In O. Niebuhr and H. Pfitzinger (eds.) *Understanding Prosody – The Role of Context, Function, and Communication.* Walter de Gruyter, pp. 1-26.

Yi Xu & Fang Liu (London)

# Intrinsic coherence of prosodic and segmental aspects of speech

## 1 Introduction

It has been long held that segmental and prosodic aspects of speech belong to very different domains. Segments refer to vowels and consonants which constitute the phonemes that make up the words of speech, and they are what most writing systems record. As such they are viewed as the core components of speech. Prosodic components, also known as suprasegmentals, refer to properties like stress, timing, duration and intonation. According to Lehiste (1996:227): *"segmental features characterize speech sounds, and suprasegmental features are properties of speech sounds or their sequences that are simultaneously present, that do not change the distinctive phonetic quality of the speech sounds, but do modify the sounds in a way that may change the meaning of the utterance"*. In this definition, speech sounds refer only to consonants and vowels, and suprasegmental properties are treated as properties that accompany the speech sounds.

Although speech research used to be dominated by segmental studies that are concerned with only consonants and vowels, in recent years speech prosody has become an increasingly active area of research. Yet prosody is still typically viewed as not only separate from segments, but also involving rather different articulatory and perceptual mechanisms. The goal of this paper is to demonstrate that this conceptual divide is unwarranted. We will show that segments and prosody are in fact intrinsically coherent, because they share essentially the same articulatory dynamics, and are both involved in encoding lexical as well as non-lexical communicative functions. We will make our case by first reviewing evidence for the basic articulatory mechanisms of tonal dynamics, and then exploring how the same principles can be extended to segmental and intonational dynamics. Finally, we will examine how tonal as well as segmental reduction can be explained in terms of both articulatory dynamics and functional requirements.

## 2 Tonal dynamics

Lexical tones are pitch patterns that are used to distinguish words that are otherwise identical in their phonetic composition (Chao 1968; Yip 2002).

Because the primary acoustic correlate of tone is F0, which is one-dimensional rather than multi-dimensional as in the case of vowels and consonants, tone offers a unique vantage point for the understanding of basic articulatory dynamics of speech. Figure 1 displays three of the lexical tones of Mandarin, High (H), Rising (R) and Falling (F), spoken in different tonal contexts in syllable 3 (between the 3rd and fourth vertical lines). As we can see, there are four rather different F0 onsets in syllable 3 in each graph, which are also ending points of the preceding tones. During the course of each tone, however, all the four contours gradually converge to a trajectory that is appropriate for the underlying tone: high-level for H, rising for R and falling for F, as indicated by the arrows.

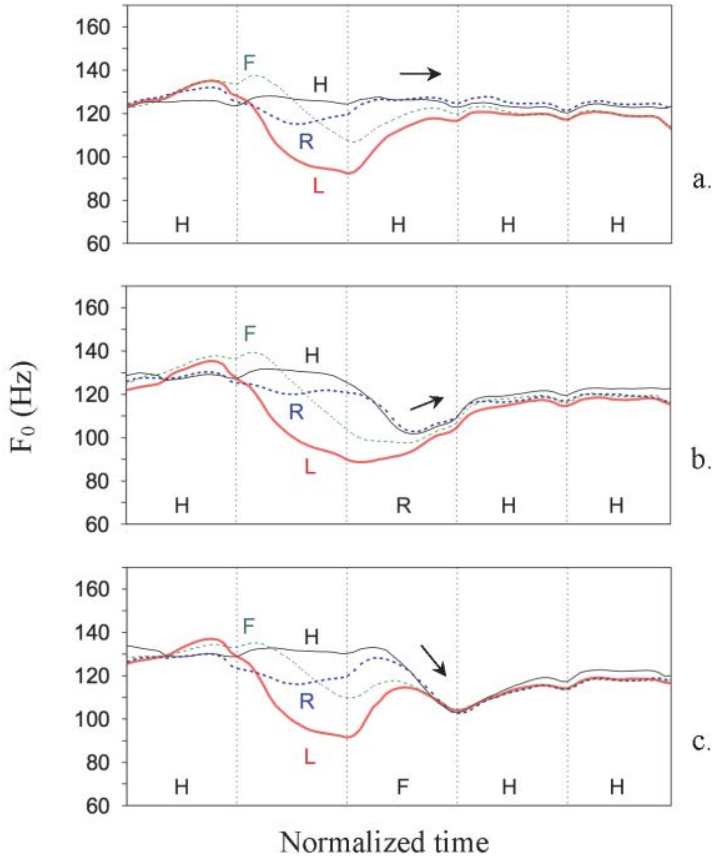


Figure 1: Mean F0 contours of Mandarin five-syllable utterances. The vertical lines mark syllable boundaries. Each plot has a different tone on syllable 3. Within each plot the four curves each has a different tone on syllable 2. H, R, L and F stand for the High, Rising, Low and Falling tones, respectively. Adapted from Xu (1999).

Because of the extensive differences in the tonal onset, each tone shows large variability due to the preceding tone. In contrast, the offset of a tone does not differ nearly as much, as can be seen in the F0 of the initial syllables in Figure 1. Also, as found in a number of tone languages (Thai: Gandour, Potisuk and Dechongkit 1994; Mandarin: Xu 1997, 1999; Yoruba: Laniran and Clements 2003; Igbo: Laniran 1997), the anticipatory influence on the preceding tone is dissimilatory rather than assimilatory.

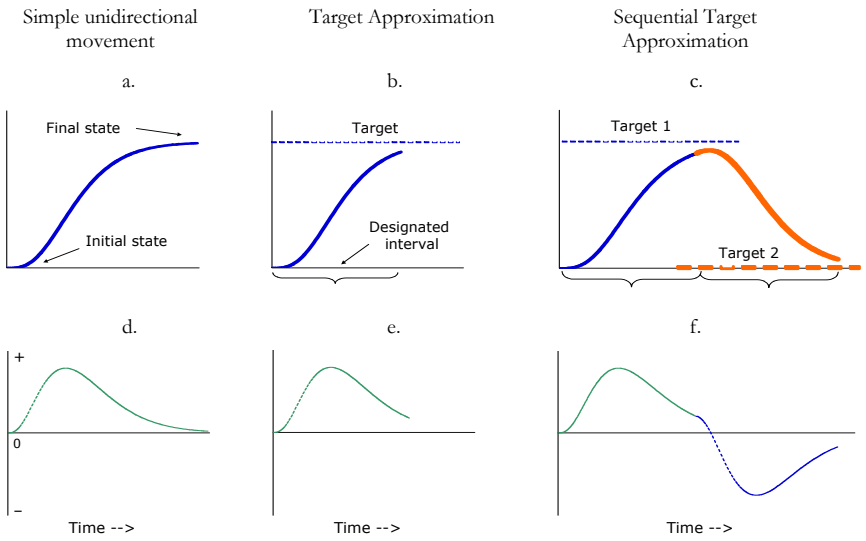


Figure 2: Illustration of simple movement (a), approximation of a single target (b), sequential approximation of two targets (c), and their corresponding velocity profiles (d-f).

The findings about tonal dynamics can be understood in terms of rather basic movement dynamics, as illustrated in Figure 2. Figure 2a shows displacement of a simple movement as a function of time, based on Nelson's (1983) definition. In such a movement, an object changes its state in a unidirectional manner. Its velocity profile shows a unimodal shape (Figure 2d), starting and ending at zero, indicating that the object is stationary at both the onset and offset of the movement. In Figure 2b, the movement is also simple but is executed with a specific goal, as indicated by the dashed line, and it is given a specific time interval which limits its duration. As a result, the target is approached but not reached by the end of the movement, although its trajectory is identical to the corresponding portion in 2a. The "early" termination of the movement results in a final velocity that is non-zero, as shown in 2e, which indicates that the movement does not come to a standstill when its designated time is over. In Figure 2c two movements are executed in succession, each approaching

a particular target within a specific time interval. The first movement is identical to the one in 2b, but its final state – consisting of relatively high position and positive velocity (Figure 2f) – is carried over to the next interval as the initial state of the second movement. As a result of such state transfer, the highest displacement (and the turning point, or peak) occurs in the second interval, despite the fact that the second target is lower than the first.

To adequately model tonal patterns seen in Figure 1, however, an additional assumption is needed. As can be seen in Figures 1b and 1c, what is approached in syllable 3 are not static registers as in Figure 1a, but dynamic trajectories with a positive or negative slope. This suggests that a target itself can be dynamic. This assumption is thus included in Xu and Wang (2001) as part of the Target Approximation (TA) model for tone and intonation, which is shown in Figure 3. Here the pitch target of the first syllable, as illustrated by the slanted dashed line, has a rising trajectory appropriate for R. An implication of the dynamic targets is that their approximation would typically generate a non-zero velocity, which would result in the turning point being delayed into the next interval by an even greater amount than in Figure 2c.

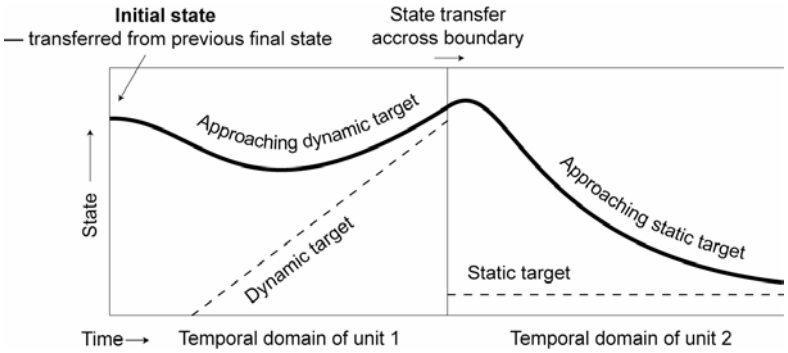


Figure 3: The Target Approximation (TA) model. The vertical lines represent boundaries between phonetic units. The dashed lines represent underlying targets of the units. The thick curve represents the surface trajectory that results from asymptotic approximation of the targets.

Given a dynamic model like TA, we can further see that the degree of each target approximation is affected by several factors: a) the distance between the initial and targeted states, b) the duration of the temporal interval assigned to the target and c) the rate at which the target is approached. Also, the amount of delay of the turning point into the upcoming interval depends not only on the final velocity of the first movement, but also the rate at which the following target is approached. A faster rate would result in an earlier reversal of the movement direction carried over from the preceding movement if a change of direction is

needed across the boundary. Evidence of all these effects has been reported in studies of various tonal phenomena (Xu 1997, 1999, 2001; Xu and Wang 2009).

Many aspects of the target approximation process have been previously discussed in both segmental and prosodic research. The notion of target has been essential to Lindblom's (1963) undershoot model, as the idea of undershoot is meaningless without fixed targets (contrary to theories assuming flexible targets, e.g., Keating 1990). Also the task dynamic model assumes that each articulatory gesture has an equilibrium point to which the articulatory state will relax by the end of the gestural cycle (Saltzman and Munhall 1989). A similar target notion is also found in the equilibrium-point hypothesis (Perrier, Ostry and Laboissière 1996). Likewise, the commands in the Fujisaki model of intonation can also be considered as targets (Fujisaki et al. 2005). Unlike TA, however, no prior models have allowed the targets themselves to be dynamic.

Some kind of target approximation mechanisms are also assumed in all these models, but an essential assumption of TA not shared by other models is the transfer of higher-order states of one movement to the next. In TA the initial state of each movement is defined in terms of not only the final displacement of the previous movement, which is assumed in all similar models, but also final velocity and final acceleration (Prom-on, Xu and Thipakorn 2009). Other models, at least in their implementation, seem to assume that velocity and acceleration at movement boundaries always reach 0.

The transferring of higher-order states across movement boundaries in TA is partly necessitated by the model's built-in allowance for target undershoot, which naturally leads to frequent occurrence of non-zero velocity and acceleration at movement offsets, as illustrated in Figure 2c and 2f. Without the state transfer, there would be abrupt changes of velocity and acceleration across movement boundaries. On the other hand, the assumption of fixed zero velocity at boundaries in some models, e.g., Fujisaki et al. (2005) and Saltzman and Munhall (1989), means treating every observed movement as if it had reached the target. There is therefore a lack of consistent representation of target undershoot in these models.

Finally, TA, like the equilibrium-point hypothesis, assumes that movements only unidirectionally approach one target or another, without returning to a baseline or a neutral position. In contrast, most other models assume that movements are bidirectional, i.e., consisting of onset and release that travel both to and from the target (Browman and Goldstein 1992; Fujisaki et al. 2005; Moon and Lindblom 1994; Saltzman and Munhall 1989; van Santen and Möbius 2000).

## 3 Segmental dynamics

Compared to tonal dynamics, segmental dynamics apparently involves greater complexity, as typically more than one articulator is involved, and the corresponding acoustic patterns are also multi-dimensional. As a result, it is much more difficult to identify the temporal intervals of segments[1]. Nevertheless, there are widely assumed segmentation conventions which are based predominantly on acoustic landmarks (Jakobson, Fant and Halle 1963; Turk, Nakai and Sugahara 2006). Although the nature of the segments divided by these landmarks may differ from segment to segment, what is common is that each segmental interval corresponds to acoustic patterns that are considered the most appropriate for the segment. For example, a typical vowel interval would consist of continuous formants bordered by abrupt spectral shifts, and a typical consonant interval would correspond to a region where vowel formants are interrupted, as illustrated in Figure 4a. Based on such segmentation, the formant transitions before and after a consonant are viewed as reflecting the coarticulatory influence of the consonant on the vowels. Likewise, if a vowel preceding an intervocalic consonant exhibits (during the transition) spectral qualities similar to those of the vowel following the consonant (Öhman 1966), it is considered as anticipatory coarticulation with the vowel. Such landmark-based segmentations run into problems when clear landmarks are absent, as in the case of glides and approximants like [j], [w] and [ɹ]. In Figure 4b, for example, because of the continuous formant movements, there is no strong basis on which the exact interval of [w] can be determined. As a result, it has been advised that these sounds should simply be avoided when studying segmental durations (Turk et al. 2006).

What seems lacking is a non-segmental landmark that can serve as a reference in the case of "difficult" sounds like glides. One possible reference is F0 events related to tonal and intonational patterns, which have been found to be consistently aligned to spectral landmarks of segments in various languages such as the onset and offset of syllables (Arvaniti, Ladd and Mennen 1998; Ladd et al. 1999; Xu 1999, 2001; Xu and Xu 2005). These findings, in turn, also suggest that given a particular kind of F0 event, for example, the F0 peak in the Mandarin tone sequence R L in Figure 5 (thick curve), one may predict where certain segmental events should occur. The onset of nasal murmur would occur about 30-50 ms before the F0 peak in a R L sequence in Mandarin (Xu 1999, 2001) as

1 Some may argue that there is no need to assume that segments have clear-cut boundaries, given the frequently reported extensive segmental overlap. Note, however, without knowing the boundaries, how can one be certain about the extent or even the existence of overlap? As the following discuss shows, the overlap assumption is actually based on conventional assumptions about segmental boundaries.

shown in Figure 5a. This knowledge should then allow us to assess the segmental intervals of glides by putting them in segmental-tonal sequences that differ from comparable sequences only in terms of the intervocalic consonant. For example, Figure 5b differs from Figure 5a only in that the intervocalic consonant is a glide with a similar place of articulation as [m]. Given that the tone sequences in Figures 5a and 5b are identical, the equivalent of the nasal murmur onset in the glide should also be about 30-50 ms before the F0 peak, as indicated by the dashed line. Interestingly, this location does not seem to correspond to any obvious spectral landmark.

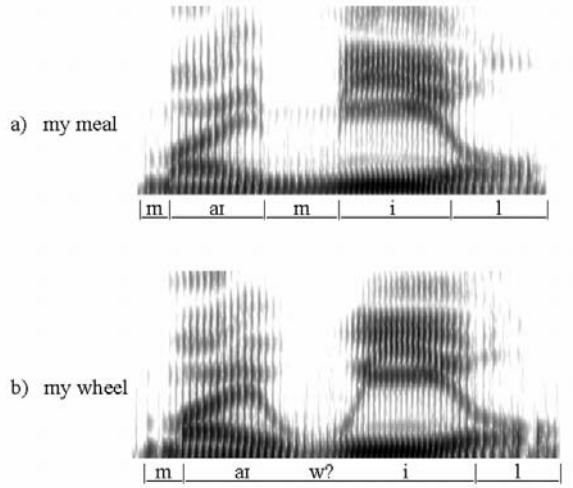


Figure 4: Landmark-based segmentations of *my meal* and *my wheel*.

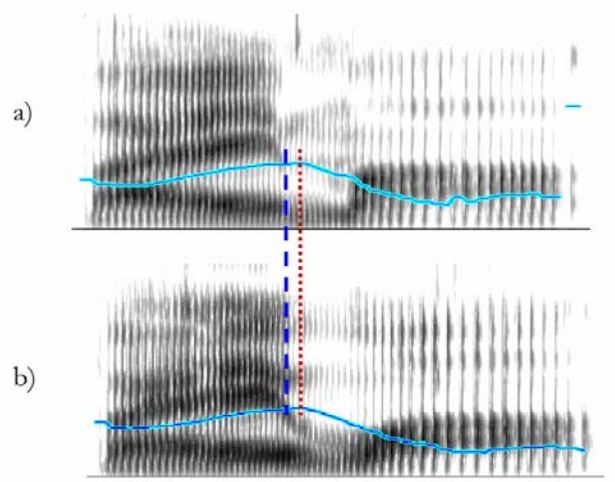


Figure 5: F0 peak as reference for determining the glide segmentation in Mandarin. a) [paɪ ma] *white horse*, tones: R L; b) [paɪ wa] *white roof-tile*, tones: R L. Data from Xu and Liu (2006).

One may notice, however, this location occurs between two turning points in F2, which divides the continuous formant trajectory into three intervals. During the first interval, F2 approaches a high frequency appropriate for [ɪ] in [aɪ]; during the second it approaches a low frequency appropriate for [w], and in the third interval it approaches a medium value appropriate for [a]. In other words, the interval during which the F0 peak occurs is one where F2 continually approaches the ideal pattern of [w]. This is reminiscent of the basic pattern of tone production seen in Figure 1: continuous approximation of an ideal target. Thus if the production of a glide is analogous to that of a tone, one may conclude that *the temporal interval of [w] is where its most appropriate pattern is being approached.* But such a conclusion is applicable not only to glides, but also to other consonants, unless we believe they are fundamentally different. In the case of [m], however, there is an issue of when exactly the ideal pattern of the consonant is best approached. In Figure 4a, the spectral pattern during the nasal murmur is largely static, thanks to the immobility of the nasal cavity, thus no continuous movement similar to that in [w] can be seen. Nonetheless, articulatory studies have discovered that the tightest occlusion occurs in the middle rather than at the beginning or end of the acoustic closure interval (Löfqvist and Gracco 1999; Westbury and Hashi 1997). This is illustrated in Figure 6, showing the spectrogram of *my meal*, as the dotted concave curve that connects the interrupted F2. Below the spectrogram are segmentations based on the traditional (upper row) or the TA model of articulatory dynamics (lower rows). According to the latter, the interval of [m] no longer coincides with the nasal murmur, but rather starts from the turning point of formant movements before the nasal murmur, and ends somewhere in the middle of the nasal murmur.

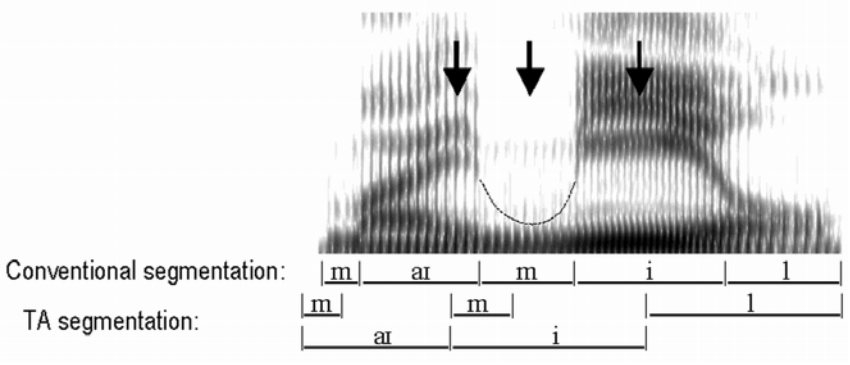


Figure 6: Conventional versus TA segmentations of *my meal*.

The application of the TA model may be further extended to vowels as well. A critical issue with the vowel is, however, where is its onset? From the perspective of target approximation, *the onset of a vowel should be where the movement toward its ideal pattern starts*. Thus the phenomenon known as anticipatory V-to-V coarticulation (Öhman 1966) can be now reinterpreted as evidence that a vowel starts much earlier than the voice onset. According to Öhman, the influence of a later vowel on an earlier one can be seen in the pre-closure formant transitions that are already in the direction of the next vowel. In fact, the classic findings that lead to the term *coarticulation* (*koartikulation* in German) (Menzerath and de Lacerda 1933, as cited by Kühnert and Nolan 1999), was described as *"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 and Nolan 1999:14). If so, the V-to-C transition in a $V_1$#$CV_2$ sequence is not only toward the underlying target of C, but also toward that of $V_2$. In other words, there is concurrent articulation (hence co-onset) of both the initial consonant and the following vowel up till the moment when the tightest consonant closure is achieved. After that, only the vowel-approaching movement continues, which then terminates when the formants change directions again.

A further illustration of TA-based segmentation can be made by the case of [l]. In its articulation, the oral cavity is not fully occluded, thus leaving the oral formants visible; and the tongue body is relatively free so that the V-to-V transition can go through the [l] murmur. As can be seen in Figure 7, before the first arrow, F2 moves continuously toward the [I] of the diphthong [eɪ]. Afterwards, it moves downward toward a very low value appropriate for the [u] in the second syllable, and this movement goes right through the intervening [l] murmur. Thus we can see clearly in the continuous trajectory of F2 that the articulatory movement approaching [u] starts well before the onset of the [l] murmur, at the time when the movement toward the [l] target starts, as has been discussed earlier, and ends well before the closure for [t].

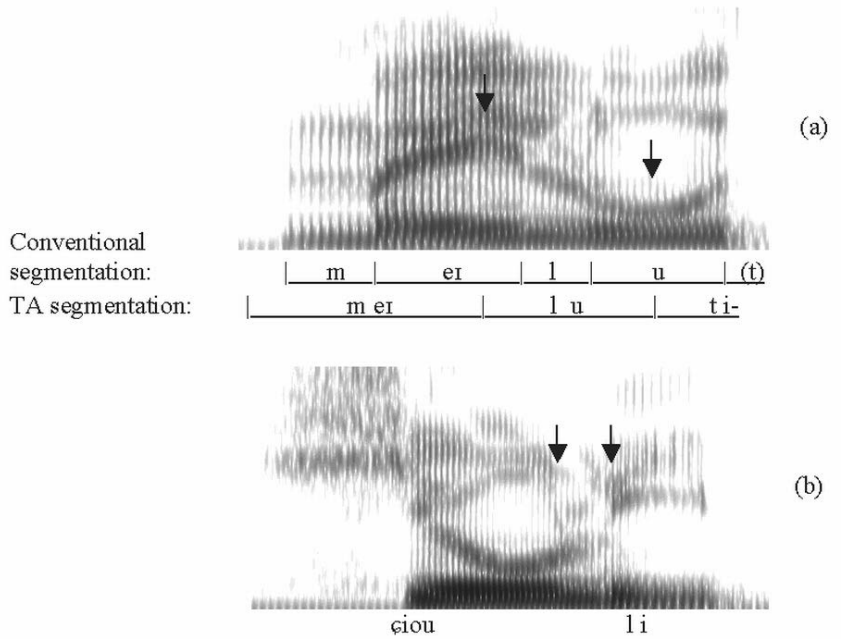


Figure 7: (a) Illustrative comparison of conventional alignment and proposed TA alignment. The entire utterance is [meɪ lu (tiɛn χuo)] (coal stove (ignition)) in Mandarin, but only the spectrogram of [meI lu] is displayed. The two arrows mark the onset and offset of the coproduced [lu], and the horizontal curly bracket marks the [l] murmur. (b) Spectrogram of [ˌiou li (pu tʂou)] (repair procedure). The two arrows mark the onset and offset of the coproduced [li], and the horizontal curly bracket marks the [l] murmur. (Partially adapted from Xu and Liu 2006.)

The TA-based segmentation discussed above has led to the *time-structure model of the syllable* (Xu and Liu, 2006), according to which the syllable serves as a *time structure* that assigns temporal intervals to consonants, vowels, tones and phonation registers, as sketched in Figure 8. The alignment of the temporal intervals is hypothesized to follow three principles:

a) *Co-onset* of the initial consonant, the first vowel, the tone and the phonation register at the beginning of the syllable;
b) *Sequential offset* of all non-initial segments, especially coda C; and
c) *Synchrony* of laryngeal units (tone and phonation register) with the entire syllable. In each case, again, the temporal interval of a segment is defined as the interval during which its target is approached.

Evidence for these principles is discussed in detail in Xu and Liu (2006).

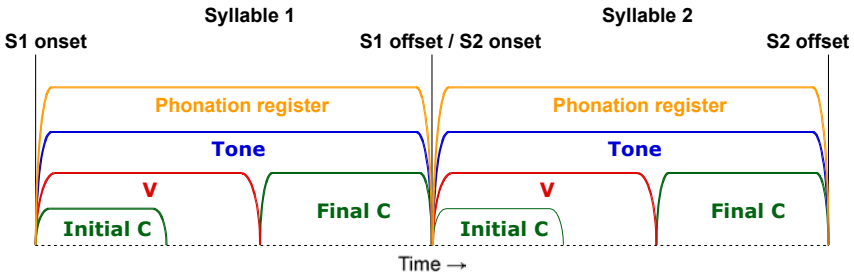


Figure 8: The time structure model of the syllable. Adapted from Xu and Liu (2006).

Several issues related to the time structure model are worth highlighting. The first is that it allows the application of the TA mechanism to segments. By defining the temporal interval of a phonetic sound as one in which its target is unidirectionally approached, and by proposing the co-onset and sequential offset principles, testable hypotheses can be formed and examined in further empirical research. The second issue is that the model conceptually eliminates a large amount of previously reported coarticulation. First, the V-to-C transition is no longer viewed as due to coarticulation, as it is part of the movement approaching the consonant and therefore is part of the consonant, as illustrated in Figure 6. Second, anticipatory V-to-V influence is no longer viewed as V-to-V coarticulation, as it is considered as part of the second V, whose target is continuously approached during this interval, as illustrated in Figure 7. Third, no carryover coarticulation of any kind needs to be assumed, because movements toward the next target are, *by definition*, within the temporal domain of that target. The only genuine coarticulation left is the co-occurrence of the consonant and vowel target approximations at the syllable onset. And even there, as far as any individual articulator is concerned, there is evidence that it can only approach one target at a time (Bell-Berti et al. 1995; Wood 1996). As for the long-distance anticipatory coarticulation reported for languages like French (Benguerel and Cowan 1974), recent evidence suggests that it is a form of vowel harmony (Fagyal, Nguyen and Boula de Mareuil 2003), which would involve a phonological process that changes the phonetic targets prior to their articulatory execution. If this is the case, the long-distance influence is due to readjustment of the vowel targets rather than overlap of their executions, and so is not genuine coarticulation.

In general, the model is a further step forward from Krakow (1999) toward a fully explicit model of the syllable. Unlike any of the previous models, the time structure model assumes that the syllable is primarily *an*

*organizing unit that assigns temporal intervals to all the syllabic components*, including not only segments, but also tones and phonation registers. Under the model, all the basic components, whether segmental or supra-segmental, are articulatory gestures that approximate specific underlying targets within their respective temporal intervals assigned by the syllable. The syllable therefore provides a structure that coherently unifies all the basic articulatory gestures. In Xu (2009), it is further hypothesized that the existence of the syllable is motivated by the need to have consistent time markers (Jones and Boltz 1989) for timing control for both production and perception of speech. The common onset of C, V, T and P at the start of each syllable would serve as such time markers. Evidence for the rigidity of co-onset can be seen in recent findings showing that a pitch target has to be categorically aligned with one syllable or another with little room for gradient alignment (Dilley and Brown 2007), and that a coda nasal is fully re-syllabified to the following vowel-onset syllable rather than remaining syllable-final or becoming ambisyllabic (Gao and Xu 2010).

An important implication of the time structure model of the syllable is that, because of the co-onset of all the syllabic components at the beginning of the syllable, each and every syllable is obligatorily assigned a local pitch target, even if the syllable is unstressed or conventionally deemed toneless, or, even in cases where identical targets occur in a row, e.g., a string of unstressed syllables, for which it might seem parsimonious to assume a single continuous target.[2] The time structure model of the syllable says nothing, however, about how the tonal targets are assigned. This is because that task belongs to the functional aspect of speech to be discussed next, in which we will again see consistency between segments and suprasegmentals.

## 4 Intonational dynamics

An important source of the conceptual divide between segments and prosody is that consonants and vowels are considered as phonemes and thus more essential to speech, while F0, duration, stress and voice quality are viewed as non-phonemic *because they are non-segmental*. The privileged status of segments may have to do with the fact that segments are often directly represented in the written language, thus seemingly essential to speech. But many ancient writing systems, such as the Sumerian, Mayan

2 We also assume that obligatory target assignment applies to segmental dynamics as well. As a result, no target underspecification (Browman and Goldstein 1989, 1992) is assumed anywhere for either segmental or intonational dynamics, even in cases where the target seems "neutral". That is, a "neutral target" is also assumed to be assigned by some specific process, and executed through target approximation.

and Chinese systems, are actually syllabaries, with no direct representation of segments (DeFrancis 1989).[3] In the case of Chinese, each monosyllabic morpheme represented by a unique character consists of not only its consonantal and vocalic components, but also its tone. Thus if segments are considered phonemic because they mark lexical contrasts, then lexical tone, lexical voice register, lexical duration and lexical stress should also be treated as phonemic.[4] Or, if we put aside the issue of phonemic status, these suprasegmental properties are at least no less important than segments in languages that use them lexically. In fact, based on the calculation of Surendran and Levow (2004), the functional load of lexical tones in Mandarin is as high as that of vowels. Given their functional importance, and assuming, as seen above, that the same basic articulatory mechanism – *syllable organized sequential target approximation* – is involved, there is no compelling reason why we should maintain an absolute divide between segmental and suprasegmental components at the lexical or syllabic level.

It could be argued that lexical tones behave more like segments because they are strictly syllabic, while non-lexical or post-lexical patterns would involve very different mechanisms. There is evidence, however, that some non-tonal languages, e.g., English, are more lexically tonal than generally believed. Although the functional load of lexical stress in English is unlikely to be as high as that of tone in Mandarin, Fry (1958) finds that listeners can easily identify minimal pairs of verb/noun that differ only in lexical stress, such as *permit*, *subject* and *object*. Fry's findings, unfortunately, are often taken as about stress *in general*, including, in particular, sentence stress (Beckman and Edwards 1994; Kochanski et al. 2005). In fact, Fry made a deliberate effort to avoid confounding lexical and sentential stress by asking listeners to judge only whether a word is a verb or a noun and not whether a syllable or word is stressed. Of the three acoustic parameters manipulated by Fry, F0 is by far the most robust cue of lexical stress, much more effective than duration and intensity. While the latter two both generated gradient word identification functions, F0 differences led to all-or-none word identification when the cross-syllable difference varied from 5-90 Hz: no matter how small the F0 difference, the syllable with higher F0 is always heard as lexically stressed, and *word identification rate does not increase with increased F0 difference*.

3 Note that this does not mean each symbol can represent only a single lexical item. In all these systems a single symbol often represents two or more homophones (DeFrancis 1989).

4 These are by no means new findings. But it is curious why they have not been more widely recognized. Just last year we were told by an editor of a highly reputable phonetics journal that lexical tones are not phonemes.

Nevertheless, it is not the case that Fry (1958) found lexically stressed syllables in English to always have higher F0 than unstressed syllables. In fact, he found rather complicated patterns when trying to identify the effects of sentence intonation on lexical-stress perception, and in many of them the perceived stressed syllable had lower F0 than the unstressed syllable. The complexity of pitch patterns in English has led to the widespread view that pitch patterns in English is fundamentally different from those of lexical tones in that,

a) they are associated with words or phrases rather than with syllables;

b) their specifications are assigned sporadically rather than syllable-by-syllable; and

c) unstressed syllables are unspecified for any pitch value.

Because of these assumed differences, fundamentally different pitch production strategies are believed to be involved in tone and non-tonal languages. In particular, unlike tones that are associated with the syllable, intonationally relevant pitch events in languages like English are proposed to be either specified as holistic contours, such as rise, fall, rise-fall or fall-rise, according to nuclear tone analysis (Crystal 1969; O'Connor and Arnold 1961; Palmer 1922), or simply in terms of peaks and valleys according to autosegmental-metrical (AM) theory of intonation (Gussenhoven 2004; Ladd 2008; Pierrehumbert 1980).

Worth particular mentioning is the argument that in English, because consistent F0 profiles can be observed across words of different lengths, there cannot be any syllable-level pitch specifications. As illustrated by Pierrehumbert (2000), words like *limb, limo, limousine* can be all said with a falling-rising contour despite the differences in the number of syllables. She explained that *"the equivalence is not captured in a syllable-by-syllable transcription of F0 levels or changes"*, because in *limb* the F0 peak occurs early in the syllable; in *limo* the peak occurs near the end of the first syllable; and in *limousine* the peak occurs beyond the end of [li] and during the nasal murmur of [m]. As a result, *"the patterns are only rendered equivalent by a representation which distinguishes the contour itself from the way that the contour is aligned with the syllables"*. This has led to the proposal that F0 peaks are only loosely associated with the stressed syllable, and that separate phonetic rules are needed to specify the exact alignment of the peaks (Ladd et al., 2009). Note that any peak-alignment-as-target account has to resolve one issue, namely, how the F0 contours between the peaks are generated. A popular assumption is that they are generated by either linear or sagging interpolation (Pierrehumbert 1981). Interpolation implies that all the points between two peaks are affected by both peaks. As found in Xu and

Xu (2005) and Chen and Xu (2006), however, the F0 of the weak syllables, i.e., unstressed syllables in English and neutral-tone syllables in Mandarin, are not affected by the F0 of the upcoming strong syllables. Two English examples are shown in Figure 9, where the two pairs of mean F0 contours each differ only in whether the final word is focused. In each pair the F0 contours with and without final-focus do not start to deviate from each other until after the onset of the focused word. Thus there is no evidence of interpolation between the final peak and either the major peak in *Lee* or the smaller peak in *know*. This suggests that a) the articulation of the focus-related pitch increase is local to the stressed syllable under focus, b) the F0 contour of the weak syllable *my* must have its own target, and c) the onset F0 of the focused, hence "strong", syllable is actually determined by the final F0 of the preceding weak syllable *my* (Xu and Xu 2005).

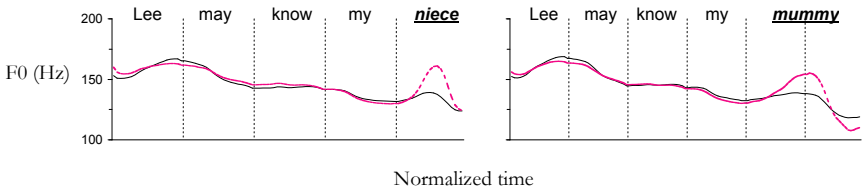


Figure 9: Mean F0 contours of two minimal pairs of English sentences. In each pair the only difference is in terms of final focus. Averaged across 49 tokens by 7 speakers. Adapted from Xu and Xu (2005).

But how can we determine the value of the local pitch target of a syllable? As seen in Figure 1, in Mandarin the most consistent F0 properties across tonal contexts and speech rates are the slope and height of the F0 trajectory near the end of the tone-carrying syllable (Xu 1997, 1998, 1999). A similar finding has been made for Cantonese (Wong 2006). Based on these findings, we have developed two measurements for assessing the underlying pitch target of the syllable – syllable-final velocity and syllable-final pitch (*final velocity* and *final pitch* for short), both of which are measured at 30 ms before the conventional syllable offset (taking into consideration the finding that all syllable boundaries should be shifted leftward as discussed earlier). These measurements have been used in a number of studies that explored the underlying targets of tone and intonation in Mandarin and English (Chen and Xu 2006; Liu and Xu 2007a, 2007b). For English, in particular, we have found that it is possible to identify syllable-bound pitch targets that are determined jointly by lexical stress, focus and sentence type (statement vs. yes/no question). The general findings are summarized as follows, which are also illustrated in Figure 10.

1) Every syllable, whether stressed or unstressed, has an underlying pitch target.

2) The pitch target of an unstressed syllable is likely [mid].

3) The pitch target of a stressed syllable is [high] in a statement (solid line, *Mi-* in Figure 10a), but [fall] if it is word final AND the word is either focused (dash-dot line, *job* in Figure 10a) or sentence final (solid and dash-dot lines, *-ssage* in Figure 10b).

4) The pitch target of a stressed syllable is [rise] in a yes/no question (dash-dot lines, *job* and *-ssage* in Figure 10a and 10b, respectively).

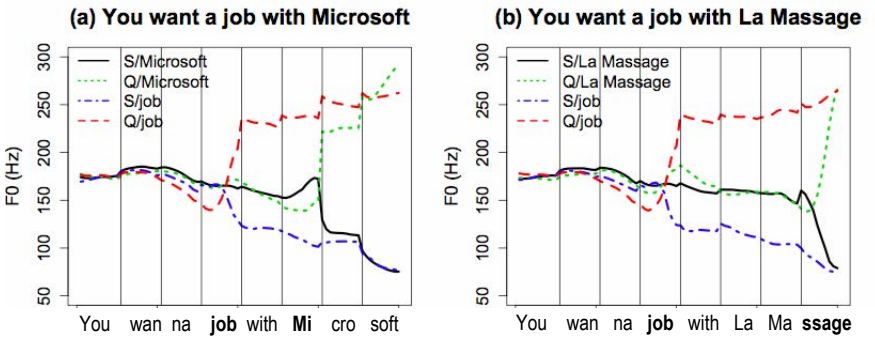


Figure 10: Time-normalized mean F0 contours of two English sentences produced as statement or yes/no question and with medial and final foci, averaged across 40 repetitions by 5 speakers (Liu and Xu, 2007a).

Figure 10 also shows that the pitch targets are also further affected by extensive changes in pitch range due to both focus and sentence type. Discussion of such pitch range modifications, however, is beyond the scope of this article. Interested readers can refer to Liu and Xu (2007a) and Liu (2009) for details.

One may question, however, how rules as complicated as these can be ever mastered by speakers of English. Judging from the highly complex tone sandhi rules in various tone languages (Chen 2000), many of which much more complicated than those just described, the complicity of tonal assignment in English should be no obstacle in either acquisition or daily operation of intonation.

The findings about the interactive patterns of F0 contours in English also offer explanations for the perceptual patterns reported by Fry (1958). Fry's pattern E, i.e., / ¯¯, for example, was recognized 70% of the times as nouns, and this is consistent with the focused *Microsoft* in a question (dotted line in Figure 10a), which is otherwise incomprehensible because

the syllable having the higher pitch is heard as unstressed. His pattern D: _ /, and P: _ _/, were recognized 61% and 66% of the times as verbs, respectively. This is consistent with the finding that in the focused word *Massage* in a question (dotted line in Figure 10b), the stressed syllable has a rising pitch target.

Finally, recent findings that different languages or even different dialects of the same language may have gradient differences in F0 peak alignment (Atterer and Ladd 2004; Ladd et al. 2009) could be due to syllable-bound underlying pitch targets that differ gradiently in their slope and height in these languages. This possibility, of course, needs to be experimentally tested, and final velocity and final pitch can be used as the measurements that can assess the properties of the possible targets.

Two caveats need to be mentioned, however, in regard to pitch targets in English. The first is that because yes-no questions in English are reported to often have falling rather than rising intonation (Hedberg, Sosa and Fadden 2004), the rising pitch targets found in Liu and Xu (2007a) are unlikely to be the direct correlates of syntactic questions. Nevertheless, the findings show that rising pitch targets are the default patterns when native speakers read aloud syntactic questions. Secondly, the [mid] targets reported in Chen and Xu (2006) and Xu and Xu (2005) are based on data rather than on assumptions, i.e., final F0 in unstressed syllable were actually measured to be midway between the highest and lowest F0 in stressed syllables. But this finding by itself says nothing about the source of the [mid] target. While it is possible that [mid] is associated with the rest position of speakers' pitch production, whether this is true has to be determined by studies specifically designed to answer the question.

## 5 Dynamics of reduction

The last commonality between segmental and suprasegmental patterns we would like to explore is reduction. Reductional phenomena have been widely reported for both segmental and tonal aspects of speech (Fourakis 1991; Kohler 1990; Lindblom 1963; van Son and Pols 1999), but the underlying mechanisms of both remain unresolved. Lindblom (1963) proposed, based on Swedish data, that vowel formant variations related to consonantal context are the result of target undershoot due to time pressure, i.e., lack of time for articulators to move from one position to another. This target undershoot account, in effect, models contextual segmental variation as a form of reduction. Lindblom's proposal, however, was questioned by a number of subsequent studies, including, in particular, Gay (1978) and Harris (1978), who reported negligible context-related vowel undershoot. In response to these criticisms, Moon and

Lindblom (1994) pointed out that target undershoot is sensitive to *locus-target distances* – the distance an articulator needs to travel between the consonant and the following vowel, and that studies that found minimal undershoot used consonantal contexts that involved only moderate locus-target distances. They demonstrated that, when locus-target distance is actually large, e.g., in the case of vowels embedded in a /w__l/ frame, reduction rate clearly varied with both consonantal context and vowel duration.

Nonetheless, Moon and Lindblom (1994) also reported that target undershoot could be reduced by adopting a clear speech style. Based on this finding, Lindblom (1990) expanded his target undershoot model to the H&H theory, according to which speakers always try to maintain a balance in a trade-off relation between sufficient contrast and articulatory effort. This modification makes H&H similar to other theories of economy of effort, in particular, Nelson (1983). In fact, economy of effort has been widely adopted as a major principle underlying phonetic reduction, although scepticisms also exist (Ladefoged 1990; Ohala 1990).

We note, however, the principle of economy of effort entails two critical assumptions: a) it is always possible to further increase velocity to avoid undershoot, and b) there is always room for articulation to go further in the direction of the target. Both assumptions would run into problems with existing findings. The first assumption is inconsistent with the finding of *minimum duration*, which, according to Klatt (1976), is *"an absolute minimum duration $D_{min}$ that is required to execute a satisfactory articulatory gesture"*. That is, unless segments are never produced shorter than their minimum duration, there will be cases where undershoot is inevitable regardless of the articulatory effort. That speakers often approach their maximum speed of articulation has in fact been suggested either directly (Sigurd 1973; Tiffany 1980) or indirectly (Adank and Janse 2009; Janse 2004). In particular, Xu and Sun (2002) show evidence that in the production of both tone and intonation, speakers often approach their maximum speed of pitch change. Furthermore, there is evidence that the degree of target undershoot is directly related to duration shortening in both tone production in Beijing Mandarin (Xu and Wang 2009) and in segment production in Taiwan Mandarin (Cheng and Xu 2009).

The problem with the second assumption entailed by the principle of economy of effort is that, in cases where the duration assigned to a target is longer than is needed for its full articulation, if the articulation does not slow down, there will be overshoot of the targets. Evidence for articulatory slowdown can be seen in several studies. Xu and Sun (2002) demonstrated that in Mandarin, speakers approach their maximum speed of pitch change only in the dynamic tones, such as R and F, where two movements are often needed within a single syllable, whereas in static

tones like H and L the speed of pitch change is much slower than the maximum speed, because only one pitch movement needs to be made within a syllable. Faster movement would have resulted in overshooting the tonal targets. Cheng and Xu (forthcoming) has found that articulatory strength, as measured by the slope of regression of peak velocity of formants as a function of movement magnitude, is actually weaker in normally produced syllables than in syllables that are severely reduced at a fast speaking rate.

In general, therefore, there is accumulating evidence against the principle of economy of effort. The alternative, we would like to suggest, is the near-ceiling performance hypothesis (Xu 2008), according to which speech is maintained near an overall performance ceiling due to its vital importance for the survival and wellbeing of human individuals. In regard to articulation, maintaining a near-ceiling performance means to optimize the efficiency of information transmission to the point that the greatest physiologically allowable speed of articulation is frequently approached even at normal speech rate. But this does not mean that the maximum speed is reached all the time, because in many cases it is unnecessary or even undesirable to use the maximum speed. Also, near-ceiling performance does not mean that there is no room for speech rate to be further increased in terms of *number of syllables or segments per second*. Not only can movements not yet at top speed be accelerated, but also those already at the speed limit can be further shortened, and the result is just more reduction, as shown by Adank and Janse (2009) and Janse (2004). It is crucial, therefore, to always clearly distinguish between speech rate in term of number of phonetic units produced and in terms of articulatory distance covered in a given amount of time.

It is important to note further that duration itself also carries information, in fact, many layers of information. These include lexical contrast related to lexical stress or quantitative vowel length, focus, and grouping (de Jong and Zawaydeh 1999; de Jong et al. 2004; Lehiste et al. 1976; Xu 2009). The grouping function alone involves multiple durational manipulations, including lengthening the initial and final syllables of a group (Beckman and Edwards 1990; Cooper, Lapointe and Paccia 1977), and shortening the group-medial syllables (Klatt 1976; Lehiste 1972). Xu (2009) proposed that these timing patterns are all based on a coding strategy (referred to as affinity index) of using the temporal distance between adjacent syllables, measured in terms of onset-to-onset interval, to iconically indicate their relational distance. The application of such a coding strategy thus entails that it is the group-medial syllables that are most likely to go through severe reductions, and this has been shown to be true of both tone (Shih 1993; Xu 1994) and segment (Cheng and Xu 2009).

In general, therefore, segmental and tonal reductions both seem to be closely related to time pressure, and there are no fundamental differences between the two in terms of their basic mechanisms other than the specific articulators involved.

## 6 Conclusion

In this chapter we have argued that a fundamental coherence exists between segmental and suprasegmental aspects of speech in terms of both articulatory dynamics and communicative functions. Articulatorily, both segmental and suprasegmental components are produced with sequential target approximation, and the assignment of the target approximation interval is organized by the syllable which guarantees co-onset of all syllabic components except coda consonants. Functionally, suprasegmental components are often used in encoding lexical contrast just like segments, and in English, at least, specific pitch targets are assigned to each and every syllable and the assignment is done jointly by the lexical, focal and sentential functions. Finally we have shown that reduction of both segmental and suprasegmental components is likely due to time pressure despite maximum articulatory effort, rather than due to economy of effort. Our discussion here is, of course, only the first step toward a unified theory of segmental and prosodic aspects of speech. Much more research is needed to further verify our proposal.

## 7 References

Adank, P. and E. Janse (2009): Perceptual learning of time-compressed and natural fast speech. *The Journal of the Acoustical Society of America* ***126***, 2649-2659.

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

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

Beckman, M. E. and J. Edwards (1990): Lengthenings and shortenings and the nature of prosodic constituency. In: J. Kingston, M. E. Beckman (eds): *Papers in Laboratory Phonology 1 – Between the Grammar and Physics of Speech* (pp. 152-178). Cambridge: Cambridge University Press.

Beckman, M. E. and J. R. Edwards (1994): Articulatory evidence for differentiating stress categories. In: P. A. Keating (ed.): *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III* (pp. 7-33). Cambridge: Cambridge University Press.

Bell-Berti, F.R. A. Krakow,C. E. Gelfer and S. Boyce (1995): Anticipatory and carryover effects: Implications for models of speech production. In: F. Bell-Berti, L. J. Raphael (eds): *Producing Speech: Contemporary Issues. For Katherine Safford Harris* (pp. 77-97). New York: AIP Press

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

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

Browman, C. P. and L. Goldstein (1992): Articulatory phonology: An overview. *Phonetica* ***49***, 155-180.

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

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

Chen, Y. and Y. Xu (2006): Production of weak elements in speech -- Evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica* ***63***, 47-75.

Cheng, C. and Y. Xu (2009): Extreme reductions: Contraction of disyllables into monosyllables in Taiwan Mandarin. P*roceedings of Interspeech 2009, Brighton, UK***,** 456-459.

Cheng, C. and Y. Xu (forthcoming): *Mechanisms of extreme reductions: Evidence from syllable contraction in Taiwan Mandarin*. Manuscript.

Cooper, W.,S. Lapointe and J. Paccia (1977): Syntactic blocking of phonological rules in speech production. *Journal of the Acoustical Society of America* ***61***, 1314-1320.

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

DeFrancis, J.F. (1989): *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu: University of Hawaii Press.

de Jong, K. (2004): Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics* ***32***, 493-516.

de Jong, K.J. and B.A. Zawaydeh (1999): Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics* ***27***, 3-22.

Dilley, L.C. and M. Brown (2007): Effects of pitch range variation on f0 extrema in an imitation task. *Journal of Phonetics* ***35***, 523-551.

Fagyal, Z.N. Nguyen and P. Boula de Mareuil (2003): From dilation to coarticulation: is there vowel harmony in French? *Studies in Linguistic Sciences 32*, 1-21.

Fourakis, M. (1991): Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America* ***90***, 1816-1827.

Fry, D.B. (1958): Experiments in the perception of stress. *Language and Speech* ***1***, 126-152.

Fujisaki, H., C. Wang, S. Ohno and W. Gu (2005): Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model. *Speech communication* ***47***, 59-70.

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

Gao, H. and Y. Xu (2010): Ambisyllabicity in English: How real is it? *Proceedings of The 9th Phonetics Conference of China (PCC2010), Tianjin.*

Gay, T.J. (1978): Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America* ***63***, 223-230.

Gussenhoven, C. (2004): *The Phonology of Tone and Intonation*: Cambridge University Press.

Harris, K.S. (1978): Vowel duration change and its underlying physiological mechanisms. *Language and Speech* ***21***, 354-361.

Hedberg, N., J.M. Sosa and L. Fadden (2004): Meanings and configurations of questions in English. *Proceedings of the 2nd International Conference on Speech Prosody 2004, Nara, Japan*, 309-312.

Jakobson, R., G. Fant and M. Halle (1963): *Preliminaries to Speech Analysis*. Cambridge: MA: MIT Press (Originally published in 1951).

Janse, E. (2004): Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication* ***42***, 155-173.

Jones, M. R. and M. Boltz (1989): Dynamic attending and responses to time. *Psychological Review* ***96***, 459-491.

Keating, P. A. (1990): The window model of coarticulation: articulatory evidence. In: J. Kingston, M.E. Beckman (eds): *Papers in Laboratory Phonology 1 – Between the Grammar and Physics of Speech* (pp. 451-470). Cambridge: Cambridge University Press.

Klatt, D.H. (1976): Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* ***59***, 1208-1221.

Kochanski, G., E. Grabe, J. Coleman and B. Rosner (2005): Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America* ***118***, 1038-1054.

Kohler, K.J. (1990): Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In: W.J. Hardcastle, A. Marchal (eds): *Speech production and speech modelling* (pp. 69-92). Dordrecht: Kluwer.

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

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

Ladd, D.R. (2008): *Intonational phonology*. Cambridge: Cambridge University Press.

Ladd, D.R., D. Faulkner, H. Faulkner and A. Schepman (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., A. Schepman, L. White, L.M. Quarmby and R. Stackhouse (2009): Structural and dialectal effects on pitch peak alignment in two varieties of British English. *Journal of Phonetics* ***37***, 145-161.

Ladefoged, P. (1990): Some reflections on the IPA. *Journal of Phonetics* ***18***, 335-346.

Laniran, Y. and C. Gerfen (1997): High raising, downstep and downdrift in Igbo. *Proceedings of The 71st Annual Meeting of the Linguistic Society of America, Chicago, USA,* 59.

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

Lehiste, I. (1972): The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America* ***51***, 2018-2024.

Lehiste, I. (1996): Suprasegmental features of speech. In: N. J. Lass (ed.): *Principles of Experimental Phonetics* (pp. 226-244). Boston: Mosby.

Lehiste, I., J.P. Olive and L.A. Streeter (1976): Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America* ***60***, 1199-1202.

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

Lindblom, B. (1990): Explaining phonetic variation: A sketch of the H&H theory. In: W. J. Hardcastle, A. Marchal (eds): *Speech Production and Speech Modeling* (pp. 413-415). Dordrecht: Kluwer.

Liu, F. (2009): *Intonation systems of Mandarin and English: A functional approach*. Ph.D. dissertation, University of Chicago, Chicago, IL

Liu, F. and Y. Xu (2007a): Question intonation as affected by word stress and focus in English. *Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrücken, Germany,* 1189-1192.

Liu, F. and Y. Xu (2007b): The Neutral Tone in Question Intonation in Mandarin. *Proceedings of Interspeech 2007, Antwerp, Belgium,* 630-633.

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

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

Moon, S.-J. and B. Lindblom (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.

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

Ohala, J.J. (1990): The phonetics and phonology of aspects of assimilation. In: J. Kingston, M.E. Beckman (eds): *Papers in Laboratory Phonology 1 – Between the Grammar and Physics of Speech* (pp. 258-275). Cambridge: Cambridge University Press.

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

Palmer, H.E. (1922): *English Intonation, with Systematic Exercises*. Cambridge: Heffer.

Perrier, P., D.J. Ostry and R. Laboissière (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].

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

Pierrehumbert, J. (2000): Tonal elements and their alignment. In: M. Horne (ed.): *Prosody: Theory and Experiment – Studies Presented to Gösta Bruce* (pp. 11-36). London: Kluwer Academic Publishers.

Prom-on, S., Y. Xu and B. Thipakorn (2009): Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* ***125***, 405-424.

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

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

Siguard, B. (1973): Maximum rate and minimum duration of repeated syllables. *Language and Speech* ***16***, 373-395.

Surendran, D. and G.-A. Levow (2004): The functional load of tone in Mandarin is as high as that of vowels. *Proceedings of the 2nd International Conference of Speech Prosody 2004, Nara, Japan***,** 99-102.

Tiffany, W. R. (1980): The effects of syllable structure on diadochokinetic and reading rates. *Journal of Speech and Hearing Research* ***23***, 894-908.

Turk, A., S. Nakai and M. Sugahara (2006): Acoustic Segment Durations in Prosodic Research: A Practical Guide. In: S. Sudhoff, D. Lenertová, R. Meyer et al. (eds): *Methods in Empirical Prosody Research* (pp. 1-28). Berlin/New York: De Gruyter.

van Santen, J. and B. Möbius (2000): A quantitative model of F0 generation and alignment. In: A. Botinis (ed.): *Intonation - Analysis, Modeling and Technology* (pp. 269-288). Kluwer, Dordrecht.

van Son, R.J.J.H. and L.C.W. Pols (1999): An acoustic description of consonant reduction. *Speech Communication* ***28***, 125-140.

Westbury, J. and M. Hashi (1997): Lip-pellet positions during vowels and labial consonants. *Journal of Phonetics* ***25***, 405–419.

Wong, Y.W. (2006): Realization of Cantonese Rising Tones under Different Speaking Rates. *Proceedings of the 3rd International Conference of Speech Prosody, Dresden, Germany***,** 198-201.

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. (2008): Multi-dimensional information coding in speech. *Proceedings of the 4th International Conference of Speech Prosody, Campinas, Brazil***,** 17-26.

Xu, Y. (2009): Timing and coordination in tone and intonation--An articulatory-functional perspective. *Lingua* ***119***, 906-927.

Xu, Y. and F. Liu (2006): Tonal alignment, syllable structure and coarticulation: Toward an integrated model. Italian *Journal of Linguistics* ***18***, 125-159.

Xu, Y. and F. Liu (2007): Determining the temporal interval of segments with the help of F0 contours. *Journal of Phonetics* ***35***, 398-420.

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

Xu, Y. and M. Wang (2009): Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics* ***37***, 502-520.

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

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

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