Italian Journal of Linguistics (2006) 18: 125-159.

Tonal alignment, syllable structure and coarticulation: Toward an integrated model

Yi Xu

University College London, London, UK & Haskins Laboratories, New Haven, CT, USA,

yi@phon.ucl.ac.uk

Fang Liu

University of Chicago, Chicago, USA, liufang@uchicago.edu

Abstract. The finding of consistent tone-segment alignment in many languages in recent research raises questions about the temporal organization of speech sounds in general. In this paper we explore the possibility that tonal alignment patterns can lead to the discovery of basic principles of temporal organization in speech. Based on a recent finding about the segmentability of approximants in English and Mandarin, we propose a general model of temporal organization, in which the syllable is the basic time structure that specifies the alignment of consonants, vowels, tones and phonation registers. All these sounds are unified under the term *phone* defined as a collection of *unidirectional* articulatory movements toward a simple or composite target. The phones are temporally organized by the syllable under three principles: co-onset of initial C and V, sequential offset of coda C, and full synchronization of tone and phonation register with the syllable. Under the time structure model, true coarticulation, in the strict sense of co-occurrence of separate phones, occurs only between initial C and V; and there is no anticipatory C to V coarticulation, no cross-consonantal V-to-V coarticulation, and no carryover coarticulation of any kind.

1. Introduction

In the "good old days" of linguistics, lexical tones were thought to be directly associated with words or syllables (e.g., Chao, 1968; Pike, 1948). Along then came auto-segmental phonology which changed this view for good, or at least as it has seemed so for a long time. Auto-segmental phonology, as the name implies, is based on the idea that the tonal components of a language form a tier that is independent of the segmental tier, hence "auto-segmental" (Goldsmith, 1976, 1990). A consequence of such independence is that many tonal phenomena are said to be explainable in terms of changed association of the underlying tones with the segmental material, usually the syllable or the vowel. Thus tones that are lexically associated with a particular syllable often break free from such association and become re-associated with a different syllable or vowel, following the operation of various proposed association rules. Such a framework has since been applied not only to many tone languages, but also to non-tone languages in a revised form known as Autosegmental-Metrical theory of intonation (henceforth the AM theory) (Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986; Ladd, 1996). In the AM theory, although the intonational tones are not moved about by complex association rules, tonal events such as pitch accent, phrase tone and boundary tone are only loosely associated with specific syllables with no strictly specified alignment rules. This situation, however, has changed with the increasing number of findings of a much stricter alignment not predicted by the classical form of the AM theory (Arvaniti, Ladd &

Mennen, 1998; Atterer & Ladd, 2004; D'Imperio, 2001, 2002; Ladd et al. 1999; Ladd, Mennen & Schepman, 2000; Xu, 1998, 1999, 2001)[1]. While its nature is still under intense debate, it is our belief that a better understanding of such alignment may be achieved only through a better understanding of temporal organization of the sounds of speech in general. This paper is thus an attempt to find the general principles of such temporal organization. We will start our discussion from one of the central issues in experimental speech research.

2. A matter of segmentation

Although, for a long time, linguists, especially those involved in experimental research, have doubted the possibility of clear segmentation of speech sounds (Joos, 1948), few have pointed out that the now widely accepted notion of coarticulation is in fact heavily rooted in implicit assumptions about segmental boundaries. The main evidence for coarticulation is the observation that the acoustic manifestation of one speech sound is affected by the properties of adjacent sounds. Note that such an observation has to be based on the assumption that we "know" where each segment is in the acoustic signal. For example, in the spectrogram of "my meal" shown in Figure 1a, the locations of the segments are generally understood as labeled in the figure. Hence, when nasality is found during an interval where the spectral pattern is clearly indicative of a vowel, the nasal consonant is said to be coarticulated with the vowel (Huffman & Krakow, 1993). Likewise, when it is shown that some characteristics of a vowel, e.g., its F2, affects the acoustic manifestation of a non-adjacent vowel, the vowels are said to be coarticulated across the intervening consonant (Öhman, 1966). In both cases, therefore, it is assumed that the proper location of a particular segment is where its most characteristic acoustic properties are found. Thus a vowel resides in an interval where clear formant patterns can be seen, a nasal in an interval where there are clear nasal resonances, a fricative in an interval with clear frication noise pattern, and a stop in an interval that starts with a sudden drop in acoustic energy and ends with a release burst. But like any other in scientific research, this assumption can be questioned. Before doing that, however, we will first take this assumption as given and try to answer a specific question: how can we segment sounds, such as [j], [w] and [ɹ], that do not display sharp acoustic landmarks?

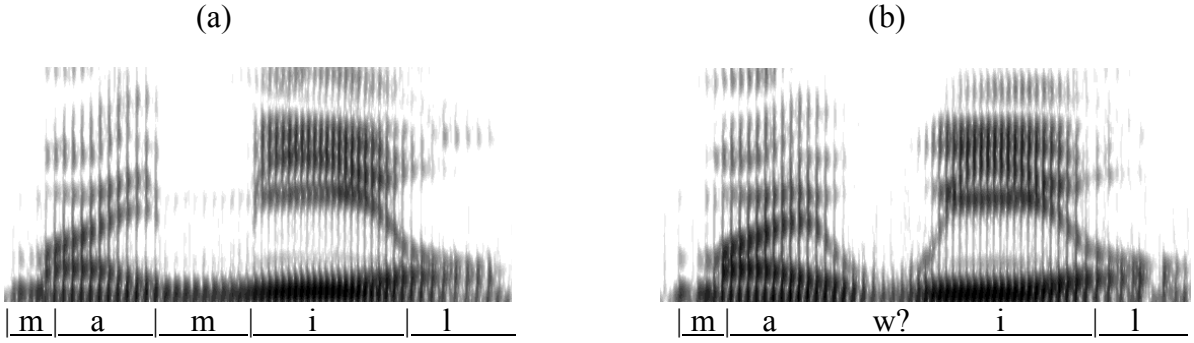

Figure 1. Spectrogram of "my meal" (a) and “my wheel” (b). The labels are based on conventional phonetic segmentation. (General American English).

Figure 1b displays the spectrogram of "my wheel." Unlike in Figure 1a, where around the location of [m] a gap, known as nasal murmur, can be clearly seen, in Figure 1b the

[1] Also see Ginésy & Hirst (1975) for an early anticipation of recent alignment findings.

approximant [w] does not show any sharp landmarks. If having to visually locate an acoustic landmark, one would be driven to point to the formant peaks and valleys, i.e., where the formants become the most extreme. Thus for [w] in "wheel," this would be where F1 and F2 are the lowest. Indeed, the landmarks proposed by Stevens (2002) for the perception of glide sounds like [j] and [w] are the peaks and valleys of the formants. Therefore, when segments are considered alone, it is reasonable to view the formant peaks and valleys in approximants as equivalent to the nasal murmur onset. Probably because there is no clear evidence to either support or oppose such understanding, no one to our knowledge has taken a firm stand on this issue one way or the other. Nevertheless, any comprehensive theory about coarticulation should not avoid the issue of how to segment approximants, or at least how their segmentation is comparable to that of other, "easier" sounds. The question is, of course, how can we do it?

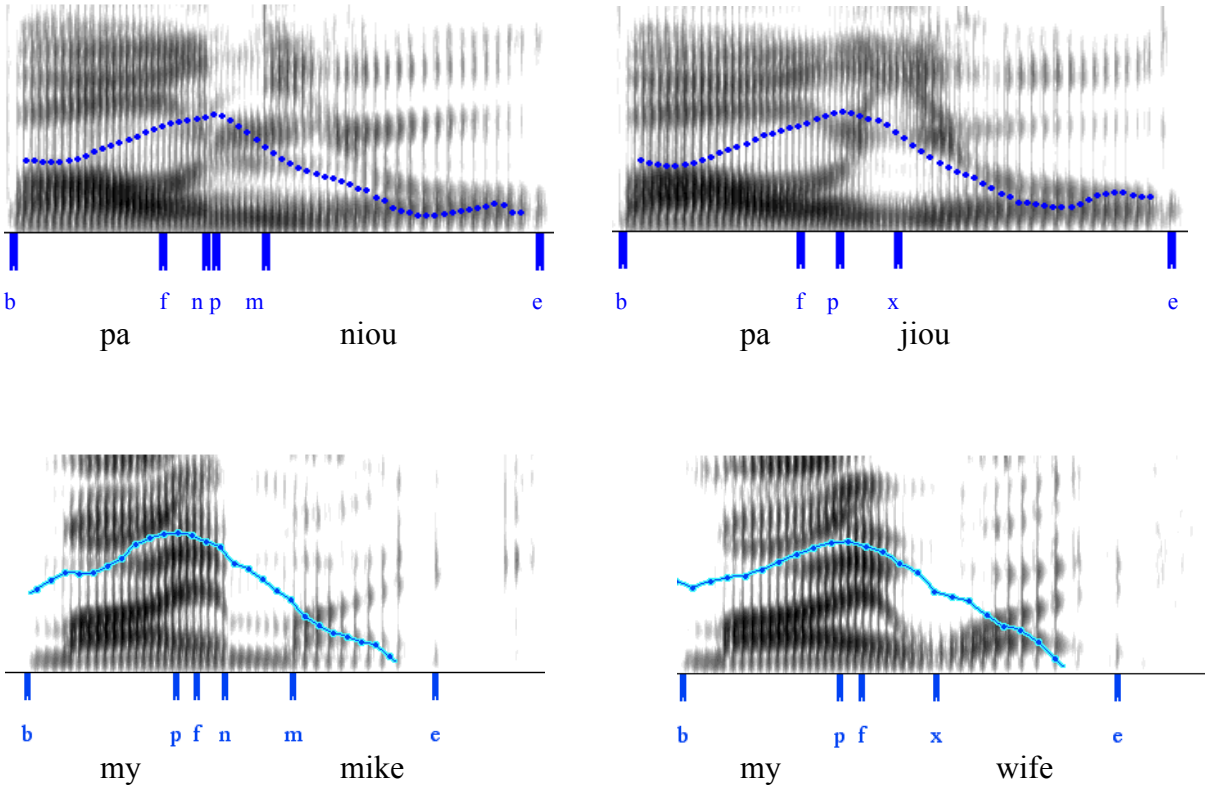

Figure 2. Illustration of markers placed in each word. Top: Mandarin nonsense words [paʊR niouʊL] (left) and [paʊR jiouʊL] (R and L stand for Rising and Low tone). The dotted lines are F_0 tracings generated by Praat. Bottom: English (General American): "MY mike" and "MY wife." (Capitalization indicates emphasis)

2.1. F_0 alignment to the rescue

As mentioned earlier, in recent years there have been a flurry of studies on various languages all showing patterns of consistent alignment of F_0 peaks and valleys with the onset and/or offset of the syllable. While those findings may be interpreted as telling us how tonal events are aligned to segmental events (Atterer & Ladd, 2004), the reverse could be true as well. That is, F_0 alignment may also tell us how segmental events are aligned to tonal events. In other words, not only are F_0 peaks and valleys consistently aligned to the onset or offset of certain segments, but also segments are consistently aligned to certain F_0 peaks and valleys. If so, F_0 alignment can be used as a heuristic for determining segmental alignment in cases where ambiguity is severe. This is the reasoning behind the two experiments reported in Xu

and Liu (2002) and Liu and Xu (2003).[2]

In those experiments we explored the segmentability of approximants [j], [w] and [ɹ] in English and [j] and [w] in Mandarin by using F_0 turning points as the temporal indicators and F_0-nasal alignment pattern as reference. The strategy was to find word or phrases in which (a) F1, F2, and/or F3 make two sharp turns in the vicinity of the target consonant, so that the formant movements associated with it are clearly visible, (b) F_0 makes a sharp turn near that consonant due to lexical tone (Mandarin) or focus (English), and (c) pairs of initial approximant and initial nasal that share similar tonal and segmental contexts, so that their comparisons could be as direct as possible.[3]

Figure 2 shows spectrograms and F_0 tracks of two pairs of disyllabic words/phrases examined in the experiments, (a) and (b) for Chinese and (c) and (d) for English. In each graph, *f* marks the point at which F2 starts to move toward the locus of the initial consonant of the second syllable; *p* marks the F_0 turning point in the vicinity of the syllable boundary; *n* and *m* mark the onset and offset of the nasal murmur in the nasal consonant; and *x* marks the point at which F2 starts to move toward the value of the following vowel in the approximant. Eight Mandarin pairs and six English pair/triplets were examined in the two experiments. They were recorded by four Mandarin speakers (2 females and 2 males) and five American English speakers (3 females and 2 males). The F_0 contours were controlled by lexical tones in the Mandarin experiment but by focus in the English experiment. The consistency of focus in the English experiment was guaranteed by leading questions that induced emphasis on either the first or the second word.

---

[2] As pointed out by a reviewer, the timing of F_0 turning points is not necessarily essential to the modeling of tone and intonation, because, in particular, F_0 peak alignment depends in all sorts of complicated ways on the temporal structure of the segments associated with the pitch movement. We fully agree with this view, and in fact in our Target Approximation model illustrated in Figure 5, there are no direct specifications for the alignment of turning points. On the other hand, what is important for the present purpose is that once the known factors affecting tonal alignment are effectively controlled, the location of F_0 turning points relative to segmental events are highly consistent. Such consistency is what we are tapping into for resolving the existing ambiguities in segmental alignment.

[3] The four lexical tones of Mandarin has characteristic pitch patterns of high level, rising, low(-rising) and falling. For a more detailed description of the Mandarin tonal system, please see Chao (1968). For the detailed F_0 pattern of Mandarin tones produced in isolation and context and their alignment with the syllable, please see Xu (1997, 1999, 1998, 2001). For F_0 alignment as related to focus in English please see Xu & Xu (2005).

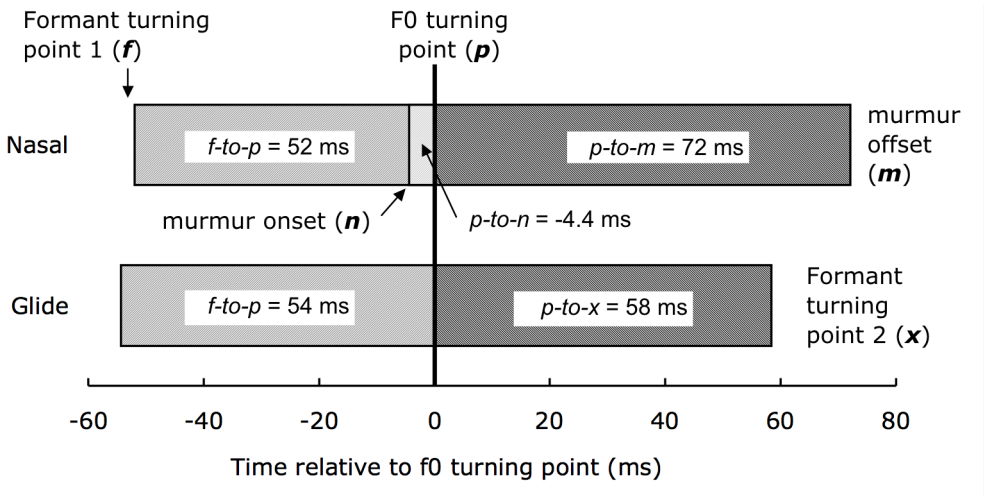

Figure 3. Mean values of *f-to-p, p-to-n, p-to-m* and *p-to-x*, averaged across four Mandarin subjects. The F_0 turning point (*p*) is plotted at time 0, which serves as the reference point for all other values.

Figure 3 is a summary plot of the mean values of the measurements, including *f-to-p*, *p-to-n*, *p-to-m* and *p-to-x* in the Mandarin experiment. The English experiment yielded comparable data. In Figure 3, the F_0 turning point (*p*) is plotted at time 0 and other measurements are plotted relative to it. Displayed this way, the time relation among the measurements provides information for determining the points in an approximant that are analogous to the onset and offset of a nasal murmur. Instead of giving us straightforward answers, however, the results surprisingly left us with two puzzles. First, as can be seen in Figure 3, when the initial consonant of the second syllable is nasal, the onset of the nasal murmur occurs right before the F_0 turning point. Using this alignment as reference, the onset of an approximant should also occur right before the F_0 turning point. As we can see in Figure 2, this inferred syllable boundary would not correspond to any clear acoustic landmark. Hence, the first puzzle: what is the nature of the inferred approximant onset?

Secondly, *x* — the point at which the formants of an approximant reach the extreme values is well after the inferred approximant onset, about 62 ms (= 58 + 4 ms; 50-58 ms in English, cf. Liu & Xu 2003). This means that the moment when formants reach the most extreme values for the approximant is unlikely to be temporally analogous to the nasal murmur onset. Furthermore, at least in terms of the order of the events, *x* appears to be sequentially close to *m*, i.e., the offset of the nasal murmur in the second syllable. But the former is somewhat earlier than the latter. The average difference is 14 ms in Mandarin and 17-32 ms in English. Hence, the second puzzle: what is the nature of the formant turning point in an approximant?

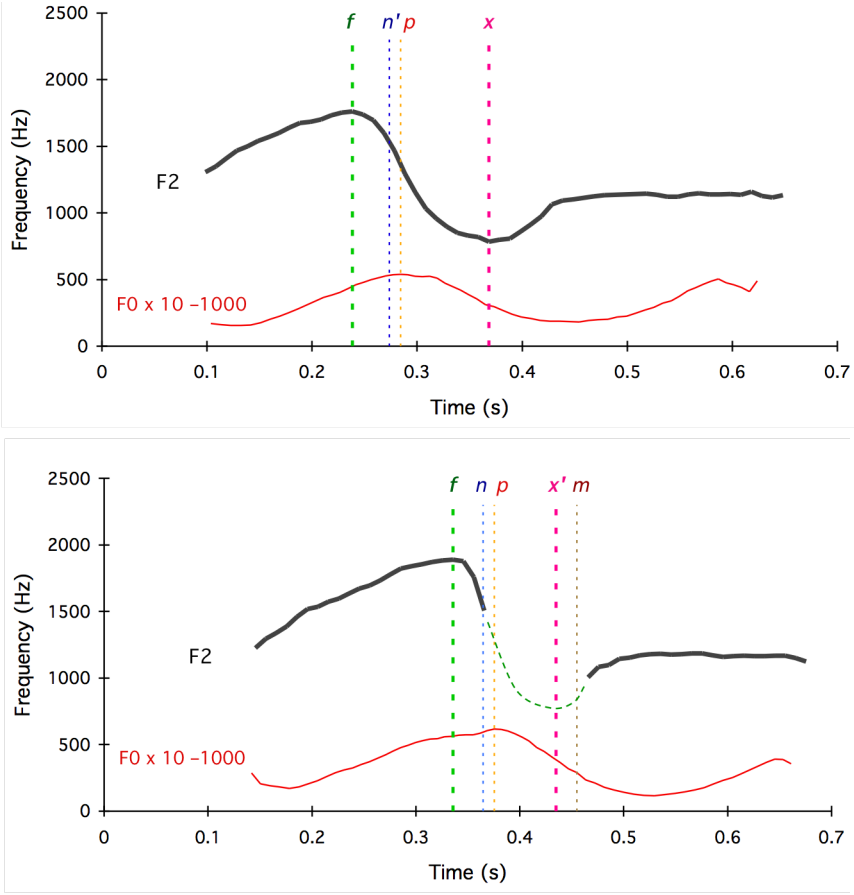

Figure 4. Schematic plots showing F2 and F_0 tracks of two Mandarin words uttered by a Mandarin speaker together with various alignment points. (a) [paɪR waR] (R = Rising tone). (b) [paɪR maR/].

To solve the two puzzles, we first drew in Figure 4 two schematic graphs showing F2 and F_0 tracks together with various alignment points. Figure 4a shows [paɪR waR] uttered by a Mandarin speaker (where R represents Rising tone). In addition to the alignment points previously seen in Figure 2, *n′* marks the point in [paɪR waR] that is temporally equivalent to *n* in [paɪR maR], i.e., nasal murmur onset, as is shown in Figure 4b. Again, this inferred point does not correspond to any apparent acoustic landmark. Nevertheless, what can be seen is that *n'* is situated in an interval during which F2 continues to move towards the lowest value, which starts at *f* and ends at *x*. Since [w] is a back rounded vowel, its canonical F2 is very low (Fant, 1960). The [ɪ] that precedes [w], being high-front and unrounded, has a fairly high canonical F2. Thus the entire *f*-*x* interval is one in which F_0 continually moves from [ɪ] to [w]. This kind of movement is very similar to the F_0 movement in a tone as simulated by the Target Approximation (TA) model (Xu & Wang, 2001), as shown in Figure 5. In the TA model, surface F_0 contours (e.g., the solid curve in Figure 5) result from asymptotic approximations of underlying pitch targets defined as simple linear functions (e.g., the dashed lines in Figure 5). Following the spirit of the TA model, the F2 movements in Figure 4a can be divided into three intervals separated by the dashed vertical lines. During the first interval, F2 moves toward a value appropriate for [ɪ], i.e., the ending element of the diphthong

[a][4]. During the second interval, F2 moves toward a value appropriate for [w]. During the third interval, F2 moves toward a value appropriate for [a], although this approximation seems to reach an asymptote and stay there till the end of the syllable. This observation thus suggests that *x* is the time when the approximation of [w] has just ended, whether or not the targeted value is actually reached, because the subsequent movement is apparently toward [a].

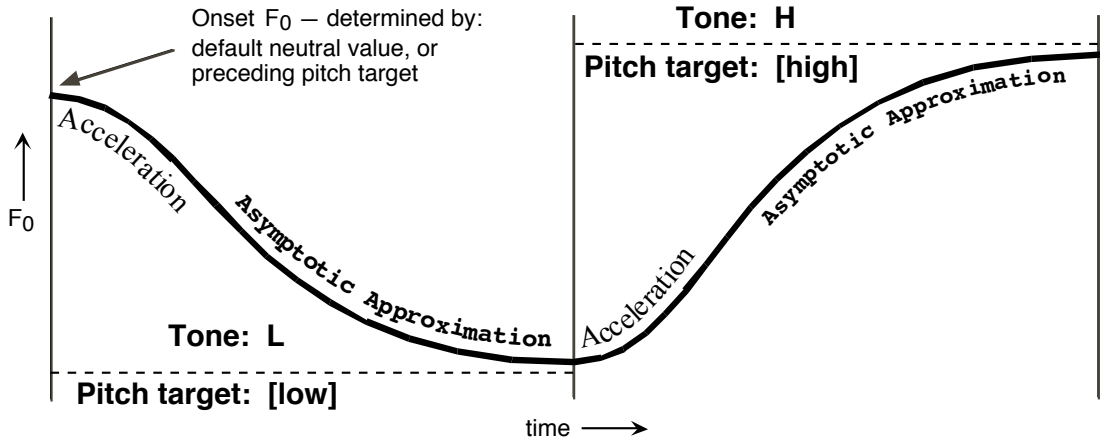

Figure 5. A schematic sketch of the Target Approximation model (Xu & Wang, 2001). The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from articulatory implementation of the pitch targets.

More interestingly, the above observation can be further applied to the case of [paɪR maR], as shown in Figure 4b. That is, we could also take *f*, rather than *n* — the nasal murmur onset, as the onset of the articulatory movement toward the nasal consonant. In other words, the point *f* is the consequence of the oral cavity starting to change toward a shape that is appropriate for the bilabial nasal [m]. But the movement toward the air-tight labial closure is not completed at the moment of the acoustic landmark: *n* — the nasal murmur onset. Rather, the movement is still at a high velocity at that moment, as found by Löfqvist and Gracco (1999) for [b] and [p]. Furthermore, *m* — the offset of the nasal murmur is probably not the end of the movement toward the articulatory goal of [m]. Rather, that movement is likely to have terminated sometime *before* the offset of the nasal murmur, as depicted by the hypothetic dotted curve adjoining the two continuous F2 curves interrupted by the nasal murmur. In other words, the end of the nasal murmur is the moment when the articulatory movement toward the following vowel, after going on for a short while, has just resulted in the parting of the lips in [m]. The point in a nasal analogous to *x* in an approximant is therefore not *m*, but rather some location inside the nasal murmur, which is illustrated in Figure 4b as *x′*. So, again, it is the *f*-*x′* interval that is likely the temporal domain of [m] implementation.

2.2. What does it mean?

[4] In fact, the underlying target being approached here is likely to be intrinsically dynamic, i.e., one that is similar in nature to dynamic tonal target such as [rise] and [fall] proposed in Xu & Wang (2001). A dynamic target is one which can be represented as a simple linear function in the form of $y = ax + b$, where *a* is the slope of the straight line and *b* its y-intercept (i.e., its vertical height).

The alignment scheme just derived, of course, is radically different from the conventional understanding of segmental alignment. As discussed in the Introduction, the widely accepted notion of coarticulation is based on the assumption that *the time interval where the acoustic pattern is most typical of a segment is the temporal domain of that segment*. Thus the nasal murmur *is* the interval of the nasal consonant, and the interval where vowel formants can be clearly seen *is* where the vowel is. The new understanding emerged from the two recent experiments as just described, however, has moved the onset of virtually all segments leftward by about 26-48 ms (based on calculation in Xu & Liu, in press) from the conventionally assumed onset. Thus a nasal consonant no longer begins with the nasal murmur onset. It begins, rather, at the point when the formants start to move toward the values appropriate for the nasal's place of articulation. A vowel no longer begins at the point when the vowel formants first appear. It rather begins at the point where the formants first start to move toward the ideal values. As we will explain later, that point is well before even the onset (by the conventional definition) of the initial consonant of the syllable in which the vowel occurs.

Such a radical change in the understanding of segmental alignment therefore calls for a new model of temporal organization of speech sounds. In the following section we will propose such a model, and we will show that in addition to the findings of the two experiments just discussed, many other lines of evidence for the model have been steadily accumulating over the years in research on the segmental aspects of speech.

3. The time structure model of the syllable

Before presenting the model, we first state a number of definitions that are essential to the model.

[1] Target — A target is an underlying goal specified in terms of ideal articulatory/acoustic patterns. A target can be either static or dynamic, and either simple or composite. A target has both positional and velocity specifications. A composite target consists of multiple positional and velocity specifications.

[2] Phone — A phone is a collection of *unidirectional* articulatory movements toward an integrated target. *Any movement away from the target is not part of the phone.* A movement does not always reach its target.

[3] Segment — A segment is a segmental phone with a target specified in terms of both vocal tract shape and spectral pattern. A segment is either a consonant C or a vowel V. A C typically has a narrower vocal tract constriction than a V.

[4] Tone — A tone is a laryngeal phone with a target specified in term of pitch (i.e., abstract fundamental frequency), abbreviated as T.

[5] Phonation register — A phonation register is a laryngeal phone with a target specified in terms of phonation type (voice quality such as breathy or pressed. cf. Ladefoged, 1983), abbreviated as P.

As will be seen, [1], which defines the target as an underlying goal and [2], which stipulates the unidirectionality of phones, are particularly critical to the time structure model. The importance of the target notion is that it represents the intended goal as opposed to observed events. This differs from models like the AM theory of intonation, in which targets are defined as surface F_0 turning points (Pierrehumbert, 1980). The target notion is not new, as it

is assumed in a number of existing frameworks, including the undershoot model (Lindblom, 1963a; Moon & Lindblom, 1994), the Equilibrium Point Hypothesis (Feldman, 1966, 1986; Perrier, Ostry & Laboissière, 1996), Articulatory Phonology (Browman & Goldstein, 1986, 1989) and the task dynamic model (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989).[5] Unidirectionality is implicitly assumed in some models, such as the Equilibrium Point Hypothesis. But in many other models, linguistically meaningful motor events are often assumed to be bidirectional, i.e., consisting of both onset and release, or movements both to and from the target (Browman & Goldstein, 1986, 1989; Fujisaki, 1983; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989; van Santen & Möbius, 2000). As we will demonstrate, these two notions, together with the explicit alignment specifications, allow the time structure model to drastically reduce the amount of coarticulation, a problem that has been troubling speech researchers almost ever since accurate instrumental observation of the speech signal was first made possible (e.g., Joos, 1948).

3.1. The Model

A simplified schematic of the time structure model is shown in Figure 6. The graph shows a two-syllable sequence as well as the segmental and laryngeal phones aligned relative to them. The vertical lines represent syllable boundaries. The horizontal axis is time. The vertical axis is the activation state of the phones. The model assumes the following principles:

[6] The syllable is a time structure that specifies the temporal alignment of all the phones, including C, V, T and P.

[7] Co-onset — The initial C, the first V, the T and P all start at the syllable onset.

[8] Sequential offset — Non-initial segments, whether V or C, are sequentially aligned after the first V of the syllable.

[9] Synchrony of laryngeal phones — Both T and P are synchronized with the entire syllable to which they are associated.

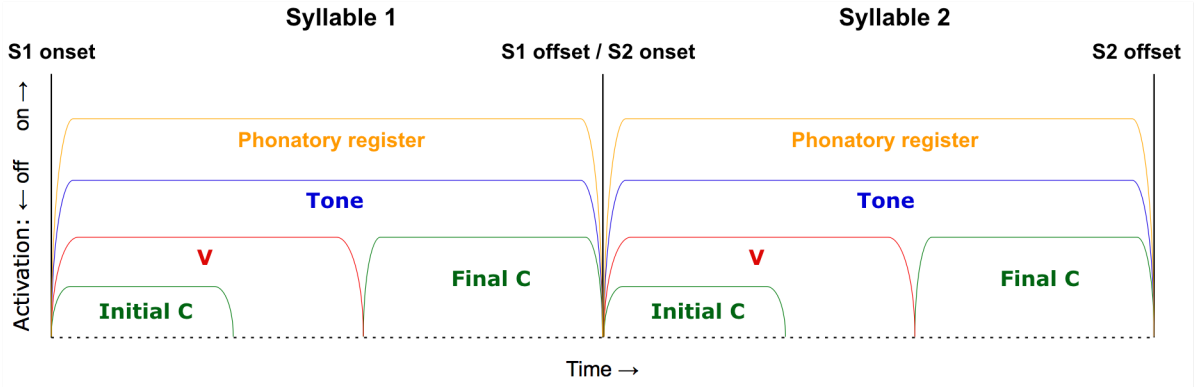

Figure 6. A simplified schematic of the time structure model.

[5] None of the earlier models, however, have allowed for dynamic targets as proposed in Xu & Wang (2001). Not being essential to the proposed model, the notion of dynamic targets will not be elaborated in this paper.

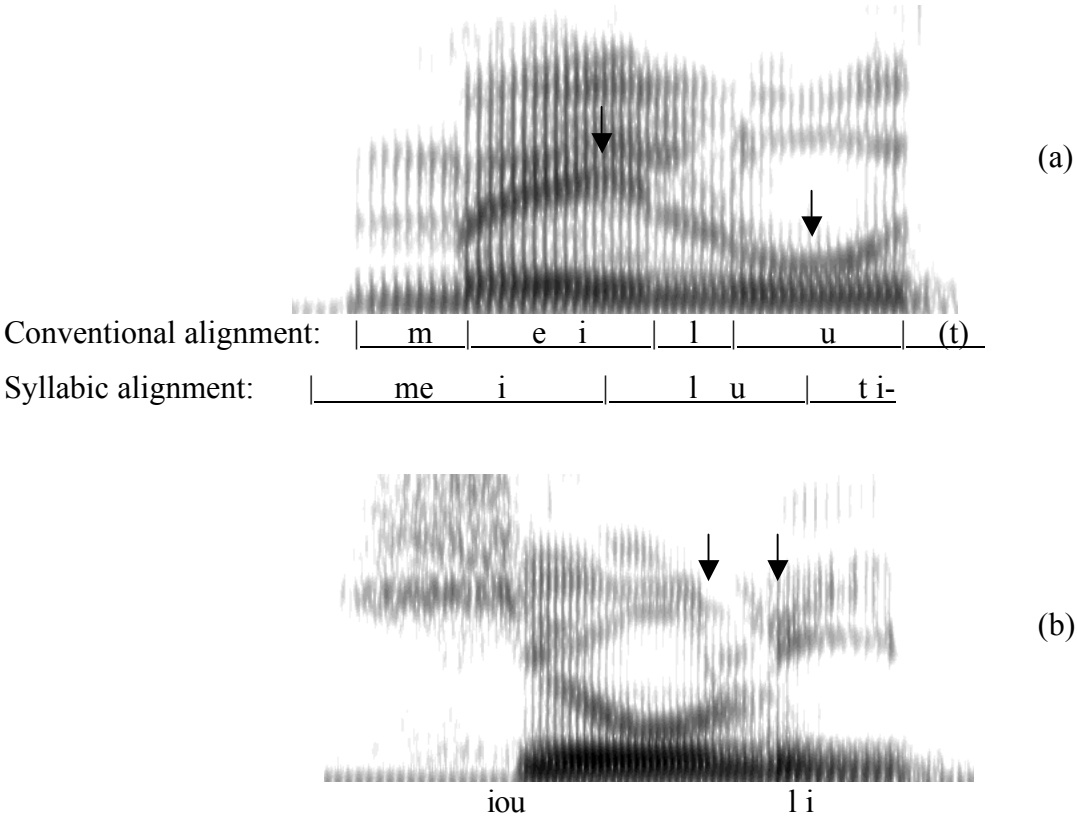

Figure 7. (a) Illustrative comparison of conventional alignment and proposed syllabic alignment. The entire utterance is [meɪR luR (tiɛnL χuoL)] (Light coal stove) in Mandarin (spoken by the first author), but only the spectrogram of meiR luR is displayed. The two arrows mark the onset and offset of the coproduced /lu/. (b) Spectrogram of [χioʊH liL (buF ʐoʊF)] (repair procedure) spoken by the first author. The two arrows mark the onset and offset of the /l/ murmur.

Figure 7a illustrates how C and V in a two-syllable sequence align differently according to the conventional view as opposed to the time structure model. The conventional view aligns the segments according to the acoustic landmarks as indicated by the upper-row labels. As a consequence, whenever a segment exhibits characteristics of a neighboring segment, the two segments are said to be coarticulated. The segmental alignment is rather different under the time structure model, as indicated by the lower-row labels. First, because the vowel of the first syllable is [eɪ], by the time F2 reaches the peak (first arrow), the underlying articulatory movements toward the [eɪ] target should have been terminated, and the subsequent formant movements should be due to articulatory movements that approach the targets associated with the second syllable. Hence, the first F2 turning point is where the first syllable ends and the second syllable starts. Secondly, since [l] and [u] start at the same time, the downward movement of F2 is not only toward [l], but also toward [u]. This explains why F2 drops so quickly during [l], which is light in Mandarin and should have a relatively high F2. As manifested by the continuous F2 movements around [l], the approximation of [u] in the second syllable actually starts from the F2 peak (first arrow) and ends at the F2 valley (second arrow). As a comparison, we can see in Figure 7b the upward F2 movement around the interval marked by the two arrows when the vowel of the second syllable is [i]. Note that, while the phenomena are the same as those observed in many coarticulation studies (see extensive reviews by Hardcastle & Hewlett, 1999), the F2 movement before the [l] murmur is no longer considered as due to long-distance cross-consonant V-to-V coarticulation under the

time structure model. Instead, it is viewed as part of [u] itself. Third, similarly, by the middle of the second vowel as marked by the second arrow in Figure 7a, the approximation of the low F2 for [u] is over, and the subsequent upward movements are toward both [t] and [i] in the following syllable. Again, there is no coarticulation between [u] and the following [t] or [i]. So, thanks to the strict definition of phone as stated in [2] as well as the principle of CV co-onset as stated in [7], the amount of coarticulation is significantly reduced from what is believed conventionally.

3.2. Evidence

3.2.1. Syllable – the basic time structure

Despite our strong intuition for the existence of syllables, efforts to find acoustic and physiological correlates of syllables had largely failed by the mid 80s, according to the extensive literature review by Krakow (1999). The quest has since continued, nonetheless, and a number of more structural accounts have been proposed. In the C/D model of phonetic implementation, sequentially ordered syllables function as the basic segmental units (Fujimura, 1994, 2000). According to the frame/content theory of speech production, all spoken utterances are superimposed on successive syllables which constitute a "continual rhythmic alternation between an open and a closed mouth (a frame) on the sound production process" (MacNeilage, 1998: 499). Some of the strongest evidence for the existence of the syllable is described by Krakow (1999), who characterizes the syllable as a physiological unit. She argues that the CV unit "provides an alternating articulatory pattern beginning with a tight constriction and ending with an open vocal tract and, as such, results in a kind of rhythm that is especially suited both to the production and [to the] perception mechanisms" (p. 50). She further shows that initial consonants have tighter constrictions and greater stability than final consonants, and that final consonants are subject to loss, either by disappearing or by being transformed into the initial consonant of the following syllable, depending on the specific language.

The time structure model proposed in 3.1 is consistent with all these developments. But it is more explicit in terms of the detailed temporal organization of the syllable and in terms of the essence of such organization. The model assumes that the syllable is the basic time structure that specifies the temporal alignment of all the phones: consonants, vowels, tones and phonation registers. Consistent with the general hypotheses of Fujimura (1994, 2000), MacNeilage (1998) and Krakow (1999), we consider CV as the most basic structure of the syllable. But more specifically, we propose that the initial consonant and the first vocalic element in a syllable share the same onset time. Or, in other words, the initial C is completely rather than partially overlapped with the first vocalic element of the syllable, although the former necessarily ends earlier than the latter. Furthermore, unlike in the other theories, the time structure model assumes no underlying strength difference between initial and final consonants (e.g., Krakow, 1999). Instead, it only recognizes that coda consonants have to be sequentially aligned *after* the vocalic phone(s) of the syllable. Their weaker status as universally observed is understood as due to the joint effects of such sequential alignment and the constraint of the maximum speed of articulatory movement. Sequential alignment means that the implementation of an element cannot start until that of the preceding element has terminated, which makes the implementation of the coda consonants much more time-consuming than that of the initial consonants. When there is not sufficient time to execute the articulatory movement toward the target because the maximum speed of articulatory movement has been approached (which is likely to happen frequently according to the findings by Janse, 2003, Xu and Sun, 2002 and Xu, in press), the final consonant simply

cannot be implemented, resulting in either their loss (via deletion or merger into the preceding vowel in the form of, e.g., nasalization) or shift in affiliation to the following syllable (Sproat & Fujimura, 1993; Gick, 2003), as we will discuss later in greater detail.

3.2.2. Co-onset of C, V, T and P

The idea of co-onset of CV was suggested as early as in the 1930s by Menzerath and de Lacerda (1933), who was also the first to use the term coarticulation (koartikulation), according to Kühnert and Nolan (1999). Thus the earliest proposal of coarticulation was based on the observation that "the articulatory movements for the vowel in tokens such as /ma/ or /pu/ began at the same time as the movements for the initial consonant" (Kühnert & Nolan, 1999:14). Similar observation was made by Öhman (1966). Based on formant transition patterns in VCV sequences in Swedish and American English, Öhman postulated that in these languages, "a motion toward the final vowel starts not much later than, or perhaps even simultaneously with, the onset of the stop-consonant gesture" (p. 165). The idea of co-onset, however, was soon overshadowed by reports of long-distance anticipatory coarticulation extending as far as six intervening consonants (Benguerel & Cowan, 1974; Sussman & Westbury, 1981. See Kühnert & Nolan, 1999 and Farnetani & Recasens, 1999 for thorough reviews). It was therefore hypothesized that the distinctive feature of a vowel is spread leftward as far as possible, so long as it is not blocked by a contrasting feature in a preceding segment (Daniloff & Hammarberg, 1973). However, a series of studies conducted at Haskins Laboratories across nearly two decades have provided strong evidence that the articulatory movement related to a segment actually starts at a rather constant time close to the acoustic onset of the segment (Bell-Berti & Harris, 1979, 1981; Bell-Berti & Krakow, 1991; Bell-Berti et al. 1995; Boyce, Krakow & Bell-Berti, 1992; Krakow, 1999). Furthermore, evidence is found that “the timing of movement onset for gestures appropriate to consonants was tightly linked to the timing of movement onsets for vowel-related gestures.” (Tuller & Kelso, 1984:1034). All this points to a pattern of tight timing relation between initial consonant and the following vowel, thus providing support for Öhman’s (1966) early observation.

It is much less clear, due to lack of direct evidence, how co-onset is realized when the initial C is a consonant cluster. There has been evidence that individual consonants in an onset cluster are more consistently aligned relative to each other than in a coda cluster (Byrd, 1996). The c-center effect reported by Browman and Goldstein (2000) is another indication of the more consistent alignment of initial consonants with the nuclear vowel. The effect refers to the phenomenon that the time lag from the temporal center of the initial C to the end of the nuclear vowel remains relatively constant as the number of consonants in the initial position varies. The same consistency was not found among the final consonants. It is not clear, however, how exactly the first V is aligned relative to the individual consonants in an onset C cluster, i.e., whether it starts with the first consonant or from the center of the entire cluster. New research is needed to establish such alignment relation.

As for consistent alignment of tonal events with the syllable onset, much evidence has been produced by recent research. For example, the findings of Arvaniti, Ladd and Mennen (1998), Atterer and Ladd (2004), D'Imperio (2001, 2002), Ladd et al. (1999), Ladd, Mennen and Schepman (2000) and Xu (1998, 1999, 2001) all suggest that an F_0 movement toward a tonal target starts at the onset of the syllable. Syllable onset is hence the point at which the greatest synchrony is achieved across the initial consonant, the first vowel and the tone of the syllable. While we are not aware of direct evidence for the co-onset of phonation register with other phones, there is no reason for us to believe that it should be exceptional. At the same time, it

is conceivable that when in direct conflict with a segment, e.g., a consonant with a particular laryngeal demand, the laryngeal event for the phonation register and the consonant would be sequential rather than blended. But such sequencing is no different in nature as coarticulation resistance to be discussed next.

3.2.3. Preservation of segmental identity

The advantage of CV co-onset is not only that it provides an accurate time reference, but also that it guarantees the release of the consonant into the following vowel that typically has a wider vocal tract opening, which, as pointed out by Mattingly (1981), enhances the recoverability of both. On the other hand, a potentially harmful consequence of co-onset is that coproduction may weaken the perceptual cues for some consonants. Due to their shorter duration than that of vowels, consonants run the risk of being hidden by the coproduced vowel. Although formant transitions provide important cues for consonants, the cues are not always highly effective by themselves, especially for stops, as found by Brancazio & Fowler (1998). There is therefore a pressure for the preservation of the consonants' identity, especially their place of articulation. Such pressure may be behind three phenomena related to coarticulation: coarticulation resistance, locus equation and trough effect.

Coarticulation resistance refers to the finding that the cross-consonant coarticulation described by Öhman (1966) is blocked to a different extent by the intervening consonant (Bladon & Al-Bamerni, 1976; Recasens, 1984a, b, 1985). Recasens proposed that the extent to which a consonant or a vowel resist coarticulation is related to the extent it constraints the tongue dorsum. This is supported by Fowler and Brancazio (2000), who found also that despite the difference in the strength of coarticulation resistance among the consonants, there is no difference in the temporal onset of the vowel-to-vowel effects. This indicates again the stability of the temporal alignment of C and V at the syllable onset as assumed in the time structure model.

An even more extreme case of coarticulation resistance is seen in Russian, where consonants with the same place of articulation differ in terms of palatalizaton/velarization. Because the targeted tongue body shapes for these consonants and the resulting acoustic patterns are contrastive, the pressure to fully realize them is naturally high. As observed by Öhman (1966: 164) "[a]n interesting acoustic feature of palatalization was observed when the formant transitions following the palatalized stops were compared with those following the corresponding unpalatalized stops in the same vowel context. The former transitions were usually convex upwards and the latter were convex downwards." Öhman remarked that "this is what would be expected in general if the release of the palatalized stops were associated with a forward motion of the point of maximum constriction of the tract and if the unpalatalized variants involved a backward motion" (p. 164). In other words, the pressure to make the palatalization/velarization difference clearly audible is so strong that speakers make sure that the maximum constriction is realized *after* voice onset. Pending further investigation, it is also possible that the palatalizaton/velarization distinction is actually realized as the first V of the syllable, i.e., it is this vocalic element rather than the nuclear vowel that shares the onset with the consonant.

Locus equation refers to the finding that for each initial consonant the F2 measured at the voice onset of the following vowel is linearly related to the F2 at the "center" of the vowel (Lindblom, 1963b; Sussman, McCaffrey & Matthews, 1991). Such linear relation has been argued to arise due to an evolutional adaptation to facilitate auditory processing of consonant place of articulation (Sussman et al., 1998). Fowler (1994) argues, however, that the linearity

is related to the invariance in coarticulation resistance across different vowels. In light of the time structure model, a locus equations can be viewed as largely a part-whole correlation. This is because the two F2 measurements are taken from different points along the same movement toward the vowel target, which originates about 26-48 ms before the consonant closure and is warped by "coarticulation resistance" of the consonant. Because the voice onset occurs roughly half way through this movement, the correlation is necessarily quite linear.

The trough effect, first reported by Houde (1967), refers to the phenomenon that in a sequence such as [ibi], the tongue dorsum temporarily lowers during the stop closure instead of remaining as high as in the flanking [i]. Although the trough effect has not been explicitly linked to coarticulation resistance, the two are likely related. That is, both seem to be related to the pressure to preserve consonant identity. Such pressure could be overridden, however, in vowel-harmony languages such as Turkish, by the pressure from a phonological rule that mandates cross-consonant vowel agreement in features such as lip-rounding (Boyce, 1990). Relating this to the findings of Fowler and Bracazio (2000), it is likely that the difference between vowel-harmony languages and other languages is not a matter of segmental alignment, but rather a matter of degree of coarticulation resistance.

Closely related to coarticulation resistance and the trough effect is the finding by Wood (1996) that even at the syllable initial position, if the consonant requires an articulatory movement that conflicts with that of the following vowel, the two movements are sequenced rather than blended into a single compromised movement. Note that such sequencing does not necessarily block CV co-onset, since there are often concomitant C and V movements that are not conflicting with each other. As explained by Goldstein and Fowler (2003:21), while "[a]n open vowel coarticulating with /b/ may pull the jaw down…", "lip closure, the essential property of the labial stop gesture is nonetheless achieved" by raising the lower lip further like what has been found to happen when the jaw is deliberately pulled down (Kelso et al., 1984). Thus it is possible that co-onset of C and V and recoverability of all involved segments are achieved with different coordination strategies.

3.2.4. Sequential nature of syllable coda

Compared to the co-onset of CV, sequential offset of VC in a closed syllable may seem to be an even stronger hypothesis. As we will show, however, not only is there already some empirical evidence for it, but also it is virtually inevitable theoretically. Empirically, one of the most direct pieces of evidence is from Lindblom et al. (2002), in which they reported very different trough effect patterns in open and closed syllables. Figure 8 displays one of the plots from Figure 4 in Lindblom et al. (2002) showing F2 values measured at four points in /i.bi/ and /ib.i/ (where the period represents the syllable boundary). The four measurement points are V1 mid point, V1 offset, V2 onset (or at the burst release for coda stops), and V2 mid point. These formant values suggest that coda /b/ is not overlapped with either the preceding or the following vowel. The lack of overlap with the preceding vowel is indicated by the similarity of the F2 values at V1 offset (second measurement point) between the open and closed syllables. Had the first /i/ in /ib.i/ been overlapped by the coda /b/, F2 at V1 offset would have been much lower than in /i.bi/. The lack of overlap with the following vowel is indicated by the third measurement point. There F2 is much lower in /ib.i/ than in /i.bi/. This suggests that during the closure the vocal track continues to approach the /b/ target, and this approximation is not blocked by the movement toward the following /i/, indicating that the

latter probably has just started at the /b/ release.[6]

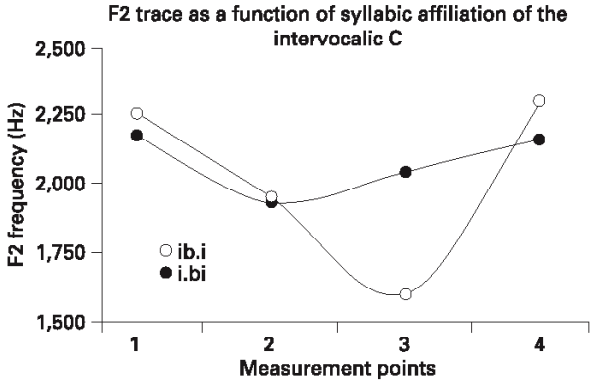

Figure 8. A plot of F2 measurements in /ib.i/ and /i.bi/ in Figure 4 of Lindblom et al. (2002), courtesy of Lindblom. See text for explanation.

A second piece of empirical evidence is related to the phenomenon of resyllabification. That is, the coda consonant of a closed syllable often becomes the initial consonant of the following vowel-onset syllable. A weaker form of resyllabification is known as ambisyllabification (Gick, 2003), by which the consonant appears to behave as both the coda of the preceding syllable and onset of the following syllable. Stetson (1951) describes a series of experiments in which he found that a repetitive sequence like "pup pup …", when said at an increasing rate, reliably changes into "pu pu…". While Stetson explained the phenomenon in terms of simplification, i.e., eliminating the coda consonant to make the movements more "in-phase", Kelso, Saltzman and Tuller (1986) show that a similar shift into "pi pi …" also occurs when speakers start from "ip ip …". They therefore argue against the in-phase account by Stetson. Nevertheless, they recognize and demonstrate that such consonant affiliation shift is essentially the same as the phase-shift that occurs when a repetitive anti-phase two-finger movement shifts to an in-phase movement at higher speed (Kelso, 1984). But the unanswered question is what has become "in-phase" when "ip ip …" shifts to "pi pi …" In light of the time structure model, and as we have seen in Figure 8, what has happened is that the sequential articulation in "ip" has shifted into the co-onset articulation in "pi". Such a shift brings two advantages, as shown in Figure 9. The first is that co-onset gives both C and V more execution time. This advantage becomes more and more critical as the rate of articulation increases. At some point, it would be simply impossible to execute two unidirectional movements in succession. The second advantage is that co-onset also makes the consonantal movement and vocalic movement more synchronized, as they now share the same onset. In other words, what has become “in phase” is the onset of the two independent movements. Such synchrony, it seems to us, is what makes CV the most stable structure in speech.

[6] In Lindblom et al.’s study “subjects were instructed to produce closed syllables with a clear release + pause” (p. 253). We believe that such an instruction was crucial, for without it, the coda consonant may easily be resyllabified with the following vowel.

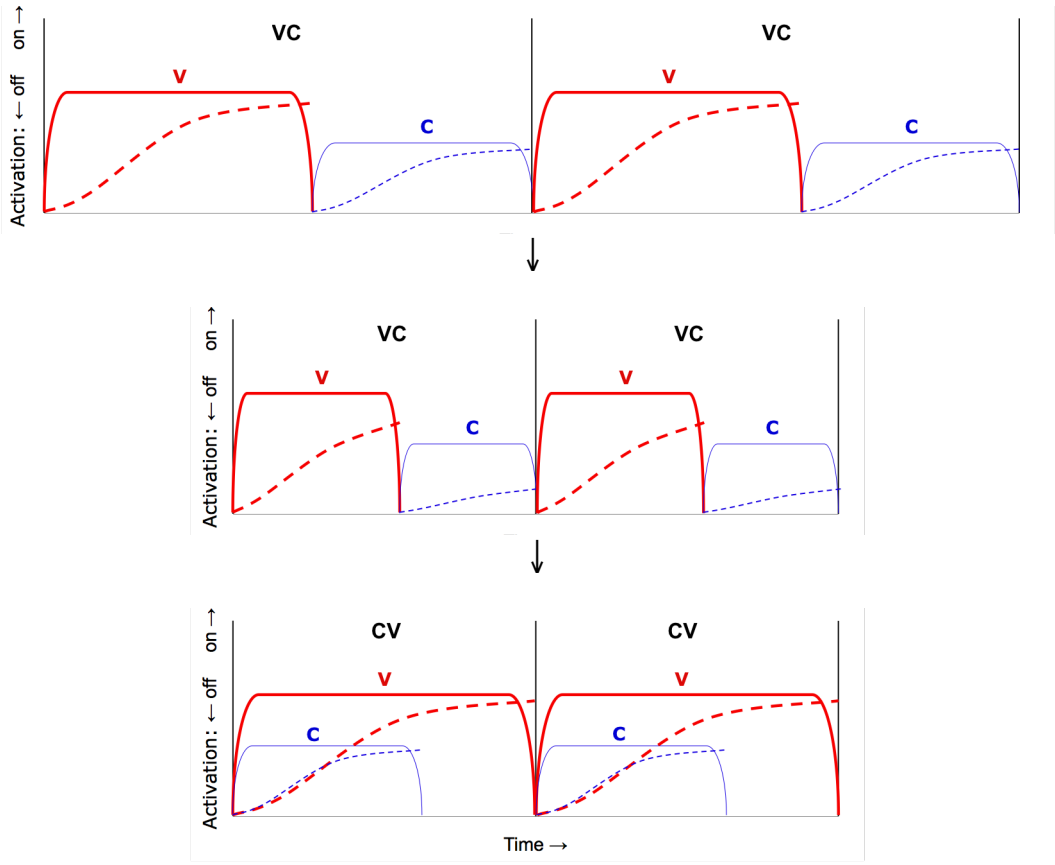

Figure 9. A schematic illustration: VC to CV transition gives both C and V more execution time. The dashed curves represent the trajectories of articulatory movement toward the C and V targets.

Theoretically, the unidirectionality of articulatory movements as stated in [2] makes it virtually impossible for a coda consonant to overlap with the preceding vowel. Consonants, by definition, have narrower vocal tract constrictions than vowels, as recognized in [3]. As a consequence, the articulatory movement toward a consonantal target is necessarily a closing gesture, while that toward a vocalic target following an onset consonant necessarily an opening gesture. Thus a coda consonant is in direct conflict with the preceding vowel in terms of the trajectory of vocal tract movement. The only way to realize a coda consonant is to start the closing movement when the opening movement of the vowel has ended. The only exception to this are cases in which part of the consonantal articulation is in less direct conflict with the vowel articulation. This seems to be the case, for example, with velum lowering in a coda nasal, which may be why the loss of a coda nasal is often accompanied by nasalization of the preceding vowel. Nevertheless, as shown by Bell-Berti and Krakow (1991), the velum movements in a syllable like /lan/ in English are in fact sequentially produced, i.e., successively approaching the specified positions for /l/, /a/ and /n/. This means that unless a vowel is fully nasalized, even the velum movement may be largely sequential from a vowel to a coda nasal.

The principles of co-onset and sequential offset may also help us better understand the phenomenon of ambisyllabification. Sproat and Fujimura (1993) and Gick (2003) have reported that the tongue tip or lip movement may lag behind the tongue body movement in a syllable final /l/ or /w/ when they become ambisyllabic. Note that, when a consonant becomes ambisyllabic, by definition they should exhibit characteristics of both coda and

onset C. It is possible that the tongue body movement is still sequentially aligned after the preceding V, but the tongue tip movement may have been realigned as part of the onset of the following syllable. If so, this would be analogous to the insertion of a consonant before an otherwise vowel-onset syllable: if not available at the lexical level, a C is inserted to facilitate co-onset at the articulation level. Future research into this issue could be facilitated by testing the principles of co-onset and sequential offset, as it could generate predictions that are more testable than before.

3.2.5. Tonal alignment — A new perspective

That lexical tones are synchronized with the entire syllable, as stated in [9], has much empirical support from recent research (Xu, 1998, 2001). As argued in Xu & Wang (2001), the synchronization is not only due to the general coordination constraints as found by Kelso (1984) and many subsequent studies, but also due to the fact that pitch changes are quite slow as compared to the typical syllable duration. According to Xu & Sun (2002), it takes at least 124 ms for an average speaker to complete a 4-semitone pitch rise or fall. This would take up much of the duration of a typical syllable in Mandarin (around 180 ms in Xu, 1999) and likely in many other languages as well. Thus full synchronization is the only possible stable coordination pattern between tone and syllable. As for the synchronization between phonation register and the syllable, we are unaware of any direct evidence for it. But such synchronization is highly probable given that the change of laryngeal state related to phonation type is unlikely to be much faster than that related to pitch change. Nevertheless, future research is needed to examine the time course of phonation type variation relative to the syllable in languages that use phonation register contrastively.

Although the time structure model of the syllable has offered a more solid foundation for our previous proposal about tone-syllable synchronization (Xu & Wang, 2001), the new understanding that all segments actually start roughly 26-48 ms earlier than their conventional acoustic landmarks seems to create a new problem in regard to the detailed F_0 alignment. The typical finding of recent studies is that, other things being equal, certain F_0 peaks and valleys are aligned relative to the *conventional* syllable onset or offset. Shifting the syllable boundaries leftward would mean shifting the F_0 turning points rightward, thus making them seem less synchronous with the syllable. While appearing to be a new challenge, this actually offers a solution to an old puzzle that has so far defied explanation. As can be seen in Figure 10, because the H tone in syllables 1 and 3 is followed by the L tone, F_0 should start to drop when the approximation of the [high] target is terminated according to the TA model. However, the drops apparently have started well before the conventional syllable boundary, as indicated by the two solid arrows. According to Xu (1999), the F_0 peak in High preceded by Low or Falling and followed by Low occurs 24 ms before the nasal murmur onset. Such an early drop could easily be interpreted as due to anticipatory coarticulation, which would be in conflict with the sequential nature of target approximation. In light of the time structure model of the syllable, this early drop is no longer a puzzle. Instead, it constitutes evidence that the onset of the syllable, hence also the onset of the tone, actually starts well before the conventional syllable boundary.

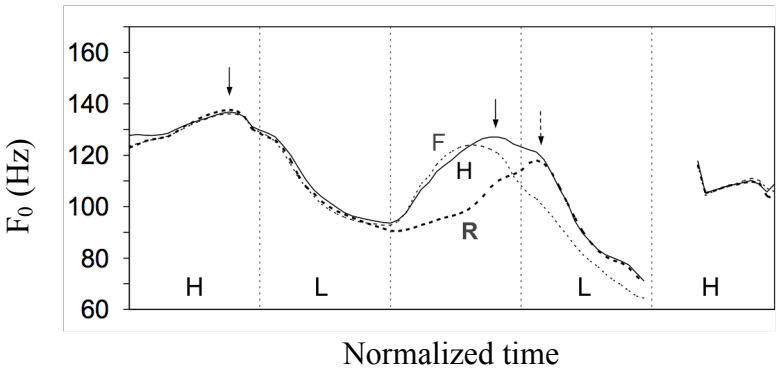

Figure 10. Time-normalized mean F_0 of Mandarin H L *x* L H tone sequences, where *x* is H, R or F. Each curve is an average of 20 tokens uttered by 4 male speakers (data from Xu, 1999). Vertical lines indicate conventional syllable boundaries. The solid arrows point to F_0 peaks related to the H tone that occur well before the conventional syllable offsets. The dashed arrow points to the delayed F_0 peak related to the R tone.

Moving the syllable boundary leftward also seem to creates a problem for the dynamic tones such as Rising and Falling. As indicated by the dashed arrow in Figure 10, the F_0 peak related to the Rising tone typically occurs after the conventional syllable boundary. Moving the syllable boundary leftward would make the "delay" appear even bigger. But as we have argued before (Xu, 2002, in particular), turning points are not direct correlates of tones. Instead, they are just the consequences of syllable-synchronized target approximation, and as such should be treated only as *indicators* of tonal alignment. In fact, an even more straightforward indicator, based on the TA model, should be the velocity of F_0 movement, as it directly reflects the nature of a movement at any particular moment in time. The velocity of F_0 is the instantaneous rate of change of F_0, which is mathematically its derivative. Numerically, velocity of F_0 can be computed by taking the difference between every two adjacent F_0 values, as shown in the following equation,

$$F_{0j}' = (F_{0j+1} - F_{0j}) / (t_{j+1} - t_j)$$

where F_{0j}' is the velocity value of *i*th F_0, and t_j is the time of the *i*th F_0.

When velocity is plotted as a function of time, the shape of the curve may help us understand the nature of the corresponding movement, according to Nelson (1983). For example, a unidirectional movement that starts at a static position and ends at another has a unimodal velocity trajectory that starts and ends at 0. Figure 11 shows a sine wave and its corresponding velocity curve. In the figure the curves are divided into five different intervals. The three intervals divided by the vertical dashes and indicated by the labels on the top of the graph are based on Nelson's above definition for unidirectional movement. During the first and last intervals, displacement approaches the minimum value (solid curve), and the velocity during those intervals shows negative unimodal profiles, both starting and ending at zero.

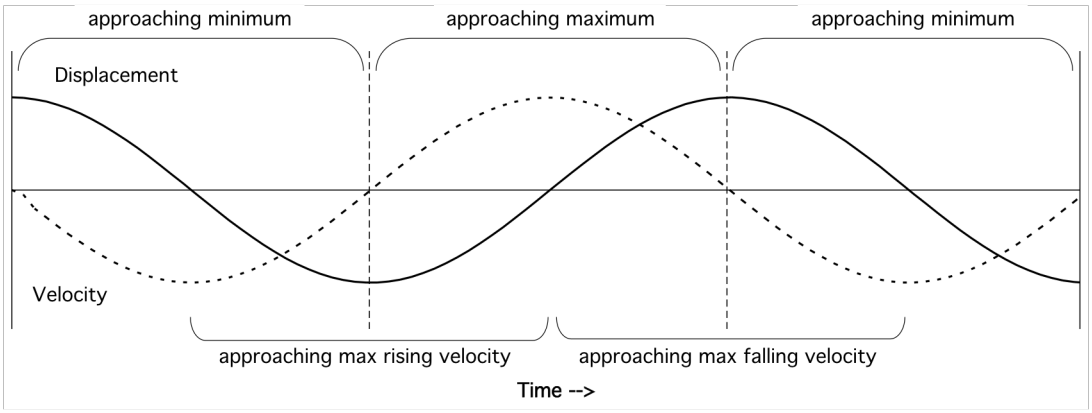

Figure 11. Velocity versus displacement. The solid curve is the displacement trajectory of a sine wave as a function of time. The dashed curve is the trajectory of instantaneous velocity of the sine wave. The trajectories can be divided into different intervals, during each of which the minimum displacement, maximum displacement or maximum velocity is approached.

Figure 12 displays mean F_0 curves of three tone sequences from Xu (1999) (solid) and their corresponding velocity trajectories (dashed). The short vertical bars on the curves mark the conventional syllable boundaries. Similar to the High tone in Figure 10, in Figure 12a, the F_0 of the second High tone has reached a peak well before the "end" of the syllable. The same is true of the High tone following the Falling tone in Figure 12b. Such early peaking is also reflected in the velocity curve in Figure 12a, which has become clearly negative by the "end" of a High-tone syllable. To compare to the alignment based on the time structure model, the locations 50 ms before the conventional syllable boundaries are marked by the dashed vertical lines in Figure 12. As can be seen in the second High tone in Figure 12a, not only are the dashed lines closer to the F_0 peak than the short vertical bars, but also it is closer to the zero crossings in the velocity curve, thus better fitting the unimodal velocity profile of unidirectional movement. Similar benefit can be seen in Figure 12b, although there 50 ms seems to be too large a shift in this case. The reason is likely that the F tone before the High tone has generated a strong negative slope that has to be first reversed when approaching the High tone target. This leaves it not enough time for reaching an F_0 asymptote and corresponding velocity zero crossing by the newly conceived syllable offset.

Unlike the static tones such as High and Low, the dynamic tones such as Rising and Falling present cases that are beyond Nelson's (1983) definition of an individual unidirectional movement. But the principle of Nelson's definition can be extended to the dynamic tones. That is, a unidirectional movement can also be one that reaches a desired velocity that is non-zero. This means that the velocity of a dynamic tone should end at either a positive or a negative value when its execution terminates, as is the case in the two intervals indicated by the bottom labels in Figure 11. In the left interval, when displacement reaches zero at the end of the interval, velocity has just reached its maximum. This should resemble the case of the Rising tone. In the right interval displacement and velocity show the opposite patterns, which should resemble the case of the Falling tone. As examples, in Figure 12a, where the third tone is Rising, at the time of the second short bar, F_0 has almost reached the peak. The dashed curve shows that, however, velocity has been going downward for a while and has almost reached 0. But the vertical line on the left is actually closer to where velocity reaches the peak. A similar situation can be seen in Figure 12c where the third syllable has the Falling tone. By the second short bar, the velocity has virtually reached 0, whereas the greatest

negative value is near the dashed vertical line on the left. This is also true for the first Falling tone in both Figure 12c and Figure 12b.

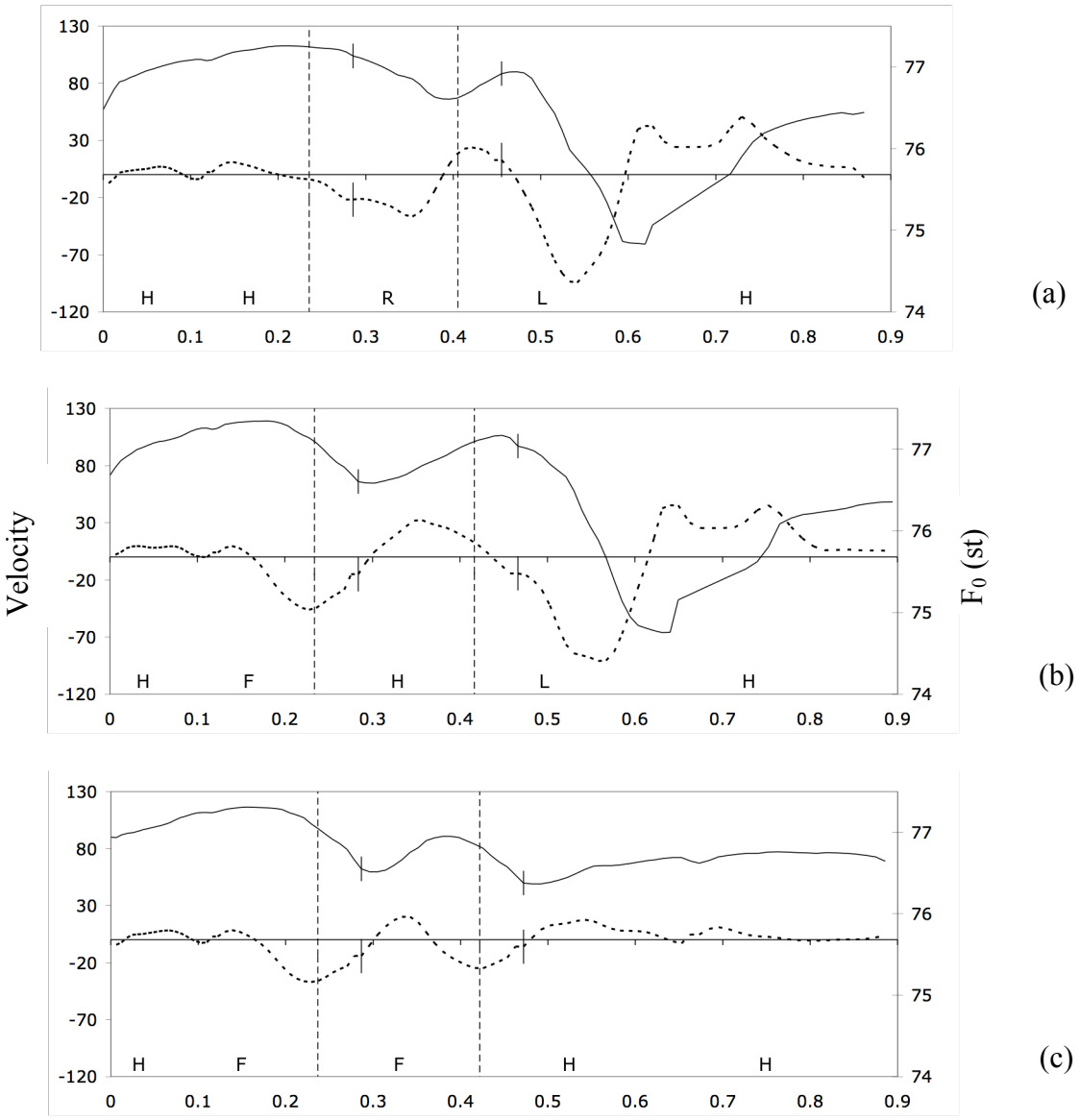

Figure 12. Solid curves: Mean F_0 trajectories of five-tone sequences from Xu (1999), averaged over 5 repetitions by 8 speakers; Dashed curves: velocity trajectories computed from the F_0 trajectories. The short vertical bars mark the conventional syllable boundaries of the third syllable. The long vertical lines are placed 50 ms before the short bars.

An important caveat is in order regarding the application of Nelson's definition and its extension to dynamic tones. That is, as mentioned in regard to the third High tone in Figure 12b, the ending velocity of the preceding tone has an impact on both the F_0 and velocity of the current tone. The magnitude of the impact depends on both the underlying targets of the two adjacent tones and their assigned articulatory strength, as found in Chen & Xu (in press). This means that the velocity of a static tone may not always start and end at zero crossing and the velocity of a dynamic tone may not always end exactly at a turning point. The precise alignment can only be simulated through quantitative modeling, which we are currently working on but is beyond the scope of the present paper (Prom-on, Xu & Thipakorn,

forthcoming).

A further note is that the velocity analysis should be applicable to formant movements as well. But an important difference between tonal movements and segmental movements is that the former is much slower than that latter. According to Xu & Sun (2002) and an ongoing study (Xu, in press), F_0 movements are about half as fast as the segmental movements. Apparently, much research is still needed to find out the exact speed difference between different kinds of articulatory movements and its impact on the alignment measurements in terms of both displacement and velocity.

3.3. Implications for speech perception

The time structure model proposed in this paper may have implications for the understanding of speech perception. In particular, many previous findings about speech perception will now need to be looked at in a new light. For example, Lee (2000) shows in a gating experiment that Mandarin listeners can identify a tone even if they hear the utterance only up to the initial sonorant of the target syllable. Such robust tone recognition can be now interpreted as because by the end of the sonorant, listeners likely have heard more than 100 ms of the target tone (48 ms + conventional duration of sonorant). The same reinterpretation can be applied to the finding of Xu (1994) that perception of tone uttered in connected speech can be drastically facilitated by the presence of the surrounding tones. For segmental sounds, van Son and Pols (1999) have shown that adding the CV transition in front of the steady-state portion of a vowel improves perception of the vowel by Dutch listeners, whereas adding the VC transition at the back of the vowel does not lead to improvement. Again this can now be interpreted as because more of the vowel itself is actually included when the CV transition is added. And that the VC transition does not lead to better vowel perception is likely because the unidirectional movement there is toward the following C. Further confirming this understanding is the finding by Warner et al. (2005) in a gating experiment that Dutch vowels in CV sequences are already fairly well perceived by one-third of the way through their conventional durations. Most of the above studies (with the exception of Lee, 2000), however, used relative rather than absolute time in preparing the stimuli. Thus it is not possible to assess how consistently in actual timing these findings are with the time structure model. Further research is needed.

4. Summary and conclusion

We have explored in this paper the implication of consistent F_0 alignment with respect to the syllable for understanding the timing and alignment of speech sounds in general. Through the discussion of the findings of Xu and Liu (2002) and Liu and Xu (2003), we have realized that the conventionally understood syllable boundaries need to be reconsidered. The reconsideration in light of the Target Approximation model (Xu & Wang, 2001) has led to the proposal of a new model of basic temporal organization of speech sounds – the time structure model of the syllable. The model assumes that the syllable specifies the temporal alignment of all the basic phonetic elements referred to as phones, which include consonants, vowels, tones and phonation registers. The phones are proposed to be temporally organized by the syllable under three principles: co-onset of initial C and V, sequential offset of coda C, and full synchronization of tone and phonation register with the syllable.

A critical factor behind the explicitness of the time structure model is the strict definition of phone, which limits it to only articulatory movements *toward* a target. This definition is inspired by the findings of contextual tonal variation in (Xu, 1997, 1999) and the findings of

Xu and Liu (2002) and Liu and Xu (2003). Following this definition, other than the true coarticulation between initial C and V due to co-onset, there is little or no coarticulation between other adjacent segments: no anticipatory C to V coarticulation, no cross-consonantal V-to-V coarticulation, and no carryover coarticulation of any kind.

The lack of articulatory overlap between coda C and the preceding V makes the coda elements weak and unstable: weak because undershoot occurs more easily under greater time pressure; unstable because the time pressure may often lead to their deletion, merger into the preceding vowel and resyllabification into the following syllable. Thus the universal stability of the CV structure, in light of the time structure model, is the natural outcome of the basic temporal alignment pattern of the syllable.

The time structure model not only offers a new perspective for the understanding of segmental variability, it also provides a more principled way of understanding the stable tonal alignment found in recent research. That is, because the syllable is the basic time structure for speech sounds, a tone can be aligned only relative to the syllable rather than to any individual segments in the syllable.

The new model also has implications for the understanding of speech perception. In particular, many of the previously reported cases of perceptual dependence on coarticulatory information can now be understood as due to hearing the complete rather than partial movements toward the phonetic targets.

Finally, although we have proposed the syllable to be the most basic time structure in speech, we are by no means ruling out possible existence of other, smaller or larger alignment and/or timing schemes. On the smaller side, phones may have internal timing specifications. In an aspirated stop, for example, the glottal opening gesture needs to terminate later than in an unaspirated stop. On the larger side, the foot may be a higher level of time structure than the syllable. Although the foot is taken as given in many phonological theories of speech prosody (e.g., Liberman & Prince, 1977; Pierrehumbert, 1980), and investigated in recent research (e.g. Hirst & Bouzon, 2005; van Santen, Klabbers & Mishra, this volume) what is needed is a model that is as explicit as the time structure model of the syllable. And for that, more experimental data are needed.

5. Acknowledgement

We would like to thank Daniel Hirst and Jan van Santen for their insightful reviews. This work was supported in part by NIH Grant DC03902 and NIH Grant DC006243.

References

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

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

Beckman, M. E. and Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook* **3**: 255-309.

Bell-Berti, F. and Harris, K. S. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding. *JASA* **65**: 1268-1270.

Bell-Berti, F. and Harris, K. S. (1981). A temporal model of speech production. *Phonetica* **38**: 9-20.

Bell-Berti, F. and Krakow, R. A. (1991). Anticipatory velar lowering: A coproduction account. *Journal of the Acoustical Society of America* **90**: 112-123.

Bell-Berti, F., Krakow, R. A., Gelfer, C. E. and Boyce, S. (1995). Anticipatory and carryover effects: Implications for models of speech production. *Producing Speech: Contemporary Issues. For Katherine Safford Harris*. F. Bell-Berti and L. J. Raphael. New York: AIP Press.

Benguerel, A.-P. and Cowan, H. A. (1974). Coarticulation of upper lip protrusion in French. *Phonetica* **30**: 41-55.

Bladon, R. A. W. and Al-Bamerni, A. (1976). Coarticulation resistance of English /l/. *Journal of Phonetics* **4**: 135-150.

Boyce, S. (1990). Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America* **88**: 2584-2595.

Boyce, S. E., Krakow, R. A. and Bell-Berti, F. (1992). Phonological underspecification and speech motor organization. *Phonology* **8**: 210-236.

Brancazio, L. and Fowler, C. A. (1998). On the relevance of locus equations for production and perception of stop consonants. *Perception and Psychophysics* **60**: 24–50.

Browman, C. P. and Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook* **3**: 219-252.

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology* **6**: 201-251.

Browman, C. P. and Goldstein, L. M. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Les Cahiers de l'ICP, Bulletin de la Communication Parlée* **5**: 25–34.

Byrd, D. (1996). A phase window framework for articulatory timing. *Phonology* **13**: 139-168.

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

Chen, Y. and Xu, Y. (in press). Production of weak elements in speech -- Evidence from F_0 patterns of neutral tone in standard Chinese. To appear in *Phonetica*.

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

D'Imperio, M. (2001). Focus and tonal structure in Neapolitan Italian. *Speech Communication* **33**: 339-356.

D'Imperio, M. (2002). Language-Specific and Universal Constraints on Tonal Alignment: The Nature of Targets and "Anchors". Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France: 101-106.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Farnetani, E. and Recasens, D. (1999). Coarticulation models in recent speech producttion theories. Coarticulation: Theory, Data and Techniques. W. J. Hardcastle and N. Newlett. Cambridge: Cambridge University Press: 31-65.

Feldman , A. G. (1966). Functional tuning of the nervous system with control of movement or maintenance of a steady posture—II Controllable parameters of the muscles. *Biophysics* **11**: 565-578.

Feldman , A. G. (1986). Once more on the Equilibrium-Point hypothesis (lambda Model) for motor control. *Journal of Motor Behavior* **18**: 17-54.

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics* **55**: 597–610.

Fowler, C. A. and Brancazio, L. (2000). Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation. Language and Speech **43**: 1-41.

Fujimura, O. (1994). C/D Model: A computational model of phonetic implementation. *Language and Computations*. E. S. Ristad. Providence, RI: American Math Society**:** 1-20.

Fujimura, O. (2000). The C/D model and prosodic control of articulatory behavior. *Phonetica* **57**: 128-138.

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

Goldsmith, J. A. (1976). *Autosegmental Phonology*, Ph. D. dissertation, MIT. [Published in 1979 by Garland Press in New York].

Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology*. Oxford: Blackwell Publishers.

Gick, B. (2003). Articulatory correlates of ambisyllabicity in English glides and liquids. *Papers in Laboratory Phonology VI: Constraints on Phonetic Interpretation*. J. Local, R. Ogden and R. Temple. Cambridge: Cambridge University Press.

Ginésy, Michel & Hirst, D.J. (1975). Formant transitions and pitch-change in English diphthongs. *Travaux de l'Institut de Phonétique d'Aix* 2, 141-148

Goldstein, L. M. and Fowler, C. (2003). Articulatory phonology: a phonology for public language use. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. A. S. Meyer and N. O. Schiller. Berlin: Mouton de Gruyter.

Hardcastle, W. J. and Hewlett, N. (1999). Coarticulation: Theory, Data and Techniques. Cambridge: Cambridge University Press.

Hirst, D. and Bouzon, C. (2005). The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). *Proceedings of Interspeech 2005*, Lisbon, Portugal**:** 29-32.

Houde, R. A. (1967). *A study of tongue motion during selected speech sounds*. Ph.D. dissertation, University of Michigan.

Huffman, M. K. and Krakow, R. (1993). Phonetics and Phonology 5: Nasals, Nasalization and the Velum. San Diego: Academic Press.

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

Joos, M. (1948). Acoustic Phonetics. *Journal of the Acoustical Society of America* **24, suppl.**: 1-136.

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

Kelso, J. A. S., Saltzman, E. L. and Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics* **14**: 29-59.

Kelso, J. A. S., Southard, D. L. and Goodman, D. (1979). On the nature of human interlimb coordination. *Science* **203**: 1029-1031.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E. and Fowler, C. A. (1984). Functionally-specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. Journal of Experimental Psychology: Human Perception and Performance **10**: 812-832.

Kozhevnikov, V. A. and Chistovich, L. A. (1965). *Speech: Articulation and Perception*. Washington, DC: Joint Publications Research Service.

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

Kühnert, B. and Nolan, F. (1999). The origin of coarticulation. *Coarticulation: Theory, Data and Techniques*. W. J. Hardcastle and N. Newlett. Cambridge: Cambridge University Press**:** 7-30.

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

Ladd, D. R., Faulkner, D., Faulkner, H. and Schepman, A. (1999). Constant "segmental anchoring" of F0 movements under changes in speech rate. *Journal of the Acoustical Society of America* **106**: 1543-1554.

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

Ladefoged, P. (1983). The linguistic use of different phonation types. *Vocal Fold Physiology: Contemporary Research and Clinical Issues*. D. Bless and J. Abbs. San Diego: College-Hill Press**:** 351-360.

Lee, C.-Y. (2000). *Lexical Tone in Spoken Word Recognition: A View from Mandarin Chinese*. Ph.D. Dissertation, Brown University.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry* **8**: 249-336.

Lindblom, B. (1963a). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* **35**: 1773-1781.

Lindblom, B. (1963b). On vowel reduction. *Speech Transmission Laboratory, The Royal Institute of Technology, Sweden*. **Report No. 29**.

Lindblom, B., Sussman, H. M., Modarresi, G. and Burlingame, E. (2002). The trough effect: Implications for speech motor programming. *Phonetica* **59**: 245-262.

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

Löfqvist, A. and Gracco, L. (1999). Interarticulator programming in VCV sequences: Lip and tongue movements. *Journal of the Acoustical Society of America* **105**: 1864-1876.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavial and Brain Sciences* **21**: 499–546.

Mattingly, I. G. (1981). Phonetic representation and speech synthesis by rule. *The cognitive representation of speech*. T. Myers, J. Laver and J. Anderson. North-Holland: Amsterdam**:** 415-420.

Menzerath, P. and de Lacerda, A. (1933). *Koartikulation, Seuerung und Lautabgrenzung*. Berlin and Bonn: Fred. Dummlers.

Moon, S.-J. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America* **96**: 40-55.

Nelson, W. L. (1983). Physical principles for economies of skilled movements. Biological Cybernetics 46: 135-147.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America* **39**: 151-168.

Perrier, P., Ostry, D. J. and Laboissière, R. (1996). The Equilibrium-Point Hypothesis and its Application to Speech Motor Control. *Journal of Speech and Hearing Research* **39**: 365-377.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. Ph.D. dissertation, MIT, Cambridge, MA. [Published in 1987 by Indiana University Linguistics Club, Bloomington].

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

Prom-on, S., Xu, Y. and Thipakorn, B. (forthcoming) Quantitative Target Approximation Model: Simulating Underlying Mechanisms of Tones and Intonations.

Recasens, D. (1984a). Timing constraints and coarticulation: Alveolo-palatals and sequences of alveolar + [j] in Catalan. *Phonetica* **41**: 125-139.

Recasens, D. (1984b). Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America* **76**: 1624-1635.

Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. *Language and Speech* **28**: 97-114.

Saltzman, E. L. and Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review* **94**: 84-106.

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology* **1**: 333-382.

Sproat, R. and Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* **21**: 291-311.

Stetson, R. H. (1951). Motor Phonetics: a Study of Speech Movements in Action (2nd edition). Amsterdam: North Holland.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* **111**: 1872-1891.

Sussman, H. M., Fruchter, D., Hilbert, J. and Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavial and Brain Sciences* **21**: 241–299.

Sussman, H. M., McCaffrey, H. A. and Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America* **90**: 1309-1325.

Sussman, H. M. and Westbury, J. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research* **24**: 16-24.

Tuller, B.; Kelso, J. A. S., 1984. The timing of articulatory gestures: Evidence for relational invariants. *Journal of the Acoustical Society of America* **76**: 1030-1036.

van Santen, J. P. H. and Möbius, B. (2000). A quantitative model of f0 generation and alignment. *Intonation: Analysis, Modelling and Technology*. A. Botinis: Kluwer Academic Publishers**:** 269-288.

van Son, R. J. J. H. and Pols, L. C. W. (1999). Perisegmental speech improves consonant and vowel identification. *Speech Communication* **29**: 1-22.

van Santen, J. P. H., Klabbers, E. and Mishra, T. (this volume). Measurement of Pitch Movement Alignment. *Ilatian Journal of Linguistics*.

Warner, N., Smits, R., McQueen, J. M. and Cutler, A. (2005). Phonological and statistical effects on timing of speech perception: Insights from a database of Dutch diphone perception. *Speech Communication* **46**: 53-72.

Wood, S. A. J. (1996). Assimilation or coarticulation? Evidence from the temporal co-ordination of tongue gestures for the palatalization of Bulgarian alveolar stops. *Journal of Phonetics* **24**: 139-164.

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

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

Xu, Y. (in press). How often is maximum speed of articulation approached in speech? To be presented at the 153rd meeting of The Acoustical Society of America.

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

Xu, Y. and Liu, F. (in press). Determining the temporal interval of approximants with help of F0 contours. To appear in *Journal of Phonetics*.

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.

Xu, Y. and Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* 33: 159-197.