as a possible contributor. To do that, again I shall examine the issue in light of the Target Approximation model.

According to the Target Approximation model, the implementation of a pitch target does not start until that of the previous target is over. This requirement would often put a lot of strain on the implementation of a contour tone. This is because the offset F_0 in the preceding tone may or may not be close to the initial pitch of the contour tone. Unless there are some very specific tonotactic rules in the language, quite often a situation like the ones depicted in Figure 10(a) would occur. That is, two pitch movements need to be made within the time interval allocated to the contour tone. The first movement is the transition toward the initial pitch of the contour tone, and the second is the movement intrinsic to the contour tone itself. This situation would arise whether the contour tone is assumed to consist of two static elements or a single dynamic element. Having recognized this situation, we may calculate how much time it would take for an average speaker to complete two consecutive pitch movements, using equations (3) and (4). For a symmetric down-up or up-down movement the size of 4 semitones,

(5) $t = 89.6 + 8.7 \times 4 + 100.4 + 5.8 \times 4 = 248$ ms

Anyone who has looked into the issue of how duration relates to contour tones would recognize this 248 ms as being quite long. This is because in syllables that can carry contour tones, the vowel is often much shorter than 248 ms. The left half of Table I lists vowel durations in syllables that carry contour tones in five languages based on several studies. All of them are much shorter than 248 ms.[3] Even if we take individual differences into consideration, the fastest speaker in Xu & Sun (2002) would need 196 ms to shift pitch up and down or down and up by 4 semitones (cf. Table V in Xu & Sun 2002), which is longer than most of the durations in Table I. This discrepancy should become less puzzling if we recognize that the syllable—rather than the rhyme or the nuclear vowel—is the domain of tone implementation, as is assumed in the Target Approximation model. The movement toward the initial value of a contour tone therefore should start at the onset of the syllable. This movement can be carried out whether or not the vocal folds are vibrating. As found in Xu, Xu, & Sun (2003), the effect of voiceless consonants such as stops and fricatives is to introduce rather local

[3] There are of course many cases where the vowels are much longer than 248 ms, as cited in Gordon (2001) and Zhang (2001). But those would not have caused the problem being discussed here in the first place.

perturbations without changing the carryover or anticipatory tonal variations reported in previous studies such as Xu (1997, 1999). Figure 14 displays the F_0 contours of Mandarin syllables /ma/, /da/, /ta/, and /sha/ 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 we put aside these local effects, the F_0 curves in /da/, /ta/, and /sha/ look very similar to those of /ma/. Hence, by the time the apparent local effect is over, F_0 is already quite low in R but quite high in F. So there is good reason to assume that the C interval in Mandarin is also used for implementing the tonal targets, whether or not voicing continues through the interval. With C included, the total duration available for tones, i.e., the duration of the entire syllable (for Mandarin and Shanghainese) are as shown in the right half of Table I.

Table I: Duration measurements of 6 languages. Left half: vowel or rime duration. Right half: syllable duration. In columns 3-5, the two numbers separated by "/" are mean duration measures from the first and second syllables in disyllabic words, respectively.

	Gordon (1999)	Zhang (2001)	Xu (1997)	Xu (1997)	Xu (1999)	Duanmu (1994a)
Hausa	133	109				
Navajo	173	209 (rime)				
Luganda		179				
Xhosa		212				
Mandarin		151/213	122/140 (R) 115/135 (F)	185/196 (R) 183/193 (F)	198 (R) 184 (F)	215
Shanghainese						162

But even these syllable durations are not very long compared to the 248 ms calculated earlier. Recall that the minimum time of pitch change obtained in Xu & Sun (2002) is for a complete pitch shift, starting at one turning point and ending at the next. As can be seen in Figure 12 as well as Figure 3, the movements in R and F do not come to a stop by the end of the syllable. In the lower left panel of Figure 3, the full reverse of the F_0 movement at the end of a dynamic tone sometimes (in fact quite regularly in Mandarin) occur in the early portion of the next syllable. When we take all these factors into consideration, it seems that, at least in Mandarin, there is just enough time for a contour tone to be effectively implemented. This observation is corroborated by the finding of Xu & Sun (2002) that in Mandarin, it is precisely during the production of the dynamic tones that the maximum speed of pitch change is approached.

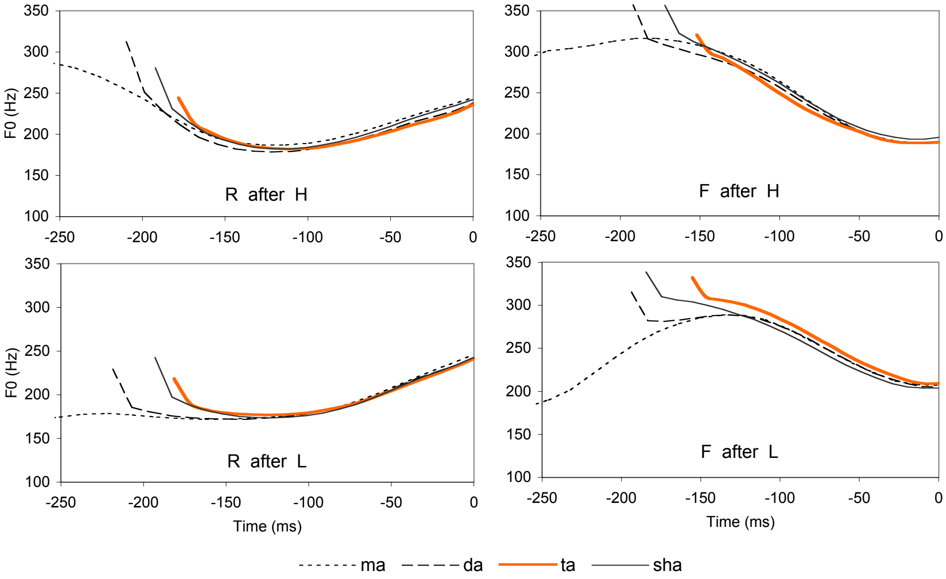

Figure 14: Effects of voiceless consonants on the F_0 contours of Mandarin R and F produced after H and L. Each curve is an average across 5 repetitions, 2 carrier sentences, and 7 female speakers. All curves are aligned to the syllable offset.

If, as has been shown, even in Mandarin where they remain functional in connected speech, the dynamic tones are often realized only with speakers' best effort (Xu & Sun 2002), and sometimes with reduced perceptibility (Xu 1994), with shorter syllable durations such as those in Shanghainese (162 ms, Table I, rightmost column), and for some types of syllables in a language that have short durations (Gordon 1999, 2001, Zhang 2001), it would be virtually impossible to implement dynamic tones without severely compromising the targeted contours. Therefore, although perception may be the last straw in the breakdown of the transmissibility of the dynamic tones in short syllables, it is the articulatory constraints that must have made the implementation of the pitch contours impossible in the first place.

Further support for this understanding comes from the findings by Janse (2003) as discussed in 3.3. Recall that what she found was that natural-fast speech actually had lower intelligibility than the linearly time-compressed normal speech. According to her report, the human perception system can easily handle resynthesized speech well over twice as fast as normal speech. This demonstrates the perceptual system's great potential in processing fast acoustic events. In contrast, naturally produced fast speech apparently contained too much undershoot for the perceptual system to fully "unwind" the human slurring. It therefore seems that articulatory constraints are probably the real bottleneck blocking the preservation of contour tones in very short syllables.

Finally, the duration sensitivity of contour tones also further confirms the strict phase relation between laryngeal and supralaryngeal movements. As shown earlier, the shortage of time for the contour tones is often in the range of tens of milliseconds. If there were ways in which speakers could micro-adjust the alignment of tonal target relative to the syllable, the occurrence of the contour tones would not have been so restricted by syllable duration.

5.2 Tone sandhi and tonal coarticulation

Tone sandhi is a term used to refer to post-lexical tonal changes that are conditioned by various phonetic, morphological, and syntactic factors (Chen 2000). The phenomena covered by the term, which are predominantly found in East Asian languages, are very diverse, as discussed in detail by Chen (2000). The mechanisms behind these phenomena, however, remain mostly unclear as of today. Tonal coarticulation usually refers to tonal variations that are strictly conditioned by tonal context and are phonetically motivated. However, partly due to the fact that the term *coarticulation* itself is yet to be clearly defined (cf. Daniloff & Hammarberg 1973, Hammarberg 1983), it is often not so easy to separate tonal variations due to coarticulation from those due to sandhi, as pointed out by Chen (2000: 25f). The Target Approximation model, since it is based on specific articulatory constraints as discussed in 2.1-2.2, makes explicit assumptions about the mechanisms of F_0 variations. It is therefore possible to reexamine tonal variations covered by both terms in light of the Target Approximation model. Under this model, it is relatively easy to make a distinction between tonal variations that are due to changes in the pitch targets, and variations that are due to implementation of the same target under different conditions. This is because the model can help us not only to identify tonal variations that are directly due to articulatory constraints, but also to rule out variations that are unlikely to be directly attributable to articulatory constraints. To be able to do that, however, detailed data are needed. I shall therefore discuss only a number of cases in Mandarin for which an extensive amount of data has been accumulated. These will include the following tonal variations, all of which were originally proposed by Chao (1968 and earlier) and widely accepted as part of the tonal phonology of Mandarin Chinese.

1. The Half L rule. L, which in isolation has the purported value of 214, loses its final rise before another tone to become 21: 214 → 21 / __ T
2. The L to R rule. L changes into R when followed by another L: L → R / __ L
3. The R to H rule. R changes into H when it is the second tone in a trisyllabic word or phrase in which the first tone is either H or R: R → H / {H, R} __ T

4. The Half F rule. F, which in isolation has the purported value of 51, loses its final portion before another F to become a half fall: 51 → 53 / __ F.
5. The neutral tone rule. The pitch value of the neutral tone varies with the preceding full tone (see 5.2.4. for detailed description).

   (The numerals in the rules are the same as proposed by Chao, with 1 indicating the lowest tonal pitch level and 5 the highest. T stands for any tone.)

As found later in a number of instrumental studies, all these patterns do manifest themselves acoustically (e.g., Lin 1985, Lin et al. 1980, Lin & Yan 1991, Shen, J. 1994, Shen, X. 1990, 1992, Shih 1988, Wu 1982, 1984, Xu 1994, 1997, 1999). However, they also appear to be different from each other in various ways, but the picture is far from clear. In the following discussion, I shall try to apply the Target Approximation model to these phenomena and try to divide them into two types, variations due to target alternation and variations due to articulatory implementation. *Target alternation* occurs in cases where the pitch target of a tone is presumably changed before being implemented in articulation. *Implementational variations*, on the other hand, are cases where tonal targets remain the same, but the acoustic realization of the targets is varied due to their implementation in different tonal contexts and/or with different amounts of articulatory effort.

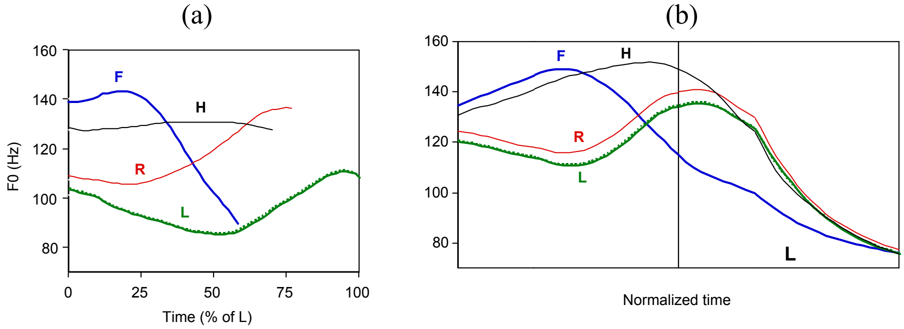

Figure 15: (a): Four Mandarin tones produced in isolation. (b): Mandarin L tone after four different tones, produced in carrier phrases. Adapted from Xu (1997)

5.2.1 L variations

Three types of variations of the tone L can be seen in Figure 15. Figure 15(a) displays the average F_0 curves of the four Mandarin tones said in isolation (Xu 1997). Note that L in this graph has a final rise, which largely agrees with Chao's (1968) description that L said in isolation has a shape of 214, although the final rise in Figure

15(a) does not go as high as the value 4 would suggest. In Figure 15(b) are disyllabic sequences produced in a carrier whose first syllable after syllable 2 in the graph carries either F or R. In the figure, syllable 2 is associated with L while syllable 1 is associated with four different tones. We can see that L in syllable 2 has no trace of the final rise as that 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 is much reduced. It is true that the duration of the L-carrying syllable in Figure 15(b) is shorter than that in Figure 15(a) (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 15(b). 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 routinely done, as in the case of the dynamic tones such as R and F discussed in 5.1. Furthermore, as found in Xu & Sun (2002), the speed of F_0 drop related to L is well within the maximum speed of pitch lowering (e.g., 35 st/s vs. 52 st/s for lowering pitch by 7 st). Thus it is rather unlikely that the lack of final rise in L in this case is due to the constraint of maximum speed of pitch change. It is more likely, instead, that the pitch target implemented for L in such a situation has no final rise to begin with. Thus the alternation between L with and without a final rise probably involves changes in the pitch targets before their actual articulatory implementation.

The second type of L variation can be seen in Figure 15(b). In the L L sequence, the first L apparently is not very different in shape from R in the same syllable, although the two differ somewhat in overall height. Wang & Li (1967) have shown that Mandarin listeners cannot distinguish words and phrases with L L sequence from those with R L sequence carried by identical CV syllables. This has been further confirmed by Peng (2000) for Taiwan Mandarin. 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 (Peng 2000, Xu 1993, 1997, Zee 1980), as is also apparent in Figure 15(b), 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, because there is no mechanism in the Target Approximation model that can generate 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 is the result of implementing a complex target that is similar to that associated with the isolated L as in Figure 15(a)? 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 syllable is to sacrifice the F_0 minimal 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 not only raises the F_0 minimum from 85 Hz in Figure 15(a) to 110 Hz in Figure 15(b), but also lifts 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 magnitude as found in Xu (1997) 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 15(b). When syllable 1 carries different tones, L in syllable 2 has rather different onset F_0, which is virtually determined by the offset F_0 of syllable 1. These variations, because they can be readily explained by asymptotic approximation of the same [low] target when having different onset F_0 values, are apparently directly related to inertia. They should therefore be considered as cases of implementational variations.

F_0 variations of L in Mandarin thus seem to be governed by two distinct processes, variation due to target alternation and variation due to articulatory implementation. The former is responsible for the alternation between the fall-rise contour seen in isolation and the low dip contour seen in most non-final positions. The latter is responsible for the variations in the amount of F_0 drop in non-final L. It is unclear whether the largely rising contour in the first L of L L is due to a complete change of the target to [rise] as in R or due to implementation of the same complex target as in isolated L with a time constraint. Further studies are needed to sort this out.

5.2.2 R variations

As discussed in 3.1, Xu (1994) reports two findings: (a) R produced on the second syllable of a tri-syllabic word 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 (referred to as the conflicting tonal context in the paper); and (b) despite the flattening, R is still correctly identified about 88% of the time when listeners hear it along with the original tonal context (with semantic information removed). Shih & Sproat (1992) report that R produced in the {H, R} __ T context on the second syllable of a trisyllabic 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 are 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 tetrasyllabic word than on the second syllable of a trisyllabic word, because syllable strength is presumably even weaker in the former than in the latter. It is therefore highly likely that the variant forms of R as described by Chao is a variation due to articulatory implementation with no change in the underlying pitch target.

5.2.3 F variations

Figure 16(a) shows mean F_0 curves of 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. Rather, it always reaches about half way toward the lowest point. This is in sharp contrast to the final F_0 reached for F produced in isolation as shown in Figure 15(a). In effect, therefore, the F_0 contour in F is a "half fall" when followed by any other tone in a disyllabic sequence. The Half F rule thus should be more appropriately stated as $51 \rightarrow 53 \ / \ \_\_ \ T$.

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 16(b), H, R, and F all start from the bottom when preceded by L, this despite the fact that the ideal starting pitches for these tones are very different: high for H and F, and low for L. But this should be exactly the case according to the Target Approximation model, which assumes that F_0 of any tone is largely determined by the offset F_0 of the preceding tone. That F_0 of F does not fall to the bottom of the pitch range has little to do with the characteristic of the following tone. Rather it is attributable logically 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 16(a), 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 articulatory limit. 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 it 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 any articulatory constraint, but because it is targeted to reach there, possibly by a language-specific rule, when the tone is not utterance final.

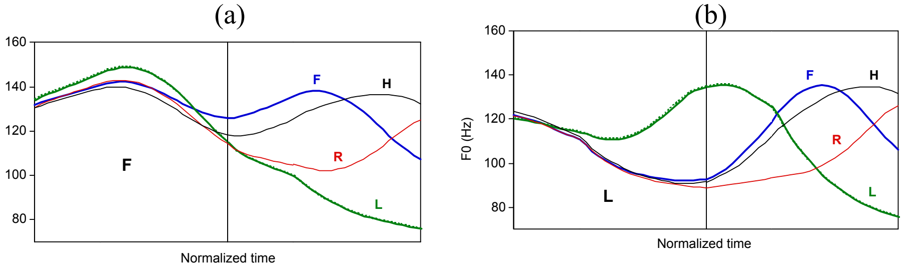

Figure 16: Effects of the following tone on the F0 contour of F (a) and L (b) in Mandarin. The vertical lines indicate the onsets of syllable-initial nasals. Each curve is a (segment-by-segment) time-normalized average of 192 tokens produced by eight speakers. Adapted from Xu (1997)

There must then be something special about final position (including in isolation) as opposed to other positions. Bruce (1977) has argued that in Stockholm Swedish the most likely canonical form of a lexical tone (often known as accent in Swedish) occurs not in isolation or in final position but in a prenuclear position (i.e., before a focus). Furthermore, F_0 of a tone in final position actually consists of the tone proper plus a boundary tone. Boundary tones have also been proposed for English (Pierrehumbert 1980). It is possible that Mandarin, too, has boundary tones. However, even if there were such boundary tones in Mandarin, their forms and interactions with the lexical tones must be rather complex. In Figure 15(a), for example, all four tones are produced in isolation without special intonation and hence would probably carry a boundary tone appropriate for a statement. It does not seem to help to assume that either the final rise of L or the final fall of F is due to the boundary tone. This is because, for one thing, they are so different, and for another, they are virtually absent in H and R, which appear to be very similar to their variants in non-final positions. The exact nature of boundary tones in Mandarin—assuming they exist—thus awaits future research.

5.2.4 Neutral tone variations

Syllables carrying the neutral tone in Mandarin are generally believed to have no tone of their own (Yip 2002). Nevertheless, in speech, they have to be said with some F_0 values. According to Chao (1968), the pitch value of the neutral tone depends on the full tone preceding it, as shown in column 3 of Table II. Shih (1988) finds that the F_0 of the neutral tone is often not static, but rather either rising or falling, as shown in column 4 in Table II. Some recent studies have reported similar findings (Wang 1997, Liang ms). What seems puzzling is why a “toneless” tone would have so many variant forms. If, as Chao and Shih have shown, these variant forms are closely related to the preceding tone, are they then governed by arbitrary rules, or are they actually predictable?[4]

Table II. Pitch values of the neutral tone after different tones according to Chao (1968) and Shih (1988).

Preceding tone	Example	Chao 1968	Shih 1988
H	ta de [his]	half-low	starts high, then falls
R	huang de [yellow]	middle	starts high, then falls, but not as low as after H
L	ni de [yours]	half-high	starts fairly low, then rises
F	da de [big]	low	starts fairly low, and falls even lower

Figure 17 shows F_0 contours of the neutral tone as compared to full tones in similar tonal contexts (Data from Chen & Xu 2002). In the figure, the number of successive neutral tone syllables varies from 0 to 3. The pinyin above each graph is for only one of the sentences in the graph, i.e., the one with H on the first syllable of the target word/phrase (henceforth syllable 1). What we can see in Figure 17 is that the behavior of the neutral tone is actually somewhat similar to that of a full tone. In Figure 17(a), F_0 of F in syllable 1 manifests the typical patterns of a full tone when preceded by different tones: starting the transition toward the pitch target (presumably a linear [fall]) at the syllable onset and continuing throughout the syllable. As a result of this continuous transition, by the end of the syllable, all four curves have effectively converged to the targeted [fall]. In Figure 17(b), the tone in syllable 2 is neutral (N). But the F_0 contours in the syllable look somewhat similar to those in Figure 17(a): starting to move away from the offset F_0 of the preceding tone at the onset of the current syllable in the direction of a converging point toward the end of the syllable. What is different in Figure 17(b) is that the four F_0 contours have not merged even by the end of the syllable. But, as can be seen in Figure 17(c) and (d), as the number of consecutive neutral tones increases, the F_0 contours become closer and closer to each other, except for the contour after L. Somehow L seems to raise the F_0 of the following tone, which is visible even in Figure 17(a) where the tone of syllable 2 is F. But this raising effect is apparently limited in time, because in Figure 17(c) and (d), by the middle of syllable 3 F_0 following L already starts to drop, thus joining the converging transitions of the other three curves.

From the perspective of the Target Approximation model, the above observations

[4] Note that the articulatory and perceptual constraints discussed in this paper cannot explain where and why the neutral tone occurs. As a language-dependent phenomenon, the origin of the neutral tone, as those of the full tones, is unlikely to be either purely articulatory or perceptual. Nevertheless, with improved understanding of the neutral tone in Mandarin, especially in regard to the role of articulatory effort, we may also improve our understanding of other related issues in both tone languages and non-tone languages.

suggest three things. First, the neutral tone is not totally targetless. If it were, there would not be such apparent converging transitions through the course of the neutral tone syllable. This transition is especially striking after R in syllable 1 in Figure 17(b)-(d), where F_0 always starts to drop in the middle of the very first neutral tone despite the fact that the initial F_0 contour in the syllable sharply rises, presumably due to the rising momentum of the preceding R (Xu 2001b). Second, the implementation of the neutral tone target is not done with a strong articulatory effort, as is evident from the much slower convergence as compared to the quick merger seen in F and L in Figure 17(a). Third, L raises the F_0 of the following tone, which is maximally manifested on the neutral tone. Leaving aside the F_0 raising effect of L, which seems to be independent of the other two observations, it seems that the variant F_0 contours of the neutral tone is actually easy to understand. They probably all result from asymptotic approximation of some simple and likely static pitch target with much reduced articulation effort. As for the value of the pitch target for the neutral tone, it seems to be somewhere in the middle of the pitch range, because it is lower than the high F_0 in F but higher than the low F_0 in L, as found in Chen & Xu (2002).

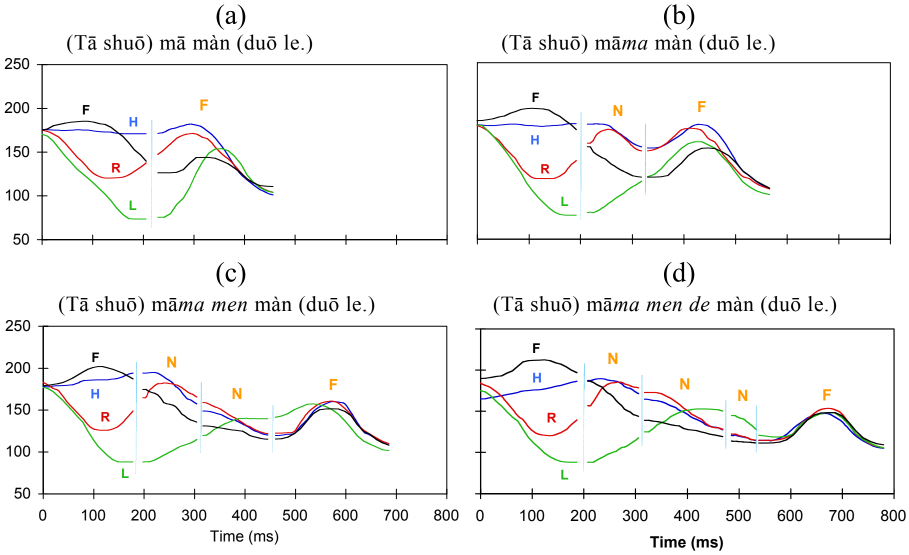

Figure 17: Mean F_0 contours of Mandarin sentences containing 0-3 neutral tone (N) syllables. In all graphs, the tone of syllable 1 alternates across H, R, L, and F. In (a) the tone following syllable 1 is F. In (b)-(c), there are 1-3 neutral tone syllables following syllable 1. Vertical lines in the graphs indicate syllable boundaries. Data from Chen & Xu (2002).

6. Conclusions

I have argued in this paper that tone research has to take articulatory constraints seriously, given recent advances in our knowledge about both speech production and speech perception. In particular, the maximum speed of pitch change has recently been found to be slower than previously thought. As a result, syllables are often shorter than the minimum time it takes to complete a pitch change of only a few semitones. Furthermore, the constraint of synchronizing laryngeal and supralaryngeal movements is likely just as strong, because it seems to prevent micro-alignment adjustment from happening. The lack of such adjustment under time pressure often results in undershoot of tonal targets. In contrast to the these articulatory constraints, there is evidence that the human perceptual system is actually very proficient in handling fast changing acoustic events as well as phonetic deviations due to articulatory constraints. But perception apparently prefers less articulatory undershoot rather than more. To put the articulatory and perceptual constraints into a cohesive system with which the generation of F_0 contours can be simulated, I presented a Target Approximation model that takes these constraints as part of the basic assumptions. Applying the model to some of the standing issues in tone research, I have demonstrated that they can be better explained in terms of the interaction between underlying pitch targets associated with lexical tones, which are language-specific, and physical constraints that are inherent to the articulatory system, which are likely universal.

Many tone related issues remain unresolved, however, not only because of lack of acoustic data, but also because, more importantly, lack of specifically designed experiments. For example, whether a particular tone in a language is inherently dynamic or composed of static elements can be seen more clearly only when speech rate is slowed down so that there is sufficient time for the underlying target(s) to be fully implemented. Future studies, therefore, should deliberately manipulate factors that may interact with known articulatory constraints to reveal the true targets that are being implemented in articulation.

References

Abramson, A. S. 1978. The phonetic plausibility of the segmentation of tones in Thai phonology. *Proceedings of the 12th International Congress of Linguistics*, 760-763. Vienna.

Abramson, A. S. 1979. The coarticulation of tones: An acoustic study of Thai. *Studies in Tai and Mon-Khmer Phonetics and Phonology in Honour of Eugenie J. A. Henderson*, ed. by T. L. Thongkum, P. Kullavanijaya, V. Panupong and K. Tingsabadh, 1-9. Bangkok: Chulalongkorn University Press.

Atterer, M., and D. R. Ladd. 2004. On the phonetics and phonology of “segmental anchoring” of F0: Evidence from German. *Journal of Phonetics* 32:177-197.

Bell-Berti, F. 1993. Understanding velic motor control: Studies of segmental context. *Nasals, Nasalization, and the Velum*, ed. by M. K. Huffman and R. A. Krakow, 63-85. San Diego: Academic Press.

Bell-Berti, F., and K. Harris. 1982. Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America* 71.2:449-454.

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

Bruce, G. 1977. Swedish word accents in sentence perspective. *Travaux de l’Institute de Linguistique de Lund*, XII, ed. by B. Malmberg and K. Hadding. Lund: Gleerup.

Caspers, J., and V. J. van Heuven. 1993. Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica* 50:161-171.

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

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

Chen, Y., and Y. Xu. 2002. Pitch target of Mandarin neutral tone. Paper presented at LabPhon 8. New Haven, CT, USA.

Daniloff, R. G., and R. E. Hammarberg. 1973. On defining coarticulation. *Journal of Phonetics* 1:239-248.

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

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

Fowler, C. A. 1984. Segmentation of coarticulated speech in perception. *Perception & Psychophysics* 36:359-368.

Fowler, C. A., and M. Smith. 1986. Speech perception as “vector analysis”: An approach to the problems of segmentation and invariance. *Invariance and Variability in Speech Processes*, ed. by J. S. Perkell and D. H. Klatt, 123-139. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Gordon, M. 1999. *Syllable Weight: Phonetics, Phonology, and Typology*. Los Angeles: University of California dissertation.

Gordon, M. 2001. A typology of contour tone restrictions. *Studies in Language* 25:405-444.

Greenberg, S., and E. Zee. 1979. On the perception of contour tones. *UCLA Working Papers in Phonetics* 45:150-164.

Hammarberg, R. 1982. On redefining coarticulation. *Journal of Phonetics* 10:123-137.

Harris, M. S., and N. Umeda. 1987. Difference limens for fundamental frequency contours in sentences. *Journal of the Acoustical Society of America* 81:1139-1145.

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

Janse, E. 2003. Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. *Speech Communication* 42:155-173.

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

Kelso, J. A. S., K. G. Holt, P. Rubin, and P. N. Kugler. 1981. Patterns of human interlimb coordination emerge from the properties of non-linear, limit cycle oscillatory processes: Theory and data. *Journal of Motor Behavior* 13:226-261.

Klatt, D. H. 1973. Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. *JASA* 53:8-16.

Ladd, D. R. 1978. *The Structure of International Meaning*. Ithaca: Cornell University dissertation.

Leben, W. R. 1973. *Suprasegmental Phonology*. Cambridge: MIT dissertation.

Lee, C.-Y. 2001. *Lexical Tone in Spoken Word Recognition: A View from Mandarin Chinese*. Providence: Brown University dissertation.

Liang, L. (ms). Putonghua qingsheng de shiyan yanjiu [An experimental study of Mandarin neutral tone].

Liberman, M., and J. Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. *Language Sound Structure*, ed. by M. Aronoff and R. Oehrle, 157-233. Cambridge: MIT Press.

Lin, M. 1995. A perceptual study on the domain of tones in Standard Chinese. *Chinese Journal of Acoustics* 14:350-357.

Lin, M., L. Lin, G. Xia, and Y. Cao. 1980. Putonghua erzici biandiao de shiyan yanjiu [An experimental study of tonal variation in disyllabic words in Standard Chinese].

Zhongguo Yuwen [*Chinese Linguistics*] 1980.1:74-79.

Lin, M., and J. Yan. 1980. Beijinghua qingsheng de shengxue xingzhi [Acoustic properties of Mandarin neutral tone]. *Fangyan* [*Dialect*] 1980.3:166-178.

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

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

Liu, F., and Y. Xu. 2003. Underlying targets of initial glides—Evidence from focus-related F0 alignments in English. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1887-1890. Barcelona, Spain.

O'Connor, J. D., and G. F. Arnold. 1961. *Intonation of Colloquial English*. London: Longmans.

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

Peng, S.-H. 2000. Lexical versus 'phonological' representations of Mandarin Sandhi tones. *Papers in Laboratory Phonology*, Vol.5: *Acquisition and the Lexicon*, ed. by M. B. Broe and J. B. Pierrehumbert. Cambridge: Cambridge University Press.

Pierrehumbert, J. 1980. *The Phonology and Phonetics of English Intonation*. Cambridge: MIT dissertation.

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

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

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

Schmidt, R. C., C. Carello, and M. T. Turvey. 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.

Shen, J. 1994. Beijinghua shangsheng liandu de diaoxing zuhe he jiezou xingshi [Tonal contour combination and rhythmic forms of Tone-3 sandhi in Beijing Mandarin]. *Zhongguo Yuwen* [*Journal of Chinese Linguistics*] 241:274-281.

Shen, X. S. 1990. Tonal coarticulation in Mandarin. *Journal of Phonetics* 18:281-295.

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

Shih, C.-L. 1988. Tone and intonation in Mandarin. *Working Papers, Cornell Phonetics Laboratory* 3:83-109.

Shih, C.-L., and R. Sproat. 1992. Variations of the Mandarin rising tone. *Proceedings of the IRCS Workshop on Prosody in Natural Speech No.92-37*, 193-200. Philadelphia.

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

't Hart, J. 1981. Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America* 69:811-821.

't Hart, J., and A. Cohen. 1973. Intonation by rule: A perceptual quest. *Journal of Phonetics* 1:309-327.

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

Wang, J. 1997. The representation of the neutral tone in Chinese Putonghua. *Studies in Chinese Phonology*, ed. by J. Wang and N. Smith, 157-183. Berlin: Mouton de Gruyter.

Wang, W. S-Y. 1967. Phonological features of tone. *International Journal of American Linguistics* 33:93-105.

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

Woo, N. 1969. *Prosody and Phonology*. Cambridge: MIT dissertation.

Wu, Z. 1982. Putonghua yuju zhong de shengdiao bianhua [Tonal variations in Mandarin sentences]. *Zhongguo Yuwen* [*Chinese Linguistics*] 1982.6:439-450.

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

Xu, C. X., Y. Xu, and X. Sun. 2003. Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association* 33:165-181.

Xu, Y. 1993. *Contextual Tonal Variation in Mandarin Chinese*. Storrs: The University of Connecticut dissertation.

Xu, Y. 1994. Production and perception of coarticulated tones. *Journal of the Acoustical Society of America* 95:2240-2253.

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

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

Xu, Y. 1999. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics* 27:55-105.

Xu, Y. 2001a. Sources of tonal variations in connected speech. *Tone, Stress and Rhythm in Spoken Chinese*, ed. by Hana Trísková, 1-31. Journal of Chinese Linguistics Monograph Series 17. Berkeley: Project on Linguistic Analysis, University of California.

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

Xu, Y., and F. Liu. 2002. Segmentation of glides with tonal alignment as reference. *Proceedings of 7th International Conference on Spoken Language Processing*, 1093-1096. Denver, Colorado.

Xu, Y., K. Liu, D. Surendran, and A. Wallace. (in progress). Maximum speed of articulatory movements.

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

Xu, Y., and Q. E. Wang. 1997. Components of intonation: What are linguistic, what are mechanical/physiological? *Proceedings of Presented at International Conference on Voice Physiology and Biomechanics*. Evanston, Illinois.

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

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

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

Zhang, J. 2001. *The Effects of Duration and Sonority on Contour Tone Distribution—Typological Survey and Formal Analysis*. Los Angeles: University of California dissertation.

[Received 18 March 2003; revised 15 June 2004; accepted 19 June 2004]

Haskins Laboratories
270 Crown Street
New Haven, CT 06511-6695
USA
xu@haskins.yale.edu

從產生和感知認識聲調

許　毅
霍金斯研究室

認真考量語音產生和感知上的種種限制，將有助於我們對聲調的認識。具體的說，音高變化的最大速率和咽喉動作之間的協調，對詞彙聲調的產生加上了一些無法超越的限制；而儘管人類的感知系統可以很有效率的處理高速變化的聲學事項並解決因發音限制所導致的扭曲訊息，但在感知上能復原多少也仍有諸多限制。有鑑於這些限制，聲調產生的「目標近似模型」乃應運而生。這個模型將基頻的產生視爲對目標音高的漸近過程，而這些目標音高就是由語言規則所推導的個別聲調。由於能對發音時的聲調變體與目標音高的聲調交替作出明確的區別，「目標近似模型」得以幫助我們認識許多與聲調有關的問題。關於聲調及聲調起伏分布的本質、不同連調變化現象的辨別、輕聲的目標音高和發音方式等等的問題，也都可以從「目標近似模型」中獲得啓發。

關鍵詞：聲調，連調變化，音高目標，目標近似，輕聲