Articulatory Constraints and Tonal alignment

Yi Xu

Department of Communication Sciences and Disorders
Northwestern University, Evanston, IL 60208, USA
xuyi@northwestern.edu

Abstract

There has been accumulating evidence in recent years that certain F_0 events are consistently aligned with segmental events such as the syllable boundary. The mechanisms for the observed alignment patterns, however, are still being closely investigated. In this paper I argue that to understand the observed tonal alignment patterns, it is imperative to first understand the role of articulatory constraints in shaping the F_0 contours in speech. In particular, the maximum speed of pitch change limits how fast F_0 movements can be produced; and the coordination of laryngeal and supralaryngeal movements limits how syllables and tones can be aligned to each other. From a different perspective, these constraints mean that the degrees of freedom speakers have are probably less than previously thought. This may actually make our understanding of the speech signal somewhat easier than before. I will demonstrate this with a theoretical model of F_0 production that is based on the new understanding of the articulatory constraints. Though conceptually simple, the model seems to be able to account for a number of phonetic patterns that have been observed in speech. Finally, I will briefly discuss the implications of the new insights on our understanding of tonal perception in speech.

1. Introduction

When we speak, we need to control both our laryngeal and supralaryngeal movements. Acoustically, much of the laryngeal movement is manifested as fundamental frequency (F_0) contours, and much of the supralaryngeal movement is manifested as spectral patterns.[1] Conceptually, these two controls could be quite independent of each other except for certain local effects such as consonantal perturbation of F_0 ([8], [10], [16], [21], [27]) and vowel intrinsic pitch ([16], [28]). In other words, articulatorially, any pitch value could be freely associated with any voiced segment. If so, we have a kind of "freedom of articulatory association." Assuming that there is indeed such freedom, then, most, if not all, observed F_0 patterns and their alignment should only result from various choices languages make, or the speakers of the languages make. That is to say, whenever an F_0 pattern is observed, we should always look for its source from linguistic/paralinguistic specifications, from psycho-acoustic constraints, or from some idiosyncratic choices of the speakers. In general, this indeed seems to be the consensus in the area of speech prosody. For example, the following implicit assumptions are often observed in the prosody literature.

A. Most, if not all, observable F_0 contours of an utterance are intentionally produced by the speaker. Thus, for instance, an observed peak is intended as a peak and a valley as a valley; likewise, an observed rise is intended as a rise and a fall as a fall.

B. All detailed alignment of an F_0 contour relative to segmental units is intentionally produced by the speaker. Thus, for instance, if a peak is observed to be aligned to a certain location in a syllable, it is intended to be aligned there.

C. If something is not observed in the acoustic signal, either it is linguistically unimportant, or it is deliberately avoided due to perceptual constraints

D. If something seems to have moved or extended, it is intended to be moved or spread. Thus, for instance, if a syllable takes on the pitch value of a preceding tone, that value is spread from the preceding syllable to the current syllable.

All these assumptions are based essentially on the general assumption that laryngeal movements can be freely associated with supralaryngeal movements. We may refer to this assumption as the *Free-Articulatory-Association assumption*. In this paper, I will argue that we should give up this *free-articulatory-association assumption* in our investigation and understanding of speech prosody in general, and tonal alignment in particular. I will show that the articulatory mechanisms involved in the production of F_0 contours in speech in fact impose many limitations on the way F_0 contours can be produced, and that these limitations may account for a large portion of the F_0 contour variations in the speech signal. I will argue that we should make a clear distinction between F_0 variations that are due to articulatory constraints and those that are due to deliberate implementation of communicatively functional pitch targets. As a first step toward incorporating this new understanding into a theoretical framework, I will present a pitch target model of F_0 contour production, which consists of a) assumptions about articualtory constraints, b) underlying pitch targets, and c) implementational effort.

1. Articulatory constraints on F_0 movement

There are many articulatory constraints on the production of F_0 in speech. It is well known that certain consonants

[1] I am putting aside laryngeal movements due to phonation type variations, and laryngeal-supralaryngeal interaction that result in VOT patterns, etc.

raise or lower F_0 of adjacent vowels ([8], [16]). It is also well known that vowel height causes F_0 to be higher or lower ([16], [28]). However, articulatory constraints on F_0 movement per se has not been seriously considered until recently. In the following, I will discuss two major types of articulatory constraints on the movement of F_0 in speech. The first is the maximum speed of pitch change, and the second is the laryngeal-supralaryngeal coordination.

1.1. Maximum speed of pitch change

Few people would actually assume that we can change pitch at whatever speed we want. However, the actual limit on speed of pitch change was seriously investigated earlier only twice, as far as I am aware, first by Ohala and Ewan [19] and then by Sundberg [24]. Both studies used similar methods — asking the subject to change from one pitch level to another as fast as possible. This method, however, makes it difficult for the researcher to determine the exact time when the pitch shift begins and when it ends. Probably because of this, in both studies, only the time it took for the subject to complete the fastest middle 75% of the pitch shift was measured. These measurements, however, have often been taken as the time for making complete pitch shifts. Consequently, the maximum speed of pitch change is often believed to be faster than it actually is. And, when the fastest speed of pitch change in real speech is found to be much slower than the exaggerated speed, it is speculated that the cause for the "slow" speed in speech must lie somewhere else: most likely, in perception ([3], [25]).

In a recent study, Xu and Sun [36] reassessed the maximum speed of pitch change using a different method. Subjects produced rapid pitch shifts by imitating a series of model pitch undulation patterns as shown in Figure 1. Because the model pitch undulation patterns are faster than what human speakers could achieve, there was virtually no steady-state plateaus in the F_0 contours produced by the subjects, as shown in Figure 2. This permitted the measurement of the entire duration of each pitch shift as opposed to the time corresponding to only 75% of the pitch change as measured in previous studies [19], [24]. As it turned out, it took our subjects nearly twice as long to complete an entire pitch shift as to execute the middle 75% of the shift.

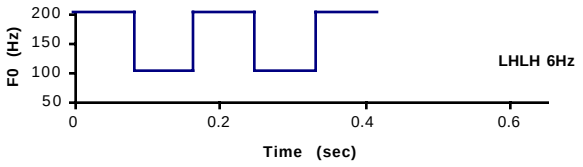

Figure 1. One of the model undulation patterns used in [36].

In the same study [36], we also found that the maximum speed of pitch change varied quite linearly (at least, for pitch shift greater than 2 semitiones) with the size of the pitch shift, and the relations can be represented by the following equations.

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

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

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 st. With (1) and (2) it can be computed that when 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 [25]. Also, the fastest pitch change speed reported by [3] was found to be comparable to the maximum speed of pitch change at similar pitch shift intervals reported by [36]. The maximum speed of pitch change reported by [36] also matched the speed of pitch change in the dynamic tones (Rising and Falling) in Mandarin recorded in [33]. For English, [25] reported 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.

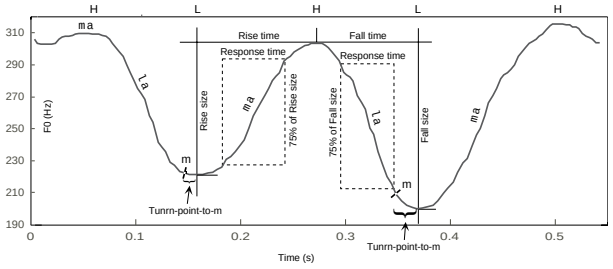

Figure 2. Illustration of measurement of rise and fall excursion time, rise and fall "response time", and turn-point-to-m in a HLHLH trial spoken with /malamalama/ in [36]. See original paper for more detailed explanations.

In [36] we also obtained the average minimum time needed to complete a pitch rise or fall as a function of the size of pitch change, as shown in Equations (3) and (4).

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

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

Here t is the amount of time it takes to complete the pitch shift, and d is the size of pitch shift in semitone. Equations (3) and (4) provide a useful tool with which we can estimate, in each specific case, how much F_0 variation due to the constraint of maximum speed of pitch change is inevitable. For example, according to (3) and (4) it takes at least 124 ms for an average speaker to complete a 4-st pitch rise or fall, and at least 142 and 135 ms to complete a 6-st pitch rise and fall, respectively, and 194 ms and 170 ms to complete a 12-st pitch rise and fall, respectively. As will be discussed later, these numbers indicate significant articulatory constraints on the shape and alignment of F_0 contours in speech.

1.2. Coordination of laryngeal and supralaryngeal movements

There is little doubt that F_0 and spectral patterns are separately controlled in speech. Without such separation, things like lexical tone and intonation would not be possible. However, separation of controls does not necessarily mean total independence from each other. For example, can a language freely put whatever pitch value at whatever location along a syllable sequence? Or, can a speaker freely adjust the micro-alignment of F_0 contours relative to a segmental sequence? If the answer is no, what are the real constraints that prevent the change of alignment from happening? Are they coming from linguistic/phonetic specifications, from perceptual limitations, or from articulatory limitations? These questions all need to be answered in our quest for true understanding of F_0 contour formation in general, and tonal alignment in particular.

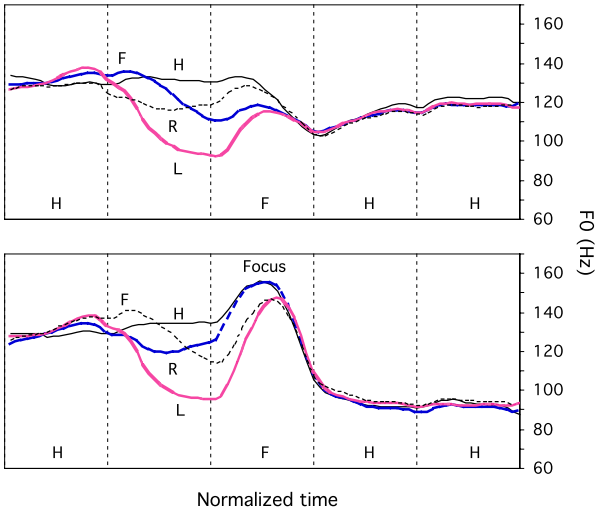

Figure 3. Mandarin F tone following four different tones. Top: no narrow focus in the sentence; Bottom, focus on the F-carrying syllable. Data from [33].

1.2.1. What is the nature of tonal alignment?

Ever since the research by Bruce and Garding ([2], [7]) on Swedish prosody, there have been many studies on the alignment of F_0 turning points with certain segmental landmarks such as the syllable boundary. In recent years, there have been a number of reports about strict alignment of F_0 events and segmental sequence. As has been observed, certain F_0 events such as F_0 peaks and valleys maintain a relatively stable alignment with the onset or offset of the syllable [1], [3], [12], [14], [20], [32], [33], [34]. In English, Greek and Dutch, the most consistent alignment is reported for the start of the F_0 rise in a prenuclear accent. It is found to occur quite regularly at the onset of the accented syllable [1], [14], [3]. In Mandarin and Chichewa, F_0 peaks are found to be consistently aligned with the offset of the tone-bearing syllable in certain situations [12], [32], [33], [34].

It has been argued that these patterns indicate that F_0 turning points are linguistically meaningful targets and are "anchored" by speakers at the onset or offset of the syllable [1], [13], [14]. In the ToBI notation, this type of prenuclear accent is labeled as LH*. Because the F_0 minimum is found to always occur at the onset of the accented syllable, it seems natural to assume that the L tonal point is deliberately placed or "anchor" there, and therefore it should belong to the accented syllable, or at least constitute part of the accent.

What we have learned from the findings about the maximum speed of pitch change, as discussed in the previous section, is that no matter what form the linguistically meaningful targets take, implementing them takes time. For example, if the linguistic task is to anchor an F_0 minimum at the onset of a syllable-initial sonorant consonant, an average speaker has to start the F_0 movement toward this low point at least 107 ms earlier even if the range of the movement is just 2 st (from equation (3)). Furthermore, the speaker has to adjust the onset of a pitch movement according to the size of the F_0 excursion toward that low point. Because we have not seen any data showing whether or not this is the case for English, Greek or Dutch, let me first turn to Mandarin for some evidence.

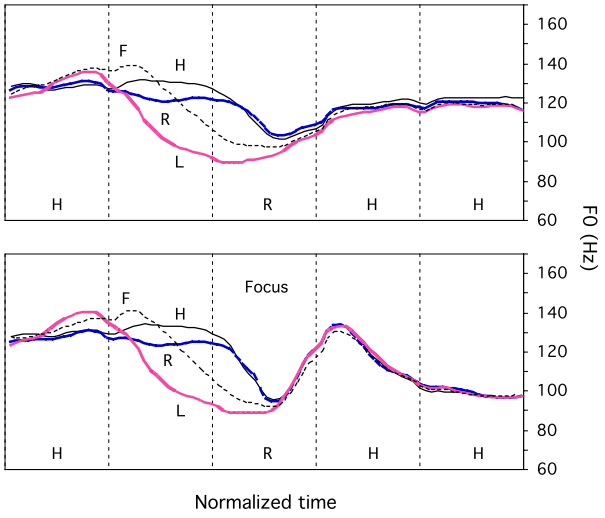

Figure 4. Mandarin R tone following four different tones. Top: no narrow focus in the sentence; Bottom, focus on the R-carrying syllable. Data from [33].

A tone language like Mandarin gives the researcher the advantage of being able to control the tonal environment easily. For example, to test the hypothesis that speakers adjust the alignment of a tonal target according to how much time it takes to complete the transition toward it from the preceding tone, one can simply place the tone after different tones and observe the variation in its F_0 alignment. Figures 3 displays the Mandarin F (Falling) tone when preceded by four different tones: H (High), R

(Rising), L (Low), and F. As can be seen in Figures 3, transitions toward the F tone always start at the onset of the F-bearing syllable regardless of the distance to be covered. This is despite the fact that there is a big articulatory constraint on the maximum speed of pitch change as discussed earlier. As a consequence, the location of the high F_0 turning point varies depending on the ending F_0 of the preceding tone: the lower the value, the later the turning point.[2]

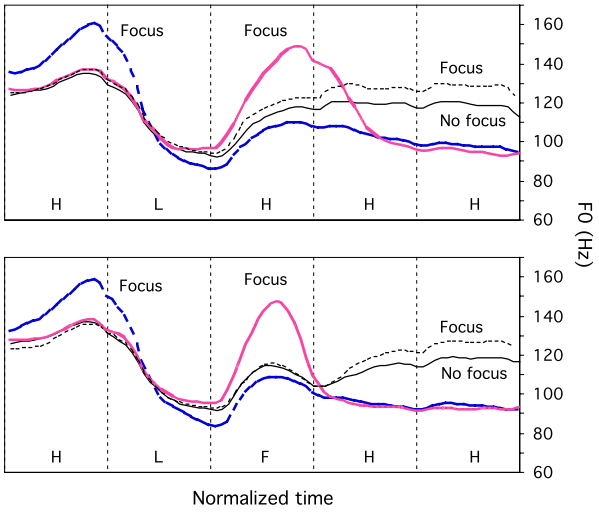

Figure 5. F_0 contours of the H (upper panel) and F (lower panel) tones when different words in a sentence are focused. The location of the focus (if there is one) is indicated by the label placed near the respective curve. Word 1 and Word 3 are disyllabic, while word 2 is monosyllabic. Data from [33].

Even the transition toward an exaggerated F_0 value due to focus does not start earlier, as shown in Figure 5. This suggests that there must be some kind of alignment constraint that is quite strong. Figure 3 also shows that, regardless of the preceding tones, the falling contour of the F tone is always best approximated near the end of the syllable. This is also true for the rising contour in the R tone in Figure 4. What is clearly illustrated by Figures 3-5, therefore, is that the implementation of a lexical tone in Mandarin starts at the onset of the host syllable and ends at the end of the syllable.

Similar evidence can be seen in Yoruba in the acoustic data reported by Laniran [15]. Yoruba is an African language with three lexical tones: H (High), M (Mid), and L (Low) [15]. Figure 6 displays selected F_0 values (two per syllable, the first one near the vowel onset and the second near the vowel offset) of three tonal sequences in Yoruba produced by three speakers. The only difference among the three sequences in each panel is the tone of the fourth syllable, which varies among H, M and L. Similar to comparable Mandarin cases shown in Figures 3 and 4, F_0 does not start to differ until after the onset of the fourth syllable. Also similar to Mandarin, F_0 differences due to the tone of the fourth syllable continue into the fifth syllable, but they gradually reduce within the syllable. Again similar to Mandarin, by the beginning of the vowel in the sixth syllable, the F_0 differences due to the tone of the fourth syllable virtually disappear. In other words, it appears that in Yoruba, too, a tonal target is implemented along with the syllable that carries it: starting as the syllable starts and ending as the syllable ends.

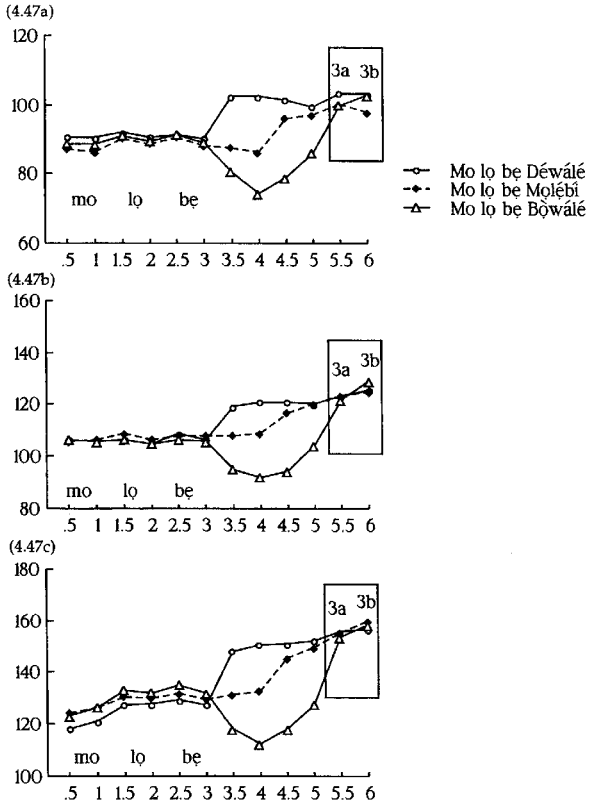

Figure 6. F_0 values (two per syllable, near the onset and offset of the vowel, respectively) of a string of Yoruba syllables in which only the tone of the fourth syllable varies among H, M, and L. Tones of other syllables are held constant. (With permission from the author [15])

1.2.2. Synchronization of tone and syllable

From the fact that the clearest evidence for strict synchronization of tone and syllable reported so far is found in Mandarin, one may say that such synchronization is simply due to a language specific feature: one tone for one syllable. But in fact, most of the phonetic/phonological accounts of Mandarin tones do not

[2] Note that, if we consider the realization of the F tone in the HF sequence to be the most ideal, its realization in the LF sequence can be then considered to be compromised. If, on the other hand, its implementation in the LF sequence is considered to be normal, then the implementation in the HF sequence may be considered to be "hyper-articulated."

interpret lexical association of tone and syllable as meaning strict alignment between the two. Some actually suggest that the tones are carried only by the rhyme [10], [21]. Others suggest that only the nuclear vowel is the tone carrier [17]. There has also been suggestions that tonal alignments vary in different tones [23].

Thus, even if a tone is lexically associated with a syllable, theoretically it does not have to be perfectly aligned with the syllable. Furthermore, due to the limit of maximum speed of pitch change, sometime it is actually quite hard to implement a tone in a "conflicting" [30] tonal context, e.g., in LH, HL, LF, HR, etc. It is imaginable that the speaker may readjust the micro-alignment of a tone to make the transition easier. However, evidence discussed in the last section indicates that speakers do not do that. When the F tone is preceded by the L tone, speakers do not start the transition toward the high pitch of the F tone earlier than when it is preceded by other tones. As a consequence, it is the size of the fall in the F tone that gets compromised, as seen in Figure 3. There must be, therefore, some stronger constraints that prevent the alignment readjustment from happening. There are two potential sources of such constraints. The first is some kind of perceptual constraint that says that a tone has to be best approximated by the end of the syllable. Such a constraint is implied in the tonal perception model proposed by House [9], although not quite directly. That model contends that pitch perception sensitivity is the lowest in zones of spectral discontinuities such as consonant release, vowel onset and rapid formant transitions in the beginning of the vowel. These regions, however, happen to be also near the syllable boundary. As found in [32], the rapid spectral change at the vowel-nasal junction within the rhyme of a syllable did not affect the shape of tonal contours in Mandarin. So, if perception is consistent with production, it is probably the syllable itself rather than its internal structure that determines the optimal regions of tonal perception, assuming that perception *is* the real constraint.

Another potential constraint comes from the fact that tone articulation and syllable articulation are concurrent movements that are controlled by a single central nervous system. To carry out such concurrent movements, as found by a number of studies on limb movements, performers have very few choices in terms of the phase relation between the concurrent movements [11], [22]. At relatively low speed, the phase angle between the two movements has to be either 180º, i.e., starting one movement after the other one 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 [26]. This means that the average syllable duration is about 143-200 ms. According to equations (3) and (4), at the fastest speed of pitch change of an average speaker, it takes at least 124 ms to complete a 4-st rise or fall, and about 107 ms to complete a 2-st rise and 112 ms to complete a 2-st fall. This means that it is virtually impossible for the speaker to maintain a 180º phase angle between pitch movement and the syllable, assuming that the syllable is managed as an articulatory cycle. Therefore, it is very likely that the only choice left for the speaker is to maintain a 0º phase angle between tonal movement and the syllable, i.e., keeping them fully synchronized.

To summarize what I have discussed so far, there appear to be two important sources of articulatory constraints for F_0 production in speech, neither of which has received much attention until recently. The first one is the maximum speed of pitch change, and the second one is the coordination of laryngeal and supralaryngeal movements. The evidence we now have seems to indicate that they both pose strong limits on the formation and alignment of F_0 contours in speech. Their recognition, nevertheless, may make our task of understanding tone and intonation easier than before. In the next section, I will present a model of tonal production that is mainly based on our newly-gained understanding of articulatory constraints.

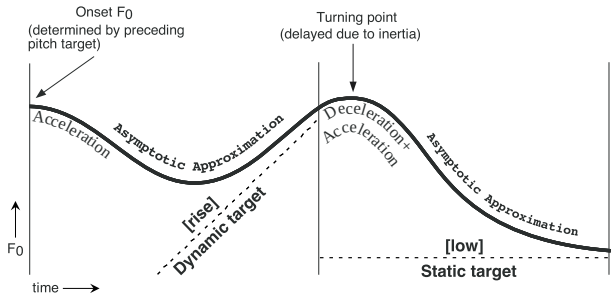

Figure 7. Illustration of the pitch target model. The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from asymptotic approximation of the pitch targets.

2. A pitch target model of F_0 contour formation

2.1. The model

The pitch target model was first proposed in [37] and was further explained in [35]. A quantitative implementation of the model was attempted in [29]. I am presenting the key elements of the model here with certain modifications prompted by a number of recent findings.

The basic assumption of the model is that surface F_0 contours are not linguistic units per se. Rather, they are the surface realizations of linguistically functional units such as tone and pitch accent as a result of their implementation under various articulatory constraints. Furthermore, similar to segmental phonemes, tones and pitch accents are considered to be abstract units, which have to be associated with articulatorially operable units to become functional in speech. These articulatorially operable units are referred to as *pitch targets* in the model. Pitch targets are analogous, to a certain extent, to segmental phones, and so we can symbolically represent them by putting them between square brackets, such as [high], [low], [rise] and [fall].

The basic idea of the model is illustrated in Figure 7. The vertical lines in the figure represent syllable boundaries. The two dashed lines represent the pitch targets [rise] and [low] which are associated with the R tone and L tone in Mandarin. The solid curve represents the F_0 contour resulting from implementing the pitch targets under articulatory constraints.

2.1.1. Pitch targets

Pitch targets can take on various forms. It can be as simple as a static target like the [low] in the second syllable in Figure 7. A pitch target can also be a dynamic one like the [rise] in the first syllable in Figure 7. For a dynamic target, the *movement* itself is assumed to be the goal. A target may also take on other forms, although so far we have found simple static and dynamic ones to be sufficient for Mandarin tones produced in non-final and non-prepausal positions.[3]

A target is associated with each syllable. In Figure 7, [rise] is associated with the first syllable and [low] with the second syllable. Note that the target is only *associated* with the syllable. It is not necessarily strictly synchronized with the syllable. Synchronization is about the physical movements that occur during articulatory implementation. The pitch target model is about how articulorially operable target is physically implemented to generate F_0 contours. During physical implementation of the pitch targets, various articulatory constraints apply. A number of such constraints are incorporated in the model as discussed next.

2.1.2. Articulatory constraints

First, the model assumes that the implementation of pitch targets and their host syllables are synchronized. Due to such synchronization, the implementation of each pitch target does not start until the onset of the host syllable, and it does not stop until the offset of the host syllable, as can be seen in Figure 7.

Second, the model assumes that articulatory transition from one pitch level to another takes discernable amount of time due to the limit of the maximum speed of pitch change [36]. As a consequence, a large portion of the early F_0 contour in a syllable is a transition from the end of the previous pitch target to the initial value of the current pitch target. Even in a 200-ms syllable, the initial transition usually takes up more than half of the syllable duration (substantial F_0 drops in both syllables in Figure 7).

Third, probably to guarantee that the target is not overshot, its approximation is always asymptotic [31], [33]. This is true for both static and dynamic targets. In the latter case, the approximation of the target achieves the highest velocity near the end of the host syllable, as illustrated in Figure 7.

Fourth, inertia of the articulators contributes to certain micro aspects of F_0 shape and alignment. In Figure 7, for example, the F_0 drop from the initial relatively high F_0 due to the preceding target takes some time to accelerate to full speed, resulting in a convex-up shape in the initial portion of the F_0 transition. Also in Figure 7, the approximation of [rise] results in a fast F_0 rise at the end of the first syllable. To implement the [low] in the second syllable, however, F_0 needs to drop quickly. Deceleration of the rising movement and acceleration toward a low F_0 both take some time, resulting in an F_0 peak in the initial portion of the second syllable.

2.1.3. Implementational effort

To implement pitch targets under various articulatory constraints, actual physical effort needs to be exerted. In our model, the amount of physical effort determines how effectively pitch targets are approximated during articulatory implementation. Other things being equal, a greater effort enables a pitch target to be approached sooner within its allocated time than a smaller effort. The amount of effort may be determined by various factors, including (but not limited to):

1. General muscle strength of the speaker as determined by individual physiological differences, the amount of speech training, such as years of speaking and whether there has been any professional training;
2. Overall effort as determined by the physical and emotional state of the speaker;
3. Sentence type such as simple statements versus exclamatory remarks;
4. Local effort as determined by stress, tone, focus, newness of information, etc.

Of these factors, not much is known yet about the first three. For the local effort, however, there has already been some preliminary data [5], as will be discussed later in this paper.

Although conceptually quite simple, the pitch target model is capable of handling a number of contextual F_0 variations reported in recent studies, as discussed in more detail in [37]. To stay close to the topic of alignment, I will discuss in more detail only how the model handles F_0 contour alignment.

2.2. Implications for F_0 contour alignment

Note first that the pitch target model assumes no micro-adjustment of F_0 contours on the part of the speaker. In terms of alignment, the model assumes that all the speaker is doing is implementing each pitch target (none of which is in the form of F_0 turning point, however) in synchrony with its host syllable. Nonetheless, the model can make certain predictions about the location of F_0 turning points, namely, peaks and valleys. In general, the location as well as the occurrence of F_0 turning points may be predicted by (a) the property of the pitch target, (b) the properties of the adjacent pitch targets, and (c) the duration of the host.

[3] The L tone in Mandarin is known to often have a falling-rising contour in pre-pausal and final positions. In those cases, this tone may conceivably have two consecutive pitch targets: e.g., [low]+[mid] or [low]+[high].

First, the model predicts that turning points often occur consistently near the syllable boundary. This is because a syllable boundary is where the implementation of one pitch target ends and that of the next begins. Whenever there is a difference between the ending F_0 of one target and the starting pitch of the next, the movement toward the second one would start at the boundary between the two host syllables. Therefore, a turning point can be often consistently observed at that location, as can be seen in Figures 3 and 4.

Second, the model predicts that a dynamic pitch target will often give rise to an F_0 turning point near the middle of the syllable, but the exact location of that turning point is dependent on the ending F_0 of the preceding syllable, as shown in Figures 3 and 4. For example, an F_0 peak will occur in a [fall]-carrying syllable if the preceding syllable has a low F_0 offset; and an F_0 valley will occur in a [rise]-carrying syllable if the preceding syllable has a high F_0 offset (or if the initial consonant of the syllable is voiceless). The exact location of the turning point will also depend on the offset F_0 of the preceding syllable: The more the ending pitch of the preceding target is different from the onset of the dynamic target, the later the F_0 turning point will occur, as was discussed earlier regarding the F and R tones shown in Figures 3 and 4.

Third, the model predicts that momentum plays an important role in determining the location of F_0 turning points. As explained earlier, the rising momentum at the syllable offset generated during the implementation of the R tone takes time to slow down and to reverse. This causes the F_0 turning point to occur consistently (about 5-50 ms) after the end of the first syllable in a RL or RR sequence, as has been found in Mandarin [32], [33], [34].

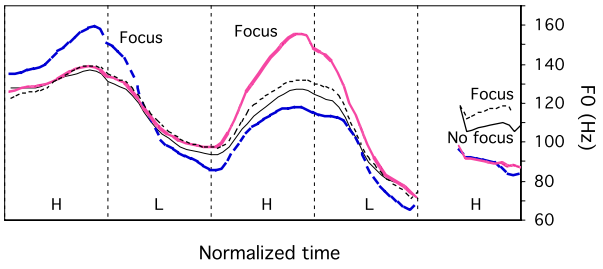

Figure 8. F_0 contour of the H tone in different focus conditions. Data from [33].

Fourth, the model predicts that the duration of a host syllable may further affect the micro alignment of F_0 turning points. For example, in Figure 8, the F_0 peak associated with the second H tone occurs before the end of its host syllable regardless of where the focus is. Because the average syllable duration without focus is about 180 ms [33], there is presumably enough time for the [high] to be reached by the end of the syllable [36]. If, however, the [high]-carrying syllable is much shortened, F_0 may be still rising quickly by the end of the syllable. As a consequence, the peak may occur in the early portion of the next syllable, resulting in the phenomenon known as "peak delay." This was confirmed in [34], in which peak delay was found to occur much more often when the H-carrying syllable was shortened at a fast speaking rate than at normal or slow rate. Further evidence can be found in Yoruba. As shown in data reported in [15], the highest F_0 in the LHL sequence usually occurs in the early portion of the second L-tone syllable. Assuming that it was due to short syllable duration that this peak delay occurs, then if the tone sequence becomes LHHL, there should be enough time for the F_0 movement to slow down before the end of the second H-carrying syllable. Indeed, both the F_0 tracings and the two-point-per-syllable measurements presented by [15] show that, in most of the L*n*HHL sequences (where *n*H indicates any number of H tones other than zero), the highest F_0 no longer occurs in the second L-carrying syllable, but inside the last H-carrying syllable.

3. The case of Mandarin neutral tone: evidence for implementational effort

As discussed earlier, implementational effort may determine how quickly an underlying pitch target is approached, other things being equal. A reduced effort may slow down the approximation of a target, for example. A recent study on Mandarin neutral tone has shed some light on the importance of understanding implementational effort. In Mandarin, beside the four distinctive lexical tones (H, R, L, and F), there is also a tone category known as the neutral tone [4]. Syllables carrying the neutral tone are traditionally considered to be toneless, because their F_0 contours appear to vary depending on the tone of the preceding syllable. Chen and Xu recently conducted a study of the Mandarin neutral tone to examine the exact nature of this tone category [5]. The study compared F_0 contours of sentences with different number of consecutive neutral-tone syllables which were preceded and followed by full lexical tones. Data from four native speakers of Mandarin show that although the F_0 contour of the neutral tone is quite dependent on the tone immediately preceding it, the influence of that tone decreases gradually as the number of neutral tone syllables increases, as shown in Figure 9.

In the top panel of Figure 9, the second syllable carries the L tone, and the tone of first syllable alternates among H, L, R and F. As can be seen, the initial portion of the F_0 curve in the second syllable varies extensively with the offset F_0 of the preceding syllable. By the end of the second syllable, however, the F_0 curves have virtually converged. In the lower panel, there are three neutral-tone syllables between the first and last syllables. Similar to the L tone, the difference due to the tone of the first syllable is substantial in the early portion of the first neutral tone syllable. Unlike the full tone, however, much of this difference still remains by the end of the first neutral tone syllable. Nevertheless, the difference is reduced over the course of the first syllable. Further more, the difference continues to reduce over the next two neutral tone syllables. By the end of the third syllable, much of the difference has disappeared. But some still remains visible. These F_0 patterns seem to reveal two important facts. First, the neutral tone is probably

associated with a pitch target of its own. Otherwise, there should not have been an apparent reduction of the influence of the preceding full tone even within the first neutral tone syllable. Second, the implementation of the neutral tone target is with a much reduced effort as compared to the full tones. Otherwise, the influence of the first tone should not have faded away so gradually, and remained visible even by the end of the third neutral-tone syllable.

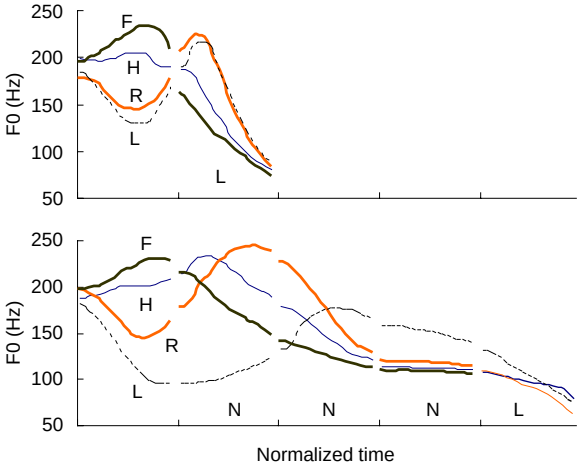

Figure 9.The top panel displays the F_0 curves of the L tone when preceded by four lexical tones in Mandarin. The lower panel displays F_0 contours of three consecutive neutral tone syllables when preceded by four lexical tones. The tone of the last syllable is again L. Each curve is an average of three repetitions. Data from [5].

Of particular relevance to the current topic, the alignment of F_0 turning points in Figure 9 can be better understood if we take articulatory constraints and implementational effort into consideration. Due to space limitation, I will discuss only two observations. First, in the case of the R tone, the F_0 peak (if it occurs at all) occurs in the following syllable that carries the L tone, as can be seen in the upper panel of Figure 9. As discussed earlier, this "peak delay" is presumably caused by the fact that it takes time to reverse a rising momentum at the syllable boundary. When the R tone is followed by the neutral tone, the F_0 after the R tone also continues to rise, as can be seen in the lower panel of Figure 9. However, like in the top panel, the rise does not continue into the next neutral tone syllable. Instead, F_0 starts to go down within the first neutral tone syllable after the R tone, resulting in a peak within that syllable. This pattern again suggests that the neutral tone probably has its own pitch target which, similar to the case of the full tones, is implemented in synchrony with the host syllable. If this interpretation is right, then the F_0 peak is not really intended as a peak, and the location of the peak is not really intended to be at a fixed location. Rather, based on the F_0 model discussed in 1.1., the location of the peak should be determined jointly by the rising momentum and the implementational effort exerted for the neutral tone. Indeed, the amount of the peak delay into the neutral tone is found to be larger than that into the L tone [5].

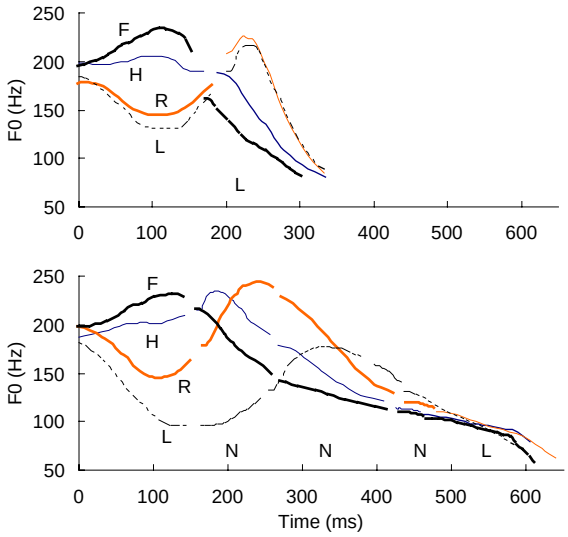

Figure 10. The same F_0 curves as those in Figure 9, but displayed here with actual time in millisecond. Data from [5].

The second alignment phenomenon is that peak delay can also be seen in Figure 9 with the H tone when it is followed by the neutral tone. That is, the highest F_0 occurs in the first neutral tone syllable after the H-tone syllable. Such peak delay occurred with two of the four subjects we examined. Similar peak delay rarely happens when the H tone is followed by the L or R tone unless at fast speaking rate [34]. The fact that it happens when the H tone is followed by the neutral tone probably also has to do with a reduced effort in implementing the neutral tone, which would have reduced the force to reverse the rising momentum at the end of the H-carrying syllable.

It might be argued that difference between the neutral tone and a full tone can be attributed to the short duration of the neutral tone syllable. It is true that the neutral tone typically has much shorter duration than a full tone [18]. However, the "recovery" from the influence of the preceding tone does seem slower in a neutral tone than in a full tone even when syllable duration is taken into consideration. In Figure 10, all F_0 contours in the upper panel converge to a low value by the end of the L tone at about 325 ms. The F_0 contours of the neutral tone in the lower panel, in contrast, do not even come close to converging by 325 ms. In other words, whatever the underlying pitch target associated with the neutral tone is, it does not seem to be implemented with the same amount of effort than in the case of a full tone. In fact, it is very likely that reduced implementational effort is a distinctive phonological property of the neutral tone. Naturally, this possibility needs to be further explored in future research.

4. Implications for tonal perception

Although articulatory constraints may play an important role in determining the shape and alignment of F_0 contours in speech, as I have been arguing so far, the importance of perception as a contributor to F_0 patterns should by no means be overlooked. In fact, the new understanding of articulatory constraints calls for new insight into the mechanisms of perceiving tone and intonation in speech. First, with the constraints of the maximum speed of pitch change and synchrony of laryngeal and supralaryngeal movements, speakers are probably under greater pressure to produce perceptually distinct F_0 contours; and listeners are faced with the task of separating the intended targets from the variations due to the speaker's articulatory limitations. One possibility is that listeners, who are also speakers themselves, are able to discover the underlying targets by factoring in the articulatory constraints. Some evidence of that already exists (e.g., [6] and [30]). More studies, however, are needed to investigate how exactly listeners can "unwind" the variations due to articulatory constraints and how successful they are in different situations. Second, the new findings also suggest that there is probably more to the perception task than to simply factor in coarticulation effects. For one thing, perception must be able to handle synchronized events. For example, how may perception expect synchronization, how can it detects synchronization, and how can it use synchronization as an aid for discovering segments and pitch targets? House's findings [9] as discussed earlier are probably an important step in that direction, because it demonstrates that listeners treat F_0 movements in different regions of a syllable differently. Of course, this raises the question as to which is recognized first: the syllable or the pitch targets associated with the syllable? Or may be the recognition of both is hand-in-hand. Third, from the manner at which a pitch target is implemented, including how much peak delay or valley delay occurs, listeners are probably able to hear how much implementational effort is exerted by the speaker, which may either encode conventionalized information, such as stress and focus, or simply reflect the physical or emotional state of the speaker. New studies are needed along all these lines.

5. Conclusions

Alignment is but one of the many aspects of tonal and intonational patterns in speech. Understanding it nevertheless requires the understanding of the basic mechanisms of speech production and perception in general. Conventionally, most of the tone and intonation literature has been tacitly assuming that articulatorially, speakers have much freedom in producing F_0 contours and align them with segmental units, which I referred to earlier as the *Free-Articulatory-Association assumption*. It is probably largely due to this assumption that whenever some F_0 shape or alignment patterns are observed, researchers tend to look into perception or the language itself for explanations. Recent findings about certain articulatory constraints in the human tonal production mechanism has convinced me that we have to give up the *Free-Articulatory-Association assumption* before we can better understand many tonal and intonational phenomena, including various alignment patterns. In this paper, I have presented what I believe to be evidence for the critical role played by articulatory constraints in producing many observed F_0 patterns. I have also presented a model in which articulatorially operable pitch targets are implemented under various articulatory constraints with different amount of implementational effort. At least conceptually, the model is able to generate a number of F_0 shape and alignment patterns reported in Mandarin and possibly also in English, including patterns of tonal alignment. Finally, our new understanding of articulatory constraints has implications for the understanding of tonal perception as well. Through the improved understanding of both the production and perception of tonal events in the speech signal, our understanding of speech communication in general may be enhanced.

References

[1] Arvaniti, A.; Ladd, D. R.; Mennen, I., 1998. Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics* 36, 3-25.

[2] Bruce, G., 1977. *Swedish word accents in sentence perspective*. Lund, Gleerup.

[3] Caspers, J.; 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.

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

[5] Chen, Y.; Xu, Y., forthcoming. Pitch target of Mandarin neutral tone.

[6] Fowler, C. A.; Smith, M., 1986. Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In *Invariance and variability of speech processes*, J. S. Perkell and D. H. Klatt (eds.). Hillsdale, NJ: LEA, 123-139.

[7] Gårding, E., 1977. The importance of turning points. In *Studies in Stress and Accent*, L. M. Hyman (ed.). Los Angeles, CA, Department of Linguistics, University of Southern California, 27-35.

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

[9] House, D., 1990. *Tonal Perception in Speech*. Lund: Lund University Press.

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

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

[12] Kim, S.-A., 1999. Positional effect on tonal alternation in Chichewa: Phonological rule vs. phonetic timing. In *Proceedings of Annual Meeting of Chicago Linguistic Society*, Chicago, 34, 245-257.

[13] Ladd, D. R.; Faulkner, D.; Faulkner, H.; Schepman, A., 1999. Constant "segmental anchoring" of F0

movements under changes in speech rate. *Journal of the Acoustical Society of America* 106, 1543-1554.

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

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

[16] Lehiste, I.; Peterson, G. E., 1961. Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America* 33, 419-425.

[17] Lin, M.; Yan, J., 1995. A perceptual study on the domain of tones in Standard Chinese. *Chinese Journal of Acoustics* 14, 350-357.

[18] Lin, T.; Wang, W., 1984. Shengdiao ganzhi wenti [Perception of tones]. *Zhongguo Yuyan Xuebao [Bulletin of Chinese Linguistics]* 2, 59-69.

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

[20] Prieto, P.; van Santen, J. P. H.; Hirschberg, J., 1995. Tonal alignment patterns in Spanish. *Journal of Phonetics* 23, 429-451.

[21] 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.). Canberra: Pacific Linguistics, 55-82.

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

[23] Shih, C.-L., 1988. Tone and intonation in Mandarin, *Working Papers, Cornell Phonetics Laboratory* No. 3, 83-109.

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

[25] 't Hart, J.; Collier, R.; Cohen, A., 1990. *A perceptual Study of Intonation — An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

[26] Tiffany, W. R., 1980. The effects of syllable structure on diagochokinetic and reading rates. *Journal of Speech and Hearing Research* 23, 894-908.

[27] van Santen, J. P. H.; Hirschberg, J., 1994. Segmental effects on timing and height of pitch contours. *Proceedings of The International Conference on Spoken Language Processing*, 94, 719-722.

[28] Whalen, D. H.; Levitt, A. G., 1995. The universality of intrinsic F0 of vowels. *Journal of Phonetics* 23, 349-366.

[29] Xu, C. X.; Xu, Y.; 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, 2359-2362.

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

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

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

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

[34] Xu, Y., 2001. Fundamental frequency peak delay in Mandarin. *Phonetica* 58, 26-52.

[35] Xu, Y., 2001. Sources of tonal variations in connected speech. *Journal of Chinese Linguistic*, monograph series #17, 1-31.

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

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