To be presented at the Eighth International Symposium on Chinese Languages and Linguistics (IsCLL VIII), November, 2002, Taipei, Taiwan

UNDERSTANDING TONE FROM THE PERSPECTIVE OF PRODUCTION AND PERCEPTION

Yi Xu

Department of Linguistics
The University of Chicago

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 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, often referred to broadly as "coarticulation" (Daniloff & Hammarberg 1973; Hammarberg 1982), 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 will show that the maximum speed of pitch change has recently been found to be slower than has been thought before, and that in tone production speakers often have to get much closer to their articulatory limits than have been recognized previously (Xu & Sun, 2002). I will also show that an additional, and probably even stronger, constraint on tone may come from the coordination of laryngeal and supralaryngeal movements. Furthermore, I will show that a number of recent perceptual findings seem to indicate that human perception is actually quite good at handling very fast acoustic events in the speech signals, better probably than we have thought before (Quené & Janse 2001; Janse, Nooteboom & Quené in press; Lee 2001). Based on these new empirical data, I will argue that many observed tonal patterns are probably more closely related to articulatory limitations than we have thought before. I will further demonstrate how these new findings can be incorporated into a pitch target implementation model for F_0 contour generation. When applied to a number of tonal phenomena, as I will show, this model may help us gain new insights into the underlying mechanisms of these phenomena.

2. ARTICULATORY CONSTRAINTS

When considering articulatory constraints, what first come to one's mind are likely static limits such as the highest and lowest pitch values, the highest and lowest jaw positions, etc. Some of those limits are probably indeed seldom approached in speech, such as the lowest jaw position and the highest pitch. Some other static limits are probably quite often approached, such as the highest jaw position and the lowest pitch. Beside those static limits, however, there are also dynamic limits inherent to the articulatory system that are probably just as important, but have not received much attention. In the following I will consider recent findings about some of the dynamic limits. In particular, I will discuss the maximum speed of pitch change and the coordination of segmental and suprasegmental movements. I will 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 by 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 the 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 used by Praat (www.praat.org), each shift appears to be completed in about 10-20 ms.

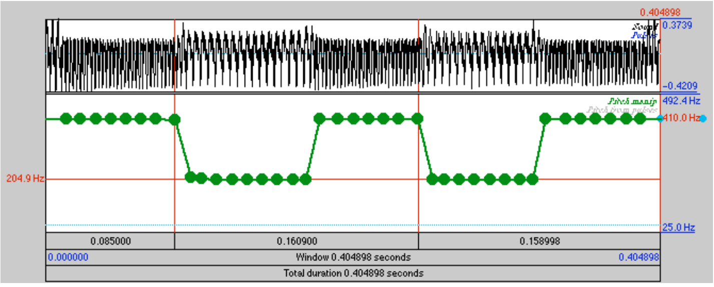

Figure 1. One of the model pitch alternation patterns used in Xu and Sun (2002). The pattern is HLHLH with a pitch range of 12 semitones (1 Octave) and a HL cycle duration of 167 ms (6 cycles per second).

Figure 2 shows pitch tracking of an actual pitch alternation pattern produced by one of the subjects in Xu and 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. Furthermore, we measured the time interval corresponding to the middle 75% of the magnitude of each pitch shift. This measurement was used by Ohala and Ewan (1973) and Sundberg (1979) as estimates of the maximum speed of pitch change.

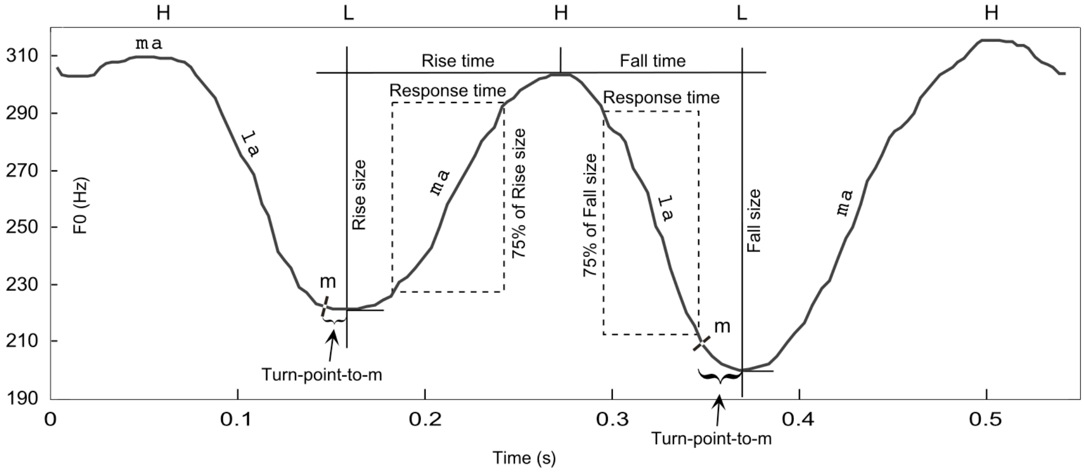

Figure 2. Illustration of measurement of rise and fall excursion time, rise and fall “response time”, and turn-point-to-m in a HLHLH trial spoken with /malamalama/ in Xu and Sun (2002). See original paper for more detailed explanations.

Our analyses of the measurements revealed several interesting results. First, we found that 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 for maximum speed of pitch rises and falls as a function of pitch change size are represented by the linear equations in (1) and (2), respectively,

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

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

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 imperative to take the size of pitch change into consideration.

The second result of interest in Xu and Sun (2002) is that the minimum time it takes to complete a pitch change is found to be 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 for the minimum time of pitch rises and falls as a function of pitch change size are represented by the linear equations in (3) and (4), respectively,

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

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

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 and 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., 't Hart et al 1990; Caspers and van Heuvan 1993, as discussed in detail in Xu and Sun 2002).

The third result of Xu and 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 found to be 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 pitch change speed reported by Caspers and van Heuven (1993) was found to be 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 and Sun (2002) also matched the speed of pitch change in the dynamic tones (R and F) in Mandarin recorded in Xu (1999) (but not in the static tones, i.e., H and L, however). 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 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 in many occasions, the fastest speed of pitch change is indeed approached in speech.

The fourth result of Xu and Sun (2002) relevant to my discussion in this paper is that, in terms of the maximum speed of pitch change, no differences were found between native speakers of American English and native speakers of Mandarin Chinese. This was found with data from 16 English speakers and 20 Chinese speakers. This indicates that, as different as English and Chinese can be in terms of their linguistic use of F_0 contours, being native speakers of either language did not result in significant physiological differences as far as 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 and Sun (2002) is applicable to languages in general. On the other hand, however, we did find differences across individuals in both our Mandarin subjects 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 could be either faster or slower.

Coming back to the example shown in Figure 2, it now becomes apparent why no pitch plateaus are formed when the speaker tries to imitate the model pitch alternation patterns consisting of static H and L. It is probably 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 will inevitably be continuous and free of consistent steady states. Furthermore, if a time slot available for reaching a pitch level is shorter than the minimally required time, the pitch movement may seem to "spill over" into the next time slot. I will return to this point in 5.1.2. concerning spreading.

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 apparent rising contour when following L (Low), as shown in Figure 3. Likewise, L is produced with an apparent 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

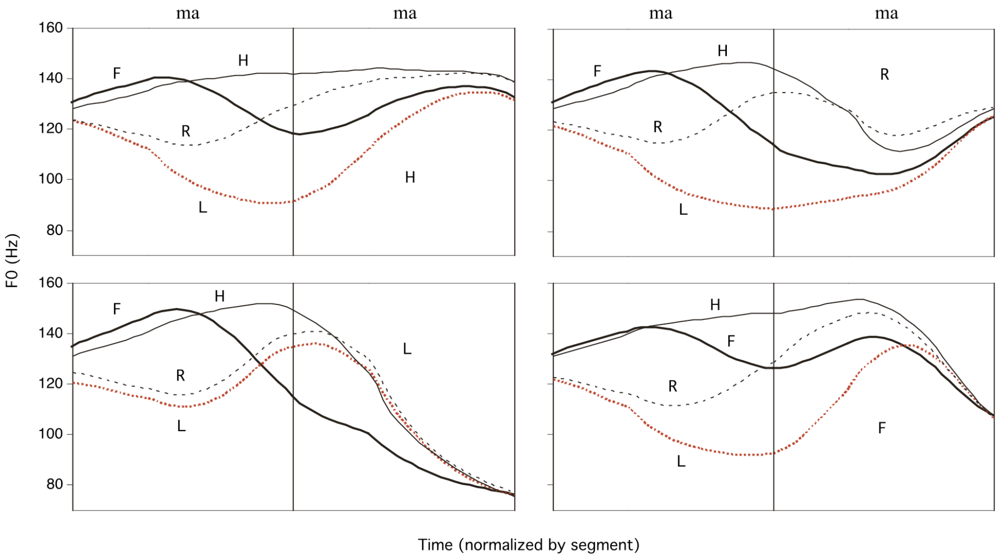

Figure 3. Effects of preceding tone on the F_0 contour of the following tone in Mandarin. In each panel, the tone of the second syllable is held constant, while the tone of the first syllable is either H, R, L or F. The vertical lines represent the syllable boundaries (at the onsets of initial nasals).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

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 achieved. The long transitions observed in H L and L H and other tone sequences in Xu (1997, 1999) therefore should mostly be attributed to the constraint of maximum speed of pitch change. There are further ramifications for dynamic tones such as R (Rising) and F (Falling), which I will discuss later in 5.2.2.

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. For Mandarin, for instance, many phonetic/phonological accounts have in fact suggested various micro adjustment schemes for tone-syllable alignment (Howie 1974; Lin & Yan 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, for example, in tone sequence such as L H, H L, L F and H R. It is imaginable that speakers can 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 and F), averaged across six repetitions produced by four male speakers. The sentences in the upper graph carry no narrow focus, while those in the lower graph 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 second syllable, the F_0 contour corresponding to the F-carrying 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 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 the upper graph at least, the height of the peak F_0 of F seems proportional to the offset F_0 of syllable 2. The fact that the F_0 of F 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, to borrow the terms more frequently associated with consonants and vowels). Assuming that such a compromise is not really desirable, for which I will 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. Such a readjustment would especially make sense 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 suggests that there must be some kind of strong constraint that has prevented the readjustment from happening.

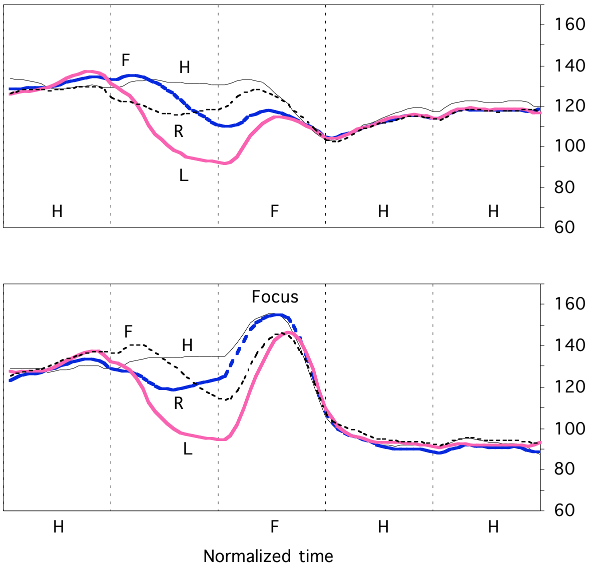

Figure 4. Mandarin tone F following four different tones. Top: no narrow focus in the sentence; Bottom, focus on the F-carrying syllable. Each curve is an average of 20 tokens produced by four male speakers (five repetitions per speaker). (Data from Xu 1999).

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 1984; Schmidt, Carello & Turvey 1990). At relatively low speed, the phase angle between two movements has to be either 180°, i.e., starting one movement after the other is half way through its cycle, or 0°, i.e., starting and ending the two movements simultaneously. At high speed, however, only the 0° 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 (3) and (4), at the fastest speed of pitch change of an average speaker, it takes at least 124 ms to complete a 4-st rise or fall, and about 107 ms to complete a 2-st rise and 112 ms to complete a 2-st fall. This means that it is virtually impossible for the speaker to maintain a 180° phase angle between pitch movement and the syllable, assuming that the syllable is managed as an articulatory cycle. Other odd phase angles would be even less likely based on the findings of Kelso (1984) and Schmidt et al. (1990). Therefore, the only likely choice left for the speaker is to maintain a 0° phase angle between tonal movement and the syllable, i.e., keeping them fully synchronized.

The recognition of this constraint may have profound impact on phonological accounts of tone-syllable alignment. What it suggests is that although tone and syllables are produced by relatively independent systems — the larynx and the vocal tract, speakers probably have very few degrees of freedom when it comes to aligning the movements generated by the two systems during production. Most likely, they can only produce them in full synchrony. Note that full synchrony of different movements does not mean that they have to be equal in duration. One movement can have half, or even 1/3 or 1/4, of the duration of the other, as long as the onset and offset of a string of faster movements coincide with those of a slower movement. Understanding this should help us understand the phase relation between a slower movement, such as that corresponding to the syllable, and faster movements, such as those corresponding to consonants and vowels. I will come back to this later in 4.1.

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 and 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 & Liu in progress). Subjects are asked to do two different tasks. In the nonspeech task, similar to what we did in Xu and 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 nonspeech tasks: ranging from 120 to 130 ms.

The minimum syllable duration of 120-130 has at least one important implication. That is, the 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. Each phase thus takes about half of the syllable cycle, i.e., 60-65 ms. (We are actually 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 in Xu and Liu (in progress), 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 to be 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 1984; 't Hart 1981). These studies, however, often use nonspeech 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 will discuss three studies which 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 phonetically 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 extensive 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'ōŋ jóu pǐŋ] (HRL) ➔ [ts'ōŋ jōu pǐŋ] (HHL) 'green onion pancake'

This description of the tonal variation implies that the tonal identity of R is changed in this environment, which means that listeners will hear the tone of [jóu] as identical to H. To see if R has indeed become indistinct from H in this case, I did a study to investigate both the acoustics and perception of R in this kind of tonal environment (Xu 1994). In the study I first examined F_0 contours of R produced in different tonal contexts. I found that in H R L: [□ / _], which I 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 semantic meanings of the original words were made neutral through acoustic manipulation, listeners were able to correctly identify R most of the time. It seems 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 H R L: [□ / _], and those that did not cause much distortion, such as 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 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 just could not help it.

3.2. Perception of tones with only initial fragments

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 the 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 __). 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 the gating experiment in Lee (2001). 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 his experiment are (a) most subjects are able to correctly identify whether the tone of the target syllable is 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 (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).

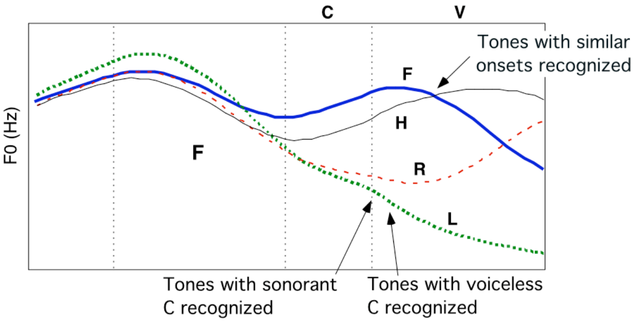

Figure 5. Schematic representation of the findings of Lee (2001). The tonal contours are adapted from Xu (1997). The arrows point to the moments when tones start to be correctly perceived.

Lee's (2001) findings may seem puzzling if we believe literally that tones are carried only by vowel or syllable rime. It would also be puzzling if we 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, of course, that the preceding tonal context is known. On the other hand, Lee's (2001) findings do not lead to the conclusion 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, it is the differences in the contours corresponding to the later portion of the syllable, which is determined by the underlying tonal target, that give rise to the differences in the initial transitions. Listeners seem to know that, because they make surprisingly effective use of this information, as is clearly demonstrated by the findings of Lee (2001).

3.3. Perception of fast speech produced by human or resynthesized with time compression

What Xu (1994) and Lee (2001) have demonstrated is listeners' amazing ability to make full use of the tonal information not only in the form of the final targets approximated but also in the form of the transitions toward those targets. If we contrast this ability with what is found in another set of recent perception studies (Quené & Janse 2001; Janse et al. in press), we can see an even clearer picture on the nature and magnitude of the effects of articulatory constraints. Quené & Janse (2001) and Janse et al. (in press) report that spoken words in natural-fast speech have reduced perceptual intelligibility. At the same time, they find that if normal speech is linearly speeded up through resynthesis, intelligibility remains very high, even at rates twice the normal rate. They also reported that intelligibility of time compressed speech is effectively reduced if normal speech is time-compressed to 35%. To investigate whether there are ways to improve the intelligibility at very fast speed, they compared three ways of achieving the time compression: (a) linearly shortening the duration of sentences produced at normal rate, (b) shortening unstressed syllables more than stressed syllables, and (c) shortening stressed syllables more than unstressed syllables. What they find in the end is that intelligibility is still the highest if the time-compression is done linearly, i.e., reducing the duration of a whole sentence by a constant factor. Based on this finding, they argue that the non-uniform way of speeding up observed in natural speech (e.g., shortening stressed and unstressed syllables differently) may be caused by the fact that speakers simply cannot do it otherwise, even though this may be harmful for perception.

The relevance of these studies on time-compressed speech to our understanding of tone perception is that the human perceptual system seems to be highly capable of processing fast changing acoustic events, provided that the magnitude of the changes is not severely reduced. On this basis, and on the basis of the findings of Lee (2001), it seems highly likely that the reduction of intelligibility reported in Xu (1994) for the "conflicting" condition and in Janse et al. (in press) for the natural-fast speech is due to excessive undershoot from the canonical phonetic targets due to constraints such as the maximum speed of articulatory movements and temporal coordination of these movements.

4. A PITCH TARGET IMPLEMENTATION 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 forces and involuntary ones. Voluntary forces originate from communicative demands, whereas involuntary forces 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. 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 has been that speakers are usually so far away from those 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 will briefly sketch a pitch target implementation 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 and Luo (1999).

4.1. The model

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 mapped* onto surface phonetic patterns. Rather, the model assumes that associated with each tone is an articulatorily operable unit called pitch target, which has a simple form such as [high], [low], [mid], [rise] or [fall]. The

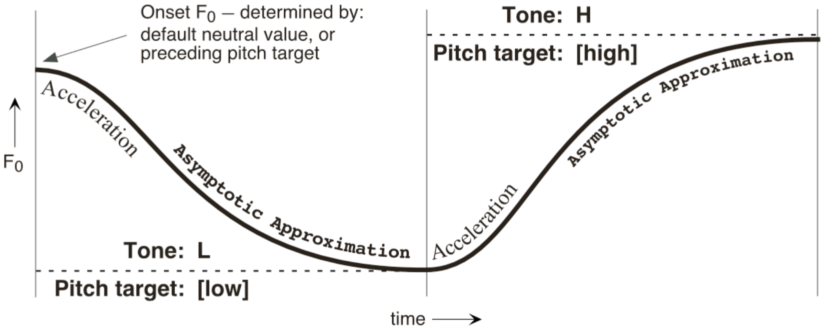

Figure 6. A schematic sketch of the pitch target implementation model. The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from articulatory implementation of the pitch targets.

process of realizing each tone is to *implement* its pitch target by articulatorily approximating the shape of the target with voluntary forces. The target approximation is assumed to be implemented as faithfully as possible. However, this implementation is done under various articulatory constraints. The first, and probably the strongest, constraint 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 7a shows the sequence said at 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 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 7b). Nevertheless, the two are still implemented sequentially within the syllable cycle.[1]

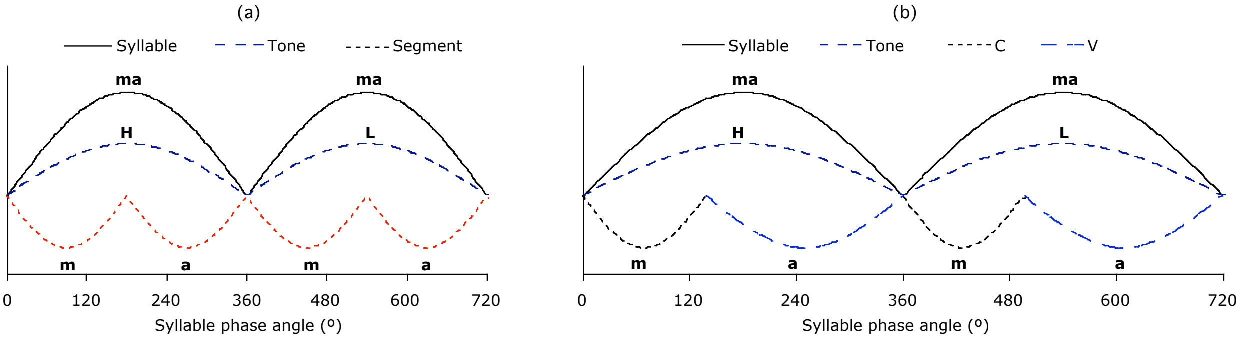

Figure 7. Schematics of synchronized phase relation among syllable, tone, consonant (C) and vowel (V): (a) fast rate, (b) slower rate.

The fitting of tones into this phase relation needs some explaining. Most phonetic and phonological accounts as of today assume that only vowels or rhymes are tone-bearing units. If we apply this assumption when considering the phase relations between tone, syllable and segments as shown in Figure 7, we would have to posit that tone cycle always skips the C cycles while keeping pace with the syllable cycle. This would be assuming that the laryngeal movement is brought to a standstill during every C. Note that even when C is voiceless, i.e., 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), they are not deactivated during C whether or not voicing continues. The phase relation depicted in Figure 7 therefore assumes that tone cycles coincide with syllable cycles.

A synchronized phase relation means that the movement toward any target does not start until its cycle begins. In other words, as suggested by Figure 7, the syllable, the first segment of the syllable and the tone associated with the syllable all start their movement toward their respective targets at phase angle 0°. 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 360° or 720°, where the syllable ends, as illustrated in Figure 7. 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. This is despite the fact that the voluntary force is pushing toward realizing each target as fully as possible. As a result, various scenarios may occur:

1. There is plenty of time for a target, e.g., a vowel is given well over 100 ms, or a tone is given well over 200 ms — The target value would be attained before the end of its allotted time; and a pseudo steady state at that value might be achieved. An illustration of this scenario is shown in Figure 8a.
2. A target is given just enough time, e.g., a vowel is given, say, 75 ms, or a tone is given, say, 150 ms — 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 8b.
3. A target is given insufficient amount of time, e.g., a vowel is given, say, much less than 75 ms, or a tone

[1] Note that C and V here do not overlap in time, which is different from many widely accepted accounts (Fowler 1984, 1986, for example). The discussion of this difference is, however, beyond the scope of the present paper.

is given, say, much less than 150 ms — Within the allotted time interval, again the movement toward the target would proceed continuously; but the target would not be reached 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. An illustration is shown in Figure 8c.

As I will discuss later, these scenarios seem to all 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.

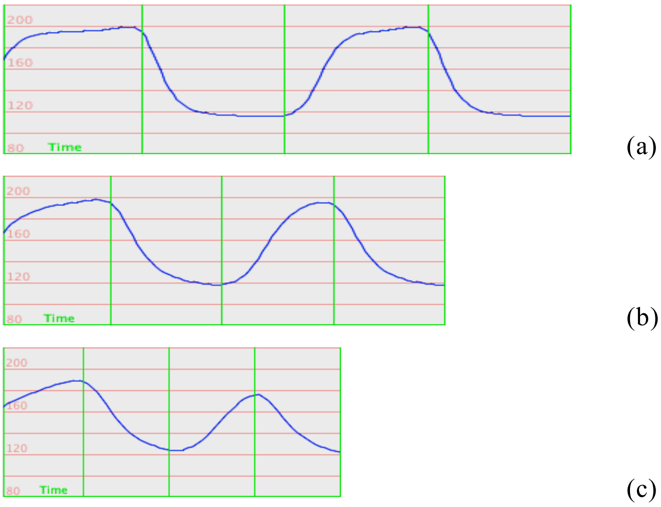

Figure 8. Degrees of attaining the targets in a [high low high low] sequence given different amount of allotted time. (a) Excessive time: full target attainment with sustained pseudo steady state. (b) Just sufficient time: virtually full target attainment but without steady state. (c) Insufficient time: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

4.2. Types of pitch targets

There is no requirement in this model that all pitch targets have to be static, mono-valued ones such as [high], [low] or [mid]. For some tones, we have observed that static targets cannot produce pitch contours similar to the observed ones. There seem to be also 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 is illustrated in Figure 9 together with a static [low]. The slanting dashed line in syllable 1 represents the target [rise] associated with, say, Mandarin R. The level dashed line in syllable 2 represents the target [low] associated with, say, Mandarin L. The solid curve again 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 F_0 peak often seen in connection with R (Xu 1997, 1998, 1999) results directly from implementing a [rise] which is 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.

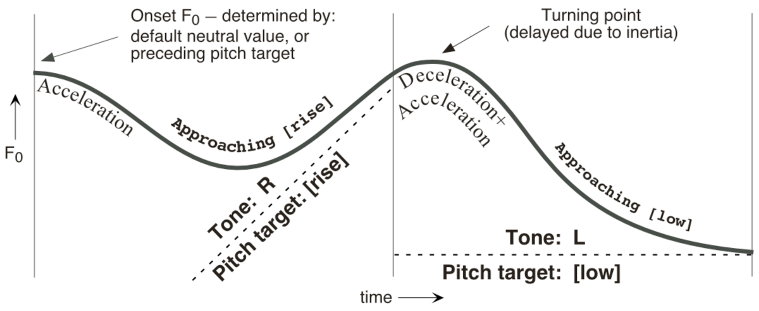

Figure 9. Dynamic and static targets and their implementation. The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from asymptotic approximation of the pitch targets.

Besides the simple pitch targets discussed so far there can be also 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 said in isolation or in a pre-pausal position. An isolated or pre-pausal syllable is often long enough to allow 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 form of the long L in Mandarin.

5. IMPLICATONS

It could be argued that the pitch target implementation 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 simply reduced to a bunch of physical factors. What I will show in the rest of the paper is that to understand tones, language specific factors need to be considered along with the physical factors that I have been discussing. And, as I will show, many tonal phenomena are better understood as resulting from speaker's effort to implement simple pitch targets associated with lexical tones under various articulatory constraints. In other words, it is largely the interaction of voluntary and involuntary forces that gives rise to many complicated surface forms that have been reported in the tone literature.

5.1. Articulatory basis for tone spreading

One of the classic universal tonal phenomena proposed by Hyman and Schuh (1974) is tone spreading. Since then, spreading has been widely accepted as an important phonological process for many contextual tonal variations. With spreading a tonal element such as H or L may spread, mostly rightward, into an adjacent tone-bearing unit. Although it is widely suspected that spreading may have some phonetic motivations due to its assimilatory nature, exactly how that is related to articulatory constraints has remained unclear. In the following, I will try to apply the pitch target implementation model to several spreading related phenomena and see whether they can be explained by the model. Since this kind of explanation is dependent on acoustic details, I will limit my discussion to spreading in Yoruba, for which there are published acoustic data available.

In Yoruba, a tone sequence L H is said to be realized as L LH and a tone sequence of H L is said to be realized as H HL, where LH and HL stand for rising and falling contours, respectively (Hyman & Schuh 1974), as exemplified in (5). To explore the phonetic basis for the tone spreading phenomenon, I will examine two behavior patterns: (1) direction of spreading, and (2) alignment of turning points.

(5) ala (LH) ➔ ala (L LH) 'dream'
rara (HL) ➔ rara (H HL) 'elegy' (from Akinlabi & Liberman 2001)

5.1.1. Direction of spreading

According to Hyman and Schuh (1974), spreading is mostly from left to right. This is also the case with Yoruba. If we think about the phonetic process involved in spreading only in terms of assimilation, it is difficult to see why it is only from left to right. In terms of the model presented in 4., however, this directional asymmetry is quite natural. In the model, a pitch target associated with a tone is continuously approximated within the time interval allocated to the target. As a direct consequence, a pitch target is best approximated near the end of its allocated time interval. A further consequence is that the following pitch target always has to start from the ending F_0 of the previous target. The transition from this ending F_0 toward the current pitch target would naturally reflect the pitch characteristic of the preceding target. This can be seen clearly in the Mandarin cases shown in Figure 3 and 4. To reverse the direction of spreading, e.g., to generate surface F L instead of H F would violate the synchronization constraint: shifting the pitch target phase 180º (or other degrees) ahead of the syllable phase. So, the synchronization constraint and the speed constraint jointly determine that the direction of the spreading has to be right-ward.

5.1.2. Alignment of turning point

Those familiar with the acoustics of Yoruba would immediately point out that the assumption of synchrony of tone and syllable is problematic. This is because in Yoruba spreading is reported to occur only in H L or L H, but not in any combination of H or L with M. Acoustically, there is a clear alignment difference between the spreading case and a non-spreading case. In Figure 10, for example, the fall toward L does not start until the release of the medial sonorant in H L, whereas in M L the fall is already well underway at the release of the medial sonorant. Furthermore, F_0 in the L-carrying syllable never reaches a steady state, while a steady state seems to have been achieved in all the M-carrying syllable in Figure 10.

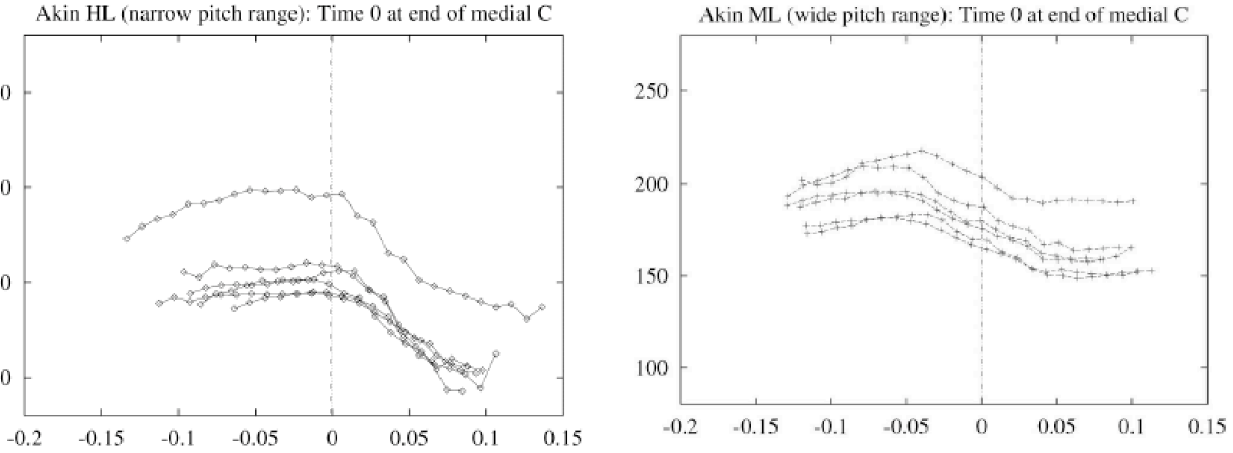

Figure 10. F_0 tracings of H L (left) and M L (right) sequences in Yoruba, presented in Akinlabi and Liberman (2001). The vertical line in the middle of each graph indicates the end of the medial sonorous C. The H L cases were spoken in a relatively narrow pitch range, while the M L examples were spoken in a relatively wide pitch range. (Courtesy of the authors).

What may not be immediately apparent in Figure 10 is that, although the contour shapes of H L and M L seem similar, the magnitudes of the pitch movements are actually quite different: about 7 semitones in H L but only 3 semitones in M L. This despite the fact that the M L cases were spoken in a relatively narrow pitch range, while the M L examples were spoken in a relatively wide pitch range (Akinlabi & Liberman 2001). Applying equation (4) with these magnitudes, we get

$t = 89.6 + 8.7 \times 7\ semitone\ = 141\ ms$ H L

$t = 89.6 + 8.7 \times 3\ semitone = 118\ ms$ M L

The 141 ms in H L means that even if the entire second syllable is used for making the pitch fall of 7 st, at this duration, only a sharp fall can be seen. In contrast to the case of H L, the minimum time needed for the M-L transition is 23 ms shorter than that needed for the H-L transition.[2] This means that there is a better chance to complete the shift, i.e., to start from a plateau or a turning point and end at a different plateau or turning point.

The above computation only addresses the question of whether a steady state can be reached in the second syllable. It does not directly address the alignment difference, however. In particular, there should be an explanation as to why the F_0 turning point in the H L sequence occurs in the L-carrying syllable. Acoustic data reported by Laniran (1992) may provide some clue. Similar to the left graph in Figure 10, the highest F_0 in the L H L sequence in Laniran (1992) also usually occurs in the early portion of the second L-carrying syllable. An example is shown in Figure 11a. However, when the tone sequence is L H H L, or when there are even more consecutive H's before the final L, in most cases, the highest F_0 no longer occurs in the second L-carrying syllable, but inside the last H-carrying syllable (Laniran, 1992: Figures 2.3, 2.4, pp. 34-35, Figures 2.11-2.12, pp. 42-43, and Figures 3.4-3.8, pp. 61-64). An example is shown in Figure 11b. Since the combined duration of H H should be above 200 ms, as Figure 11 indicates, according to equation (3), there should be sufficient time to complete the pitch movement even if the shift magnitude is as large as 7 st. This means that the delay of F_0 movement toward the next H or L when there is only one L or H before it may be due to the short duration of the "spreading" tone. As soon as the duration is no longer short by virtue of having an identical tone before it, there does not seem to be a clear delay of the turning point any more.

[2] Had the pitch ranges the words in the H L and M L cases been similar, the difference in magnitude of the pitch movements as well as the minimum time needed for the movements could have been even greater.

The duration sensitivity of turning point alignment has also been seen in Mandarin, where in an L H L sequence the F_0 turning point usually occurs before the L-carrying syllable (Xu 1997, 1999). Xu (2001b) shows, however, that the turning point in L H L in Mandarin can be forced to delay into the second L-bearing syllable by increasing speaking rate, thus reducing the syllable duration. An example is shown in Figure 11c. The three curves in Figure 11c are mean F_0 curves produced by one of the speakers at slow, normal and fast rate. each averaged across 7 repetitions. The vertical line indicate the onset of the initial nasal in the last syllable which carries L. Note that at the slow rate, the turning point is clearly in the H-carrying syllable. At normal rate, the turning point occur around the H-L border. At the fast rate, the turning point occurs in the beginning of the L-carrying syllable. Note also the similarity between Figure 11a and the fast rate case in Figure 11c, and the similarity between Figure 11b and the slow rate case in Figure 11c. For the subject shown in Figure 11c, his average duration of syllables 1 and 2 combined is 221.6 ms at the fast rate but 418.2 at the slow rate. This also seems to agree with what equation (3) would have predicted, i.e., a pitch rise of 8 semitones needs at least 159 ms to complete.

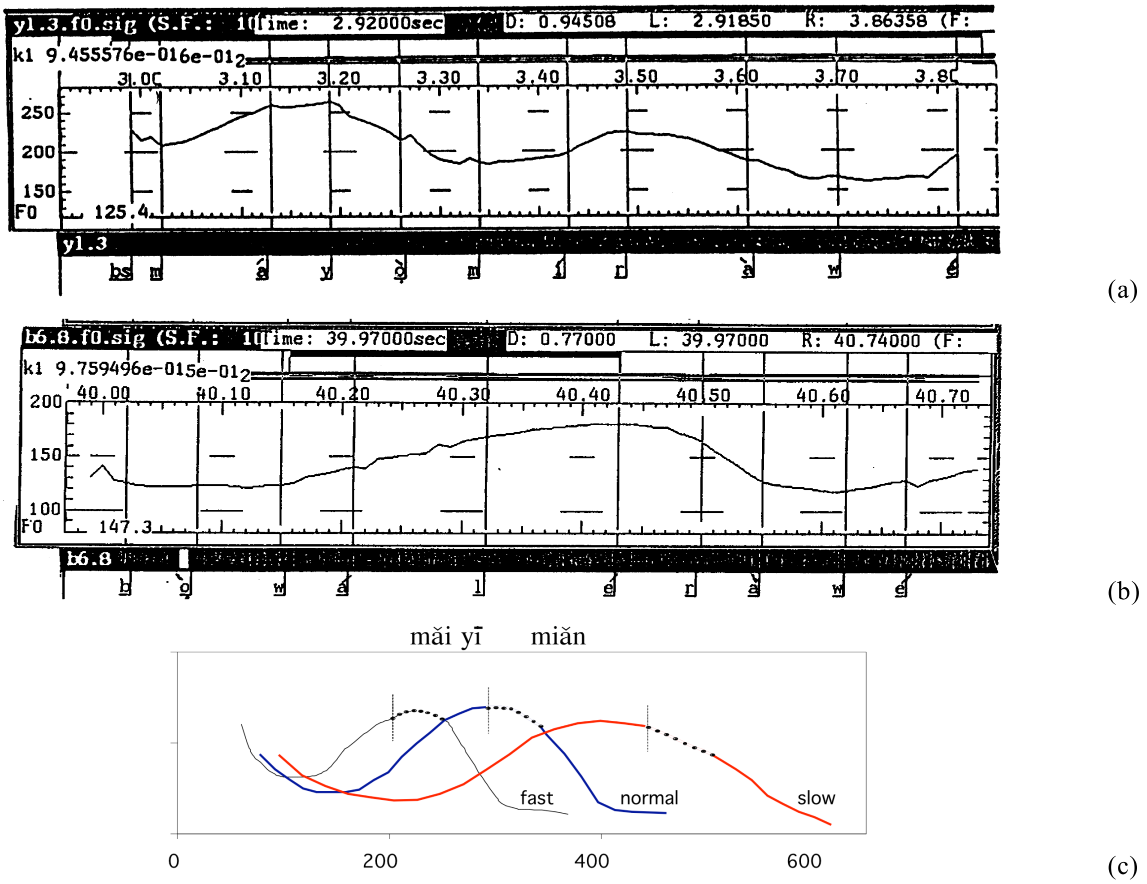

Figure 11. (a) F_0 tracing of a H L H L H sequence in Yoruba (Laniran 1992; courtesy of author). (b) F_0 tracing of a L H H L H sequence in Yoruba (Laniran 1992; courtesy of author). (c) Mean F_0 contours a L H L sequence in Mandarin spoken at fast, normal and slow rate. The gap between time 0 and the 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).

Despite the evidence discussed so far, however, I still cannot conclude whether or not the delay of the high-low transition in L H L in Yoruba is deliberately produced. But I think I have raised enough doubt to call for rigorously controlled experiments to settle the issue of alignment in Yoruba and other languages having similar phenomena. Hopefully, the articulatory and perceptual factors discussed in this paper can be helpful both in designing the experiments and in interpreting the data.

5.2. The nature and distribution of contour tones

Many tone languages are known to have contours tones. But the nature of the contours 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 will try to apply the pitch target implementation model to the case of contour tones and see if it can shed some new light on this old problem.

5.2.1. Two consecutive static elements or a single dynamic element

There has been a long standing debate over whether a phonetically dynamic F_0 contours corresponding to a intonational component is composed of a single contour component such as rising or falling or successive level elements such as high-low or low-high. The more traditional views of intonation describe intonation patterns as consisting of both pitch levels and pitch contours such as rise and fall (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 studies (e.g. Pike, 1948; Wang, 1967; Abramson, 1978) argue that contour tones found in languages such as Thai and Mandarin should be considered as single units, others (e.g. Woo, 1969; Leben, 1973; Gandour, 1974; Duanmu, 1994b) treat contour tones as sequences of H and L targets.

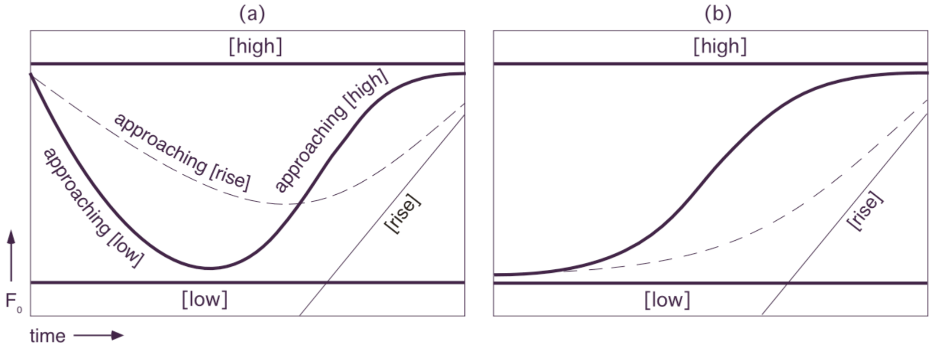

Figure 12. Hypothetical F0 trajectories (curved lines) that asymptotically approach either two consecutive static targets (solid line) or a single dynamic target (dashed line). In (a) the previous tone ends high, while in (b) the previous tone ends low.

As anyone who has 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 consists 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 acoustic pattern in speech. To reexamine the issue of contour vs. level, I would therefore like to apply the pitch target implementation model, because it incorporates the major articulatory constraints relevant for F_0 contour production.

According to the model, if we assume that a particular tone consists of two elements, then each of them is a target by its own right. (Note that the in-phase requirement wouldn't prevent two static pitch targets from being fit into one syllable.) When there is sufficient time assigned to the consecutive targets, there should be a clear transition between the two elements. By sufficient I mean 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 thick solid curve in Figure 12a, assuming that the previous tone happens to end high. If, however, the target itself is dynamic, like the dotted slanting line in Figure 12a, 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 12b, respectively. As shown in Figure 11c, 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 indeed resembles the solid curve in Figure 12a.

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 12a or Figure 12b. Rather, it is mostly like the thick curve in Figure 13, which is more like the dashed curve in Figure 12b (or 12a 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 14, and that the maximum velocity of the rise does not change systematically with variations in 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.

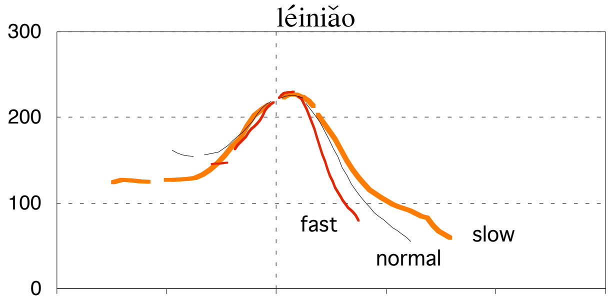

Figure 13. Mandarin tone sequence R L spoken at three different rates. The curves are aligned to the onset of [n] in the second syllable as marked by the vertical line. Each curve is an average of five repetitions. Data from one subject in Xu (1998).

5.2.2. 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) 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." (p.331) Gordan (2001) 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.” (abstract). What is still not yet clear from these mostly typological studies is the exact phonetic mechanisms for this duration dependence. So far, most accounts are treating perception as the determinant factor. The discussion about articulatory constraints, however, suggest that we should also take a closer look at the articulation process as a possible contributor. To do that, again I will put things into the pitch target implementation model.

According to the model, the implementation of a pitch target starts only when the implementation of the previous target is over. This assumption would put a lot of strain on the implementation of a contour tone. This is because the offset F_0 generated by the preceding tone may or may not be close to the initial target 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 12a 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 internal to the contour tone itself. This situation would arise whether the contour tone consists of two static elements or a single dynamic element. Having recognized this situation, we may calculate how much time it would take an average speaker to complete two consecutive pitch movements, using equations (3) and (4). For a down-up or up-down movement the size of 4 semitones,

$$t = 89.6 + 8.7 \times 4 + 100.4 + 5.8 \times 4 = 248 \text{ ms}$$

Anyone who has looked into the issue of how duration relates to contour tones would recognize this 248 ms as quite long. This is because in syllables that can carry contour tones, the vowel or vowel plus sonorant coda, which is generally believed to be the tone bearer, is often much shorter than 248 ms. The left half of Table 1 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 and Sun (2002) still would need 196 ms to complete shifting pitch up and down or down and up by 4 semitones (cf. Table V in Xu & Sun 2002), which is still longer than most of the durations in Table I. This discrepancy should become less puzzling in light of the pitch target implementation model described in 4. According to the model, the implementation of a pitch target associated with a tone would coincide with the implementation of the syllable. 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

[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 cause the problem being discussed here in the first place.

found in Xu, Xu and Sun (2002), the effect of voiceless consonants such as stops and fricatives is to introduce rather local perturbations without changing the carryover or anticipatory tonal variations reported in previous studies such as Xu (1997, 1999). Figure 15 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 amount of local raising at the voice onset. Nonetheless, if we ignore these very local effects, the F_0 curves in /da/, /ta/ and /sha/ are 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, we have good reasons to assume that the interval corresponding to C in Mandarin is also used for implementing the tonal targets. With C included, the total duration available for tones, i.e., the duration of the entire syllable for Mandarin and Shanghai 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.

	Gordon (1999)	Zhang (2001)	Xu (1997)	Xu (1997)	Xu (1999)	Duanmu (1994a)
Hausa	133	109				
Navajo	173	209 (rime)				
Luganda		179				
Xhosa		212				
Mandarin		151/231	122/140 (R) 115/135 (F)	185/196 (R) 183/193 (F)	198 (R) 184 (F)	215
Shanghai						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 and Sun (2002) is for a complete pitch shift, starting at 0 velocity and ending at the next 0 velocity. As can be seen in Figure 15 as well as in Figure 3, the movements in R and F

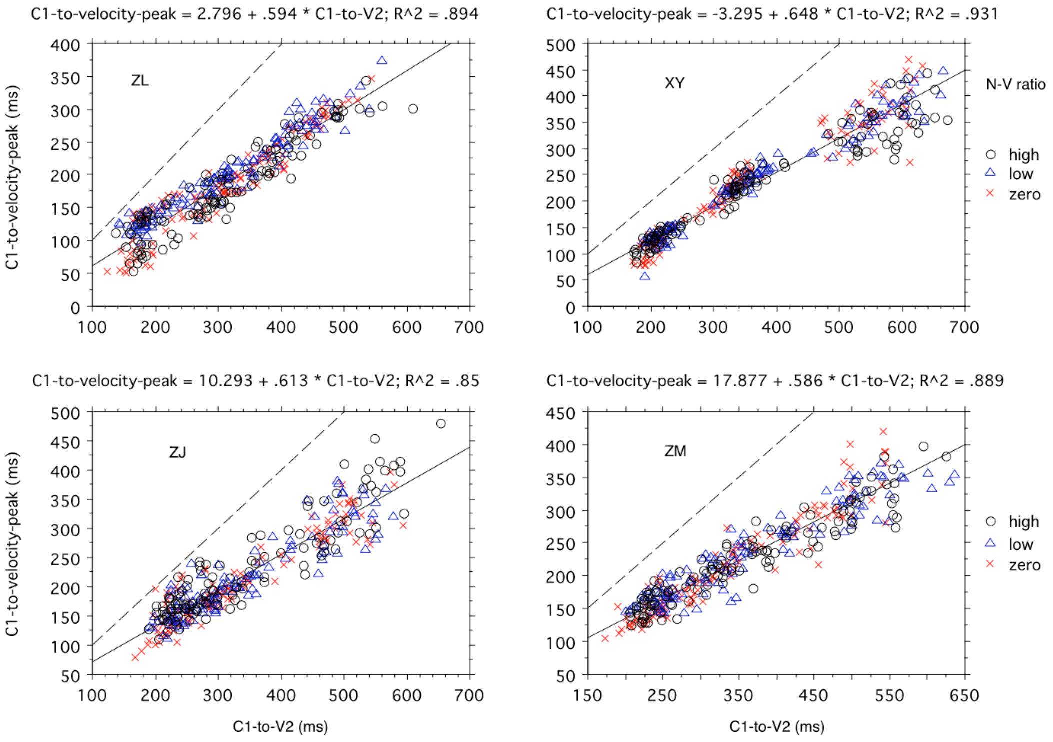

Figure 14. Location of peak velocity of F_0 rise relative to the onset of syllable 1 (y-axes) plotted against the distance between the onset of syllable 1 and the S-V boundary in syllable 2 (x-axes) for each subject. The plotting symbols represent different N-V ratios. The dashed line has a slope value of 1 and an intercept of 0. (Xu 1998)

have 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 only just enough time for a contour tone to be effectively implemented. This observation is corroborated by the finding of Xu and Sun (2002) that in Mandarin, it is precisely during the production of the dynamic tones that the maximum speed of pitch change is actually approached.

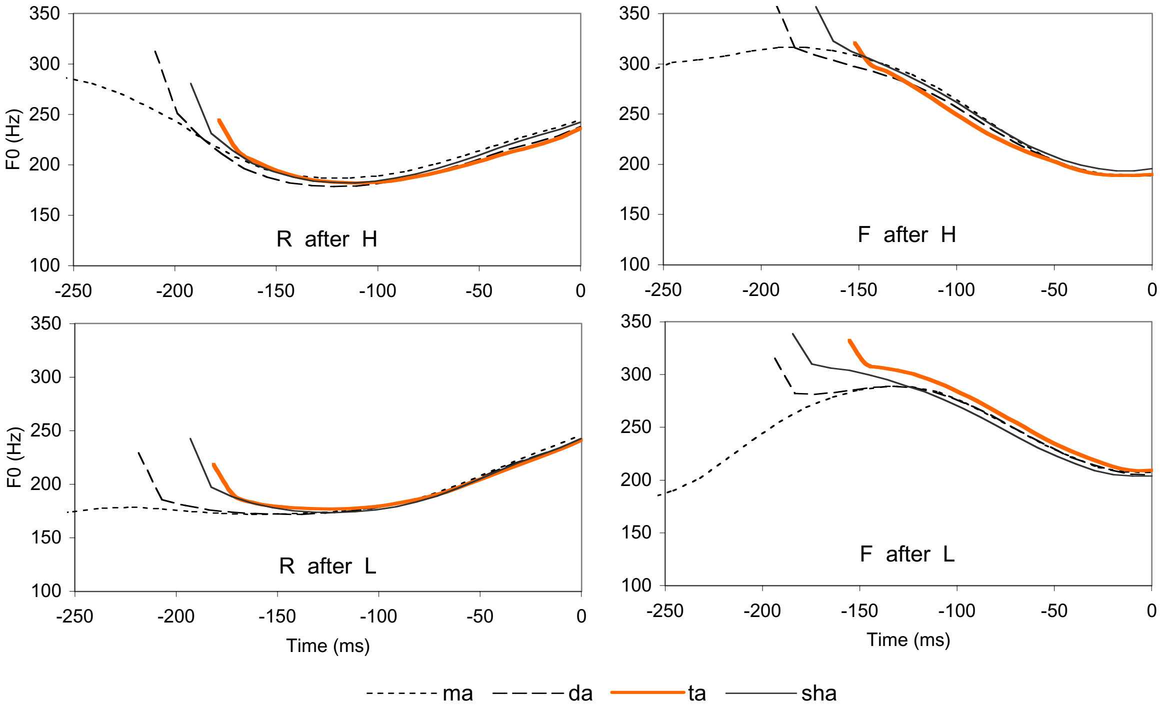

Figure 15. Effects of voiceless consonants on the F_0 contours of Mandarin R and F produced after H and L. Each curve is an average across 5 repetitions, 2 carrier sentences and 7 female speakers. All curves are aligned to the syllable offset.

If, as I have shown, even in Mandarin, where dynamic tones remain functionally intact in connected speech, dynamic tone can often be realized only with speakers' best effort (Xu & Sun 2002), and sometimes with reduced perceptibility (Xu 1994), with shorter syllable durations such as those in Shanghai (162 ms, Table 1, right most 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 transmitability of the dynamic tones on short syllables, it is probably the articulatory constraints that make the implementation of pitch contours impossible in the first place.

As for the perception account, findings by Janse and colleagues (Janse et al. in press; Quené & Janse, 2001) as discussed in 3.3. may provide some indirect argument against it. Recall that what they found was that natural-fast speech actually had lower intelligibility than the linearly time-compressed normal speech. According to their report, the human perception system can easily handle resynthesized speech 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 probably contained too much undershoot for the perceptual system to “unwind” successfully. This would agree with the finding of Xu (1994) that Mandarin listeners could not fully compensate for the flattening of R and F due to contextual variations even when the utterances were spoken at a normal rate. It therefore seems that articulatory constraints are probably the real barrier to maintaining contour tones in very short syllables.

Finally, the recognition of duration sensitivity of contour tones also confirms further the strict phase relation between laryngeal and supralaryngeal movements. This is because, as I have 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 adjust the micro-alignment of tone relative to the syllable, the occurrence of the contour tones would not have been so restricted by syllable duration.

6. CONLCUSIONS AND FUTURE DIRECTIONS

I have argued in this paper that tone research has to take articulatory constraints seriously. This is because not only is the production of tones subject to various physical limits just as the production of vowels and consonants is, but also the constraints on tone are often probably much *greater* than those on vowels and consonants. In particular, I have shown that the maximum speed of pitch change has recently been found to be slower than previously thought, and that, as a result, syllables can often be shorter than the minimum time it takes to complete a pitch change of only a few semitones. I have also shown that the constraint of synchronizing laryngeal and supralaryngeal movements is probably even stronger than the constraint of maximum speed of articulatory movement, because it seems to prevent micro-alignment adjustment from happening even if target undershoot would result without such adjustment. In contrast to the newly established articulatory constraints, I have shown evidence that the human perceptual system is actually very proficient in handling fast acoustic events as well as phonetic deviations due to articulatory constraints, although it does seem to prefer less undershoot rather than more undershoot. 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 pitch target implementation 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 voluntary and involuntary forces, i.e., between underlying pitch targets associated with lexical tones and physical constraints that are inherent to the articulatory system.

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 consisting 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. In regard to tone spreading, whether an underlying target is indeed shifted into the next tone-bearing unit can be also examined by slowing down the speech rate, and by conducting perception tests in which the subjects' tasks are linguistic rather than psychoacoustic in nature. Similar experiments can be also designed to investigate the nature of the floating tones, downstep, etc.

7. REFERENCES

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*. T. L. Thongkum, P. Kullavanijaya, V. Panupong and K. Tingsabadh. Bangkok, Chulalongkorn University Press, pp. 1-9.

Abramson, A. S. (1978): The phonetic plausibility of the segmentation of tones in Thai phonology. *Proc. The twelfth Int. Congr. Ling.*, Vienna, pp. 760-763

Akinlabi, A. and Liberman, M. (1995). On the phonetic interpretation of the Yoruba tonal system, *Proceedings of The 13th International Congress of Phonetic Sciences*, Stockholm, pp. 42-45.

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

Caspers, J. and van Heuven, V. J. (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*. University of California Press, Berkeley, CA.

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

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

Duanmu, S. (1994b). Against Contour Tone Units. *Linguistic Inquiry* 25: 555.

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 of speech processes*. J. S. Perkell and D. H. Klatt. Hillsdale (eds.), NJ, LEA, pp. 123-139.

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*. Ph.D. dissertation, UCLA.

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

Greenberg, S. and Zee, E. (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.

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

Janse, E., Nooteboom, S. and Quené, H. (in press). Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Communication*.

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

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America* 53: 8-16.

Ladd, D. R. (1978). The *Structure of Intonational Meaning*. Cornell University, Ithaca.

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

Leben, W. R. (1973). *Suprasegmental Phonology*, Massachusetts Institute of Technology.

Lee, C.-Y. (2001). *Lexical Tone in Spoken Word Recognition: A View from Mandarin Chinese*. Ph.D. Dissertation. Brown University.

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

Lin, M. and Yan, J. (1991). Tonal coarticulation patterns in quadrisyllabic words and phrases of Mandarin. *Proceedings of the 12th International Congress of Phonetic Sciences*, Aix-en-Provence, France, 3: 242-245.

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

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

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

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

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

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

Quené, H. and Janse, E. (2001). Word perception in time-compressed speech. *Journal of the Acoustical Society of America* 110: 2738.

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

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

Shih, C.-L. (1988). Tone and intonation in Mandarin, *Working Papers, Cornell Phonetics Laboratory* No. 3: 83-109.

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

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

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

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

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

Wang, W. S. Y. (1967). Phonological features of tone. *Int. J. of Am. Ling*. 33: 93-105.

Woo, N. (1969). *Prosody and phonology*. Ph.D. dissertation, Massachusetts Institute of Technology.

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

Xu, C. X., Xu, Y. and Sun, X. (2002) Consonant perturbation on F_0 in English and Mandarin. Presented at the 9th Meeting of ICPLA, Hong Kong.

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. *Journal of Chinese Linguistics,* monograph series #17: 1-31.

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

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

Xu, Y. and Q. E. Wang (1997). What can tone studies tell us about intonation? *Intonation: Theory, Models and Applications, Proceedings of an ESCA Workshop. European Speech Communication Association*. A. Botinis, G. Kouroupetroglou and G. Carayannis. European Speech Communication Association, Athens, Greece, pp. 337-340.

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

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

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