Available online at www.sciencedirect.com
ScienceDirect
Lingua 119 (2009) 906–927
Lingua
www.elsevier.com/locate/lingua

Timing and coordination in tone and intonation—An articulatory-functional perspective

Yi Xu *

University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK

Received 15 February 2007; received in revised form 17 July 2007; accepted 13 September 2007
Available online 13 May 2008

Abstract

Timing is of critical importance for speech in general and for tone and intonation in particular. Yet our understanding of timing is still limited. In this paper I explore timing-related issues from an articulatory-functional perspective, which views speech as communicative functions encoded through an articulation process. Based on this view, timing in speech can be seen as of two kinds, obligatory timing—timing as obligated by articulation, and informational timing—timing that encodes communicative meanings. I will show that the articulatory constraints of minimal movement duration and syllable-bound temporal alignment severely limit the freedom of using timing for information coding, leaving duration as the only controllable aspect of timing. I will further show that duration, as other aspects of speech, is used to encode multiple layers of communicative meanings in parallel, allowing concurrent encoding of lexical contrast, focus, and inter-constituent affinity. Such information coding seems to account for previously reported duration patterns such as polysyllabic shortening and constituent-edge lengthening. Furthermore, to the extent the reported weak isochrony tendency can be explained by information coding and articulatory mechanisms that have little to do with isochrony, speech rhythm seems to be an epiphenomenon rather than a basic mechanism.

Keywords: Target approximation; Synchronization; Time structure model; Time marker hypothesis; Affinity index hypothesis

1. Introduction

Speech takes place in time. Timing is therefore an inescapable aspect of speech. This is true of both speech in general and tone and intonation in particular. Yet our understanding of speech timing is still quite limited. This is despite the fact that there has been much research on

* Tel.: +44 20 7679 5011.
E-mail address: yi.xu@ucl.ac.uk.

0024-3841/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.lingua.2007.09.015

timing-related issues, including, in particular, rhythm, prosodic hierarchy, alignment of intonational events and durational patterns. In this paper, I explore the idea that timing in speech could be more adequately understood from an articulatory-functional perspective (Xu, 2005). This perspective is motivated by the recognition that articulation and information conveyance are two indisputably essential aspects of speech and thus should be part of any serious account of speech. In particular, given the need to convey information and the fact that such information conveyance can be done only through articulation, there has to be a continuous link between communicative meanings and surface acoustic patterns through an articulatory process.

The need to establish such a continuous link has motivated the Parallel Encoding and Target Approximation model of speech coding (PENTA) (Xu, 2005), which was originally proposed for tone and intonation and has since been extended to other aspects of speech (Xu, 2007a). A brief sketch of PENTA is provided next, but more information will be presented as the discussion proceeds. A block diagram of PENTA in its latest form is shown in Fig. 1. The boxes on the far left represent the specific communicative functions to be coded. They are directly linked to respective encoding schemes as represented by the boxes to their right, which specify the values of the articulatory parameters listed in the open box. These parameters control the articulatory process of sequential Target Approximation, consisting of consecutive unidirectional movements toward successive targets, which eventually produce the surface acoustics of speech.

The key hypothesis of PENTA is that speech coding is done by transmitting multiple communicative functions in parallel through encoding schemes that specify the values of the parameters that control the articulatory process of target approximation. This hypothesis differs from conventional views in two critical respects. First, communicative functions, as the driving force of the system, are assumed to be in direct control of the articulatory (but not acoustic) parameters through the encoding schemes. This means that all formal patterns that are under a speaker's control are generated for the sake of signaling specific communicative meanings rather than for the sake of satisfying formal requirements as conventionally assumed (Chomsky and Halle, 1968; Ladd, 1996; Pierrehumbert, 1980, 1990). Second, only articulatory parameters can be directly specified by the encoding schemes, while the generation of surface acoustic forms is always derivational and has to obey the constraints of the articulatory process. This differs from the widespread but often implicit assumption that surface acoustic forms can be directly specified either by communicative functions or by formal rules (Bolinger, 1989; Gussenhoven, 2002; Ladd, 1996, 2004; Pierrehumbert, 1990).

PENTA thus provides a mechanistic embodiment of an articulatory-functional view of speech, although it may not be the only possible embodiment. What is critical in this view is the

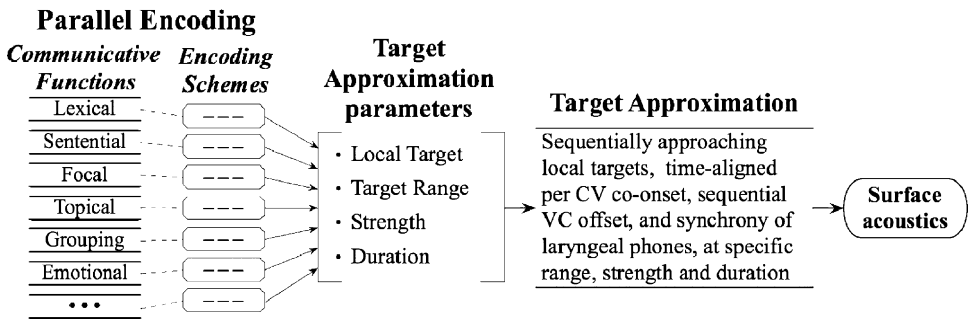

Fig. 1. A sketch of PENTA.

explication of what is articulatory and what is functional. That is, for any given observed phenomenon or proposed mechanism, it is necessary to ask, is it due to articulatory constraints or information coding? In the case of speech timing, it is thus necessary to distinguish between those aspects of timing that are obligated by articulatory mechanisms, which I will refer to as *obligatory timing*, and those that are part of the encoding schemes of communicative functions, which I will refer to as *informational timing*.

2. Obligatory timing

The human articulatory apparatus is a physical system, as such it obeys physical laws. Speech is produced by manipulating the state of this system: changing the location of the articulators, reshaping them, or adjusting their physical property such as tension and stiffness. These state manipulations are dynamical processes which necessarily take time. Part of the timing patterns of these processes is directly determined by the properties of the articulatory system and its neural control mechanisms, and is hence obligatory. There are at least two kinds of obligatory timing: minimum duration of articulatory movements, and syllable-bound temporal alignment.

2.1. Minimum movement duration

An articulatory movement is defined here as a unidirectional movement toward a targeted articulatory state. The minimum duration of a movement is dependent on a number of factors, including, most importantly, maximum net muscle force exerted in the direction of the movement, magnitude of the movement, and precision of the movement goal (Nelson, 1983; Tanaka et al., 2006). The minimum duration is negatively related to maximum muscle force, but positively related to both movement magnitude and precision of movement goal, according to Tanaka et al. (2006), who have developed a quantitative model for computing the minimum duration for eye saccades and target-reaching arm movements. The minimum time can be also empirically assessed, as has been done in Xu and Sun (2002) for pitch movements. In that study, native speakers of American English and Mandarin Chinese were asked to imitate fast alternating high–low steady-state synthetic pitch sequences as closely as possible. Because the synthetic sequences were very fast, subjects had to sacrifice accuracy for speed. As a result, a highly linear relation was found between the excursion size and peak velocity of individual pitch movements.

The importance of minimum movement duration for speech depends on how much impact it has on the surface trajectories of the acoustic variation. In regard to F_0 movement, for example, the minimum duration could be so short that it is irrelevant to most of the surface trajectories. This has been the understanding of Ladd (1996:148) when he stated that: "More 'happens' in pitch contours in Chinese, because the lexical tones occur at nearly every syllable and the transitions between them span only *milliseconds*..." (my emphasis). Fig. 2a shows average F_0 contours of the four lexical tones of Mandarin Chinese produced in isolation by 8 native speakers (Xu, 1997). Because they are produced free of contextual influences, they could be assumed as being close to the canonical forms of the tones. In Fig. 2b, the F_0 contour of the H tone is artificially concatenated to four different tones, with an imaginary 20 ms of transitions connecting the offset of the preceding tones to the onset of the H tone, so as to simulate what is envisaged by Ladd (1996). In reality, however, the transitions are much longer, and it happens mostly during the H tone itself, as seen in Fig. 2c. It could be argued that these seemingly long transitions are deliberate. But they are apparently inevitable given the minimum duration of pitch

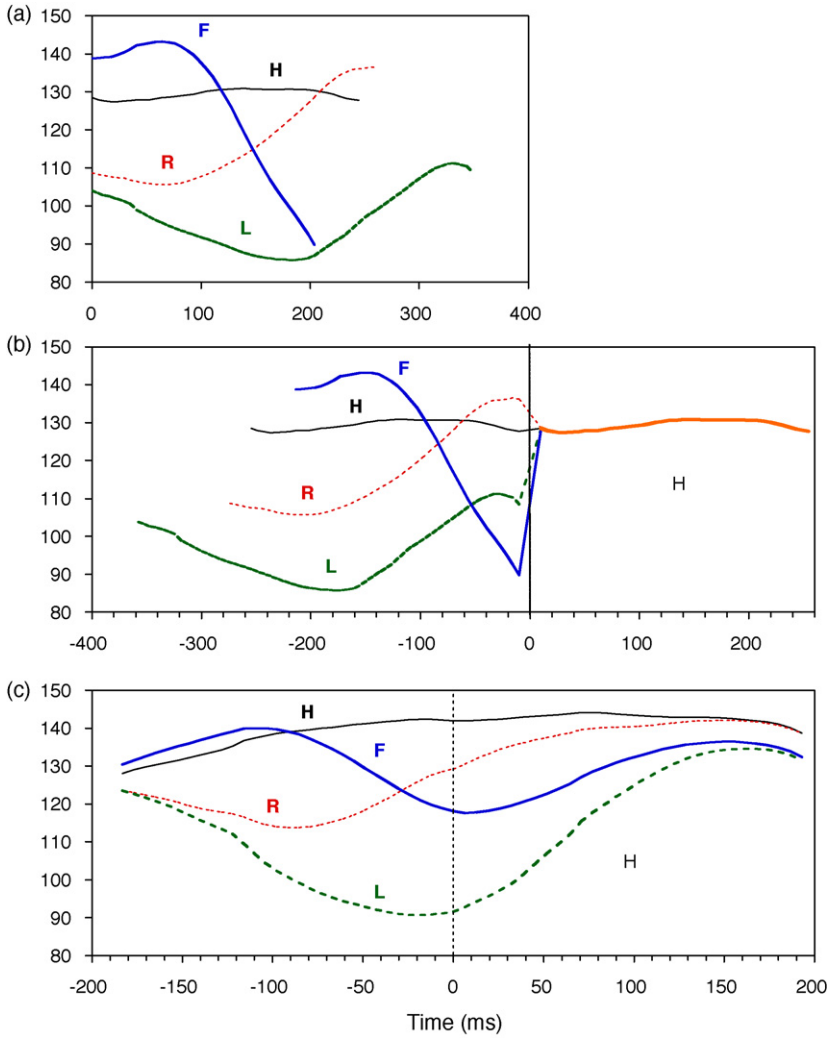

Fig. 2. (a) Mean F_0 contours of Mandarin tones produced in isolation. (b) Isolated Mandarin H tone concatenated with different preceding tones, linked by 20 ms of arbitrary linear transition. (c) Mean F_0 contours of the same two tone sequences in actual production. Adapted from Xu (1997).

movement. According to Xu and Sun (2002), the mean minimum duration of a pitch rise or fall can be estimated with the following equations:

$$t = 89.6 + 8.7d \quad \text{(rising)} \tag{1}$$

$$t = 100.4 + 5.8d \quad \text{(falling)} \tag{2}$$

where t is time in ms, and d is the size of a unidirectional pitch movement delimited by turning points.

In Fig. 2c, the amount of pitch increase from the end of the L tone to the highest point of the H tone is about 6 semitones. From Eq. (1) it takes at least 142 ms for an average speaker to complete

such a movement. Yet the mean duration of the H tone after the L tone is only 200 ms in that study, which means that the transition would take up most of the syllable duration. Thus, much of the transition in Fig. 2c is obligatory, because speakers cannot make the pitch change much faster than that.

Note that the constraint of minimum duration is quite different from what is known as the *economy of effort* hypothesis (Lindblom, 1990), according to which speakers often avoid employing full muscle force in order to conserve energy. But as noted by Tanaka et al. (2006:3883), for relatively easy and brief movements like reaching for a target, ''factors such as energy consumption and muscle fatigue are not important.'' If we assume that the F_0 contour of the H tone after a preceding H in Fig. 2c resembles the underlying target based on the conventional description of the tone, then the realization of the H tone after the L tone would be a case of undershoot. But such undershoot is unlikely due to the need to conserve energy, but rather because the articulators cannot move much faster even if driven by full muscular force.

2.2. Implications of minimum movement duration

Because of its mandatory nature, and because it is so long relative to the duration of many units in speech, the minimum movement duration has severe consequences on the surface patterns of speech. Here I will discuss only a few major ones. The first is that extensive carryover influence of one movement on the next is inevitable, in terms of both magnitude and temporal span, as is apparent in Fig. 2c. More examples can be found in Xu (1997, 1999) for Mandarin, Gandour et al. (1994) for Thai, and Wong (2006) for Cantonese.

The second implication is that dynamic tones such as Rising (R) and Falling (F), since they require two movements within a syllable when preceded by a tone with conflicting F_0 offset, are impossible to articulate if the syllable duration is too short. Several recent studies have reported typological findings that the ability of a syllable to carry contour tones is directly related to the duration of its tone-bearing portion (Duanmu, 1994; Gordon, 2001; Zhang, 2002). As discussed in detail in Xu (2004), based on duration data reported by Duanmu (1994), Gordon (1999) and Xu (1997, 1999), in languages like Shanghai where mean syllable duration is close to 150 ms, it is virtually impossible to execute two consecutive F_0 movements within a syllable as required by a dynamic tone. In fact, even in languages like Mandarin, where mean syllable duration is around 180–200 ms (Xu, 1997, 1999), in many cases, it is only barely possible to realize a dynamic tone at normal speaking rate. At fast speaking rate, syllable duration can become so short that dynamic tones are realized with virtually flat F_0 contours (Kuo et al., 2007 Xu and Wang, 2005). Thus the distribution of dynamic tones across languages is likely closely related to the constraint of minimum duration of pitch movements.

The third implication is that minimum movement duration probably underlies the phenomenon of intrinsic duration of segments (Klatt, 1976; Lehiste, 1972; Port, 1981), which, according to Klatt (1976:1215), is ''the property of an absolute minimum duration D_{min} that is required to execute a satisfactory articulatory gesture.'' For any segment to be observable, it has to be longer than the minimum duration allowed by the maximum speed of the involved articulatory movements. Otherwise, no evidence of the segment would be generated, and it would be in effect articulatorily omitted. In support of this view, we have found evidence in a preliminary study that the maximum speed of articulatory movement for vowels and consonants is often approached in normal speech (Xu, 2007b). But this happens not in stressed syllables, but in unstressed syllables, accompanied by severe undershoot. Janse (2003) finds that the intelligibility of very fast speech is much reduced when speakers are asked to speed up, but they are not able to even double their speech rate from the

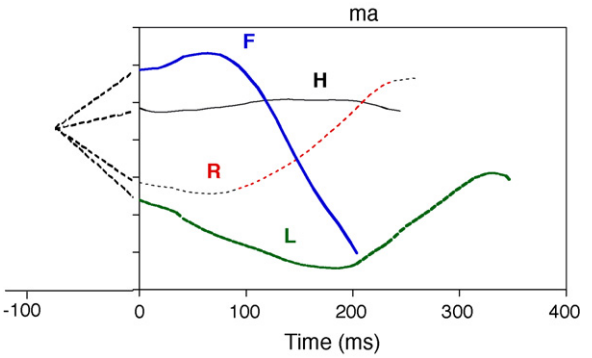

Fig. 3. Isolated Mandarin tones with hypothetical silent initial F_0 movements.

normal rate. In contrast, when normal speech is accelerated through computer resynthesis to almost three times the original rate, intelligibility still remains high, indicating that human perception is capable of processing unnaturally fast-changing speech signals. Thus it seems that it is the articulatory rather than perceptual constraints that constitute the bottleneck for the speed of information coding in speech.

Finally, minimum movement duration means that there are long articulatory transitions not only from one sound to the next, but also at the onset of an utterance. In Fig. 2a, the F_0 values of the four Mandarin tones are well separated at the voice onset. It could be the case that these values reflect vocal fold tension at the onset of articulation. But it is also possible that the tension adjustment actually starts well before the voice onset. Indeed the initial portions of the contours all seem to point back in time to an invisible common value. In Fig. 3, four back projections are added to the left of the F_0 curves to indicate an imaginary common onset point. The significance of this conjecture is not obvious at this point, but will become clearer when I discuss the second kind of obligatory timing.

To summarize, manipulating the state of the articulatory system during speech takes time, and part of that time is directly due to the minimum duration of articulatory movements. This aspect of the speech timing is thus obligatory, and as such it is not available for information coding.

2.3. Syllable-bound target alignment

The minimum movement duration just discussed does not directly determine when a movement starts and when it ends. It could be the case that there is full articulatory freedom in adjusting the onset and offset time of articulatory movements. If so, movement timing can be freely used to encode information. However, there is accumulating evidence that this is not the case. Instead, the timing of all the articulatory movements seems to be closely tied to the syllable in one way or another. In the following I will first briefly discuss several lines of evidence for syllable-bound temporal alignment. I will then explore a number of possible underlying mechanisms of such alignment.

2.3.1. Syllable-synchronized tonal alignment

The initial evidence for the lack of freedom of movement alignment comes from F_0 patterns of Mandarin tones. It is found that not only are the transitions between the adjacent tones quite long, but also they take place during the target tone itself rather than during a dedicated

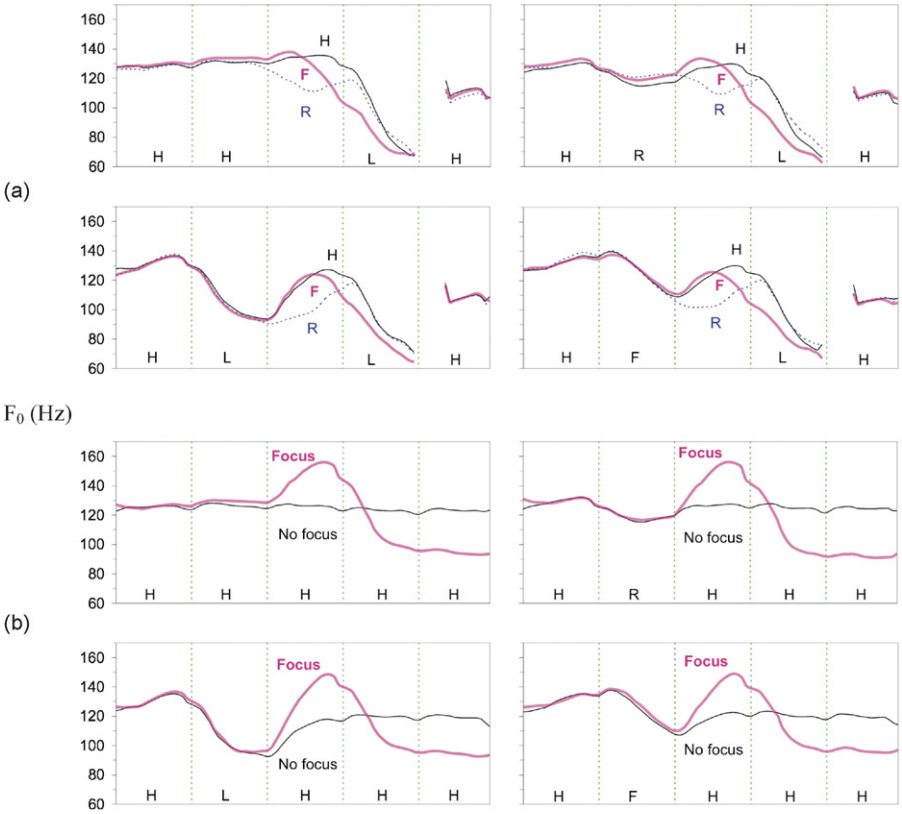

Fig. 4. Mean F_0 contours of Mandarin 5-syllable utterances. Adapted from Xu (1999).

transition interval (Xu, 1997, 1999). In Fig. 2c, for example, the F_0 contours of the H tone start at very different heights depending on the preceding tone. They then all converge to a high-level shape appropriate for the H tone. Thus the entire syllable bearing the H tone seems to be the domain of the F_0 movement toward the canonical H tone pattern. In other words, the execution of the tone-approaching movement coincides, or is synchronized, with the entire syllable.

Syllable-synchronized tonal execution can be seen more clearly in Fig. 4a, where the tone sequences in each plot differ only in the third syllable. It can be seen in these plots that the magnitude of the F_0 movements toward a tonal target varies substantially with the preceding tone. But the movements all start at a similar time around the onset of the syllable, as indicated by the vertical lines. Recall that the speed of pitch change is often as fast as possible in a dynamic tone because of minimum movement duration. The fact that the tone approaching movement does not start earlier in a dynamic tone than in a static tone or vary with the preceding tone indicates that such onset timing, just as minimum movement duration, is quite mandatory.

Syllable-synchronization of tone approaching movement is further evident in that the onset timing is not affected whether the movement is tonal or intonational. This can be seen in Fig. 4b, where the difference between the two curves in each plot is due to focus conditions on the third

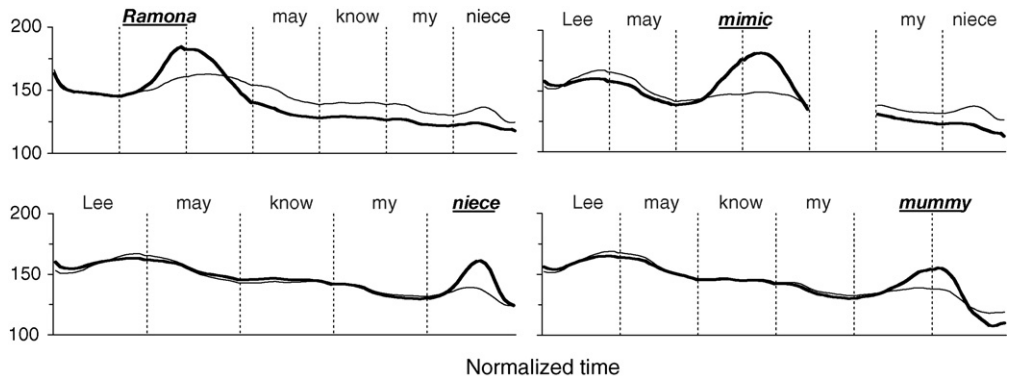

Fig. 5. Mean F_0 contours of minimal pair English sentences which differ in focus conditions. Adapted from Xu and Xu (2005).

syllable, which is a monosyllabic verb. The movement magnitude is especially large when the H tone following the L tone is focused. The onset of the sharp F_0 rise is nevertheless no earlier than in cases where the magnitude of F_0 movement is much smaller. Similarly, in English, as shown in Fig. 5, an F_0 rise starts around the onset of a stressed syllable whether or not the syllable is focused (Xu and Xu, 2005). Note that the consistent onset of target approaching movement at the syllable onset also indicates a lack of anticipatory movement toward a pitch target prior to its temporal interval, as has been discussed in Xu and Wang (2001) and Xu (2005).[1]

Further evidence comes from studies of tonal alignment variation with the composition of the syllable. Xu and Xu (2003) find that even in a syllable with a voiceless initial consonant, the target approaching movement starts from syllable onset rather than from voice onset. Similar evidence has been reported for English (Xu and Wallace, 2004) and Cantonese (Wong and Xu, 2007a). Furthermore, Xu (1998, 2001) finds that the interval of target approximation is not affected by coda consonants. Thus the interval of target approximation is the entire syllable rather than only the vocalic, or even only the voiced portion of the syllable.

2.3.2. The Target Approximation (TA) model

Based on the finding of syllable-synchronized target realization in tone production, Xu and Wang (2001) proposed the Target Approximation (TA) model, which later became a critical component of PENTA mentioned in the Introduction. Originally proposed only for lexical tones, it is presented here in a more general form, as shown in Fig. 6. According to the model, to produce a phonetic unit is to articulatorily approach its underlying target as depicted by the dashed lines. The surface form, as depicted by the solid curve, is the trajectory of such approximation. The approximation movement always starts from the initial state of the articulator, as indicated by the leftmost arrow, which is either a neutral state at the onset of articulation, or the end state of approximating the previous target. The approximation proceeds asymptotically and unidirectionally until the end of the temporal interval of the unit, by which time the target is

[1] It is theoretically possible that in languages like Japanese, the mora is also a domain of target assignment and approximation. However, it is an empirical question as to whether and how this is done independent of the syllable. What is needed are studies with systematic experiment controls like those done for Mandarin, English and Cantonese as mentioned in this paper.

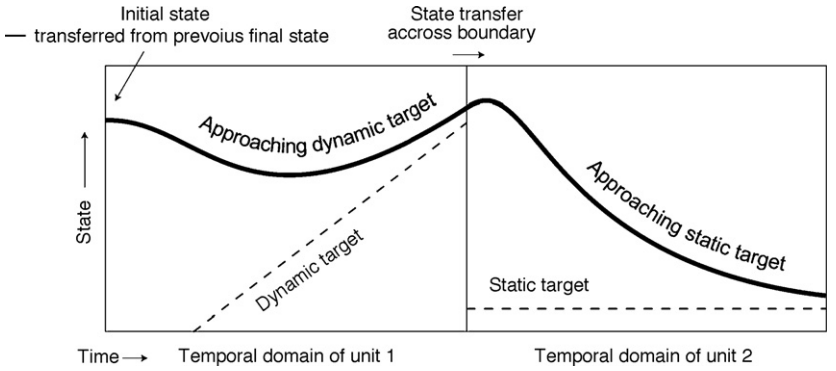

Fig. 6. The Target Approximation (TA) model. The vertical lines represent boundaries between phonetic units. The dashed lines represent underlying targets of the units. The thick curve represents the surface trajectory that results from asymptotic approximation of the targets.

either fully reached or only partially approached. The degree of approximation depends on the distance between the initial and targeted states, the temporal interval assigned to the target and the strength with which the target is approached. At the boundary of two adjacent units, the final articulatory state of the earlier unit (which may include the final displacement, velocity and acceleration of the movement, cf. Prom-on et al., 2006) is transferred to the later unit to become its initial state. The approximation of the later target starts at the boundary. But the effect of the earlier target lingers on until it is overcome by the execution of the current target. Such *carryover* effect is especially substantial in the case of a dynamic target, whose approximation results in a high final velocity which takes a long time to fully reverse. In Fig. 4a, for example, the final F_0 of the R tone in syllable 2 is consistently lower than that of the H tone, but the post-R F_0 quickly catches up with the post-H F_0 in syllable 3. In Chen and Xu (2006) it is further found that the influence of dynamic tones on the following neutral tone syllables actually surpasses that of static tones.

Two aspects of the TA model are worth particular mentioning as they are directly relevant for timing and coordination. The first is that the temporal interval of target approximation is equivalent to the temporal interval of the unit to which the target is associated. Thus the model assumes no discrepancy in the temporal interval of the unit and that of the execution of its target. The second is that the model assumes no general oscillatory mechanism that governs the approximation of successive targets. Each target approximation is initiated anew, and there is no return phase toward a neutral state either within or after each target approximation as assumed in most of the existing artilatory-based quantitative models (e.g., Fujisaki et al., 2005; Saltzman and Munhall, 1989).

2.3.3. Defining the temporal interval of segments in terms of target approximation

The proposal of the TA model has also enabled the reexamination of temporal alignments in the segmental aspect of speech. As mentioned above, the TA model assumes that the temporal interval of a unit is equal to the interval during which its target is approached. Based on this assumption, Xu and Liu (2007) examined the temporal alignment of segments and found that it follows a very restricted pattern bound to the syllable. As an illustration, in Fig. 7, F2 changes movement directions at each of the arrows (F3 at the 3rd arrow). Before each change, F2 moves toward the most characteristic pattern of the segment: [aɪ], [w] and [i]. Thus each turning point

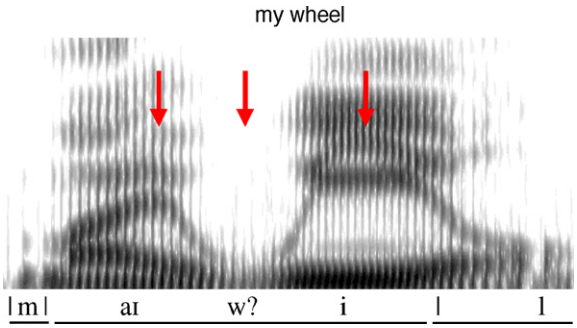

Fig. 7. Spectrogram of "my wheel". The labels are based on conventional phonetic segmentation.

can be viewed as the offset of a previous segment and onset of the following segment. In other words, the four regions divided by the three arrows are the temporal intervals of [aɪ], [w], [i] and [ɬ], respectively.

A glide like [w], however, involves incomplete closure of the vocal tract, allowing formants to be manifested continuously. Most other consonants involve complete closure of the vocal tract, which interrupts the formant movements in the spectrogram. What Xu and Liu (2007) have found is that, with the help of F_0 alignment as independent time reference, similar division as in "my wheel" can be found in "my meal" as shown in Fig. 8.

In the TA segmentation in Fig. 8, the onset of consonants is about 50 ms earlier than the conventional onset marked by landmarks such as the onset of the nasal murmur. The onset of vowels is earlier than the conventional onset by an even larger amount: in each case starting at the same time as the new onset of the preceding consonant. The new offsets of both consonants and vowels are also much earlier than the conventional offsets: at a point when the formants start to move toward the following segment(s).

The onset or offset of a movement is sometimes acoustically hidden, however. For example, in Fig. 8, the dotted curve adjoining the interrupted F2 movements is what F2 would look like had the oral cavity not been completely closed, based on articulatory data reported by Westbury and Hashi (1997) for nasals and Löfqvist and Gracco (1999) for stops. In other words, the segmental division here is no different from that in Fig. 7. Also, at the beginning of an utterance, the onset of the articulatory movements is not acoustically manifest, as it takes time for the vocal folds to be set into vibration. As a result, the acoustic onset is presumably much later than the articulatory onset, as has been illustrated in Fig. 3.

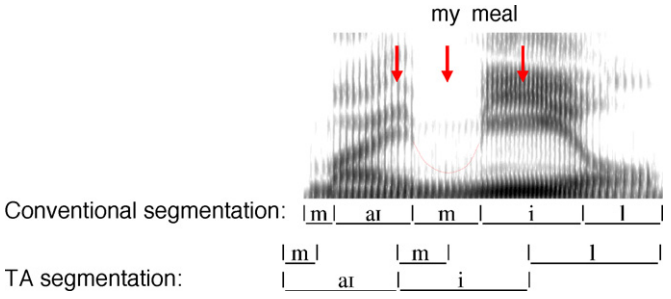

Fig. 8. Conventional versus TA segmentations of "my meal".

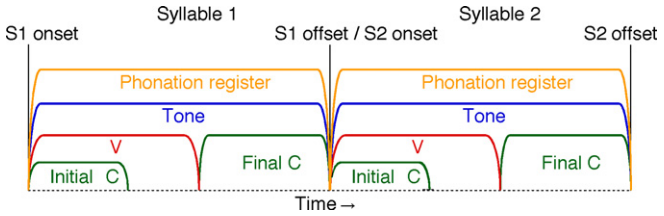

Fig. 9. The time structure model of the syllable. Adapted from Xu and Liu (2006).

2.3.4. The time structure model of the syllable

The findings of Xu and Liu (2007) have motivated the proposal of 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 of consonants, vowels, tones and phonation registers, as illustrated in Fig. 9. 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 being approached. The TA segmentation shown in Fig. 8 is actually an illustration of the segmental alignment based on the time structure model of the syllable.

2.4. Possible mechanisms of syllable-bound target alignment

While syllable-bound target alignment has empirical support as discussed above (see also detailed discussion in Xu, 2005 and Xu and Liu, 2007), its underlying mechanism is far less clear. Here I will briefly discuss a number of possibilities: entrainment, action synchronization, and time marking.

2.4.1. Entrainment

Entrainment refers to a process whereby two nearby oscillating systems, such as clocks, having similar periods, fall into synchrony (Huygens, 1666 as cited by Spoor and Swift, 2000). Many synchronization phenomena in motor movements, including those of speech, have been likened to this process (Haken et al., 1985; Kelso et al., 1986). Haken et al. (1985) proposed a non-linearly coupled oscillator model to account for entrainment in motor movements. The coupled oscillator model has also been used to explain motor movements in speech (Saltzman and Munhall, 1989; Cummins and Port, 1998; Port, 2003). There are several considerations, however, that make entrainment an unlikely mechanism underlying the synchronization behavior in speech. First, entrainment is a rather slow process, taking many cycles for two oscillators to be synchronized. In contrast, the phase shift phenomenon in both limb and speech movements, whereby a 180° phase relation between two movements is shifted to 0°, is completed in just a couple of cycles (Kelso, 1984). Such near instantaneous phase shift is not well simulated by the coupled-oscillator model, which takes quite a few cycles to complete the phase change (Haken et al., 1985). Secondly, the classical entrainment occurs between two physical systems that are independent of each other, whereas the proposed motor movement equivalents typically occur within the same person with a single central control. Even in cases where two individuals are involved (Schmidt et al., 1990), the participating individuals are watching each other's action, and thus exchanging control information. Such information exchange is critical according to

Mechsner et al. (2001), who demonstrate that motor synchronization is achieved through high-level sensorial control rather than low-level mechanical or neural linkage. Finally, and probably most importantly, in speech the syllable-bound temporal alignment occurs even when the movements are not repetitive, such as in a monosyllabic word said in isolation. Even in continuous speech, recurrent motor movements are not oscillatory in the strict sense. For example, the size of the vocal tract opening and degree of its closure vary from syllable to syllable, depending on the vowels and consonants that happen to be involved. Also, syllable duration is neither constant, nor gradually changing over time, but changing abruptly from one syllable to the next, as will be discussed later. So, any proposed synchronization mechanism has to be applicable to both recurrent and non-recurrent movements, which would apparently rule out the entrainment hypothesis.

2.4.2. Action synchronization

Kelso et al. (1979) find that, when asked to reach two targets of different degrees of difficulty, as determined by the size of the target and its distance from the reaching hand, subjects always synchronize the two manual movements, such that their onsets, offsets and peak velocities all coincide in time. They further find that such synchronization is achieved by slowing down the easier, thus faster, movement to accommodate the more difficult one. Reaching as a goal oriented movement is similar to what is captured by the TA model (Xu and Wang, 2001). The target approximation process can be simulated as a second order forced oscillation system, in which the target is the force function while the oscillator is a damped second order system (Prom-on et al., 2006). In such a system, the adjacent force functions are mechanically unrelated to each other, and each target approximation movement is thus initiated anew. Being critically damped or overdamped, there is no return phase after each target approximation. This is unlike a true oscillatory system, in which the system enters the return phase each time after the equilibrium point is reached.

There is, however, one aspect of the syllable-bound alignment pattern that does not fit the typical pattern of synchronized actions. That is, as depicted by the time structure model of the syllable, while the laryngeal and vocalic movements are largely synchronized (i.e. both with the entire syllable), the consonantal movements share only onset time with the rest of the movements. The consonantal offsets occur at a non-fixed time during the other movements, relatively earlier when the syllable is long, but later when the syllable is short.

2.4.3. The time marker hypothesis

One of the fundamental but unresolved issues about speech is how the segmentation of a continuous speech signal is done. No one to my knowledge has yet to offer a clear solution to this problem. What is often assumed is that the signal is first segmented into syllable-sized units. But it is not clear how this is done.

The timing pattern specified by the time structure model of the syllable may provide a clue as to how initial segmentation could be done. That is, syllable onsets, where the unidirectional movements toward the initial C, the first V, the tone, and possibly the phonation register all start simultaneously, probably serve as time markers in speech. A time marker is an event that functions as a reference for the measurement of time and timing (Jones and Boltz, 1989). A clock is a device that generates time markers for judging the relative timing of other events. Speech involves many timing patterns. But to control timing in production and to detect timing in perception, a reference system is needed. It is impossible to use an external clock for such purpose, for obvious reasons. A common biological clock shared by all speakers is not likely either, because the speed of articulation varies extensively not only across speakers, but also

within the same speaker, as will become clear later. What is needed are re-current events in the speech flow itself, generated by the speaker, that can serve as unambiguous time markers. Such time markers are critical for the perception of timing in events in which the temporal components are not fixed, as in music (Jones and Boltz, 1989; Large and Jones, 1999). Common onsets of multiple unidirectional movements would conceivably serve this purpose. According to this hypothesis, which has yet to be specifically tested, it is the need to have recurrent co-onsets of events to serve as time markers that gives rise to the temporal organization of the syllable as captured by the time structure model.

3. Informational timing

From the articulatory-functional perspective, actively controlled timing is for the sake of conveying information. This means that for any timing pattern that is suspected of being actively controlled, it is always necessary to ask what is the specific communicative function that is being conveyed, which of the specific articulatory parameters is being controlled and what particular value of the parameter is being specified. But to be able to do that, it is necessary to first explicate what is actually controllable given the articulatory constraints discussed thus far.

3.1. What is available as means of information coding, what is not

Any aspect of the acoustic signal produced in speech, including timing, is potentially useful for encoding information. From the articulatory-functional perspective, however, there is a critical difference between timing that is directly controlled for coding purposes and timing that is the consequence of controls that are not temporal in nature. The articulatory constraints discussed in the first half of this paper seem to set severe limits on what aspects of timing can be directly controlled and what aspects cannot. First, due to the syllable-bound temporal alignment, a pitch target assigned to a syllable has to be implemented in synchrony with it. There is thus no room for micro-adjusting this alignment, making it unavailable for information coding. Secondly, due to minimum duration of articulatory movement, the synchronous implementation of a target with the syllable necessarily generates certain F_0 turning points whose alignment is dependent on the characteristics of both the current target and the surrounding targets. In Fig. 10a, for example, the three apparent F_0 peaks in Mandarin are the consequences of not only the underlying pitch targets of [high], [rise] and [fall] for the H, R and F tones, respectively, but also the fact that they happen to be surrounded by the L tone. When they are surrounded by the H tone as in Fig. 10b, only the F_0 turning point in the F tone can marginally qualify as a peak. Furthermore, the exact locations of F_0 valleys in Fig. 10c and peaks in Fig. 10d vary depending on the preceding tone. The nature of the variation is quite mechanical: the greater vertical distance F_0 has to travel before changing directions, the later the turning point. Such variability is in sharp contrast with the consistency of target approximation movements in Fig. 10c toward [rise] and in Fig. 10d toward [fall], which remain the same regardless of tonal contexts. In general, therefore, the temporal interval of target approximation is determined by the assignment of the target to a particular syllable; and target property, surrounding tones and syllable duration jointly determine the detailed peak/valley alignment. This means that temporal alignment of F_0 turning points also cannot be directly controlled for information coding.

Note that such lack of freedom of microscopic temporal control could be contradicted by a number of reported production and perception data on alignment. For production, it has been found that there are cross-language or even cross-dialect differences in F_0 turning point alignment (Arvaniti and Garding, 2007; Atterer and Ladd, 2004; D'Imperio et al., 2007). Such differences

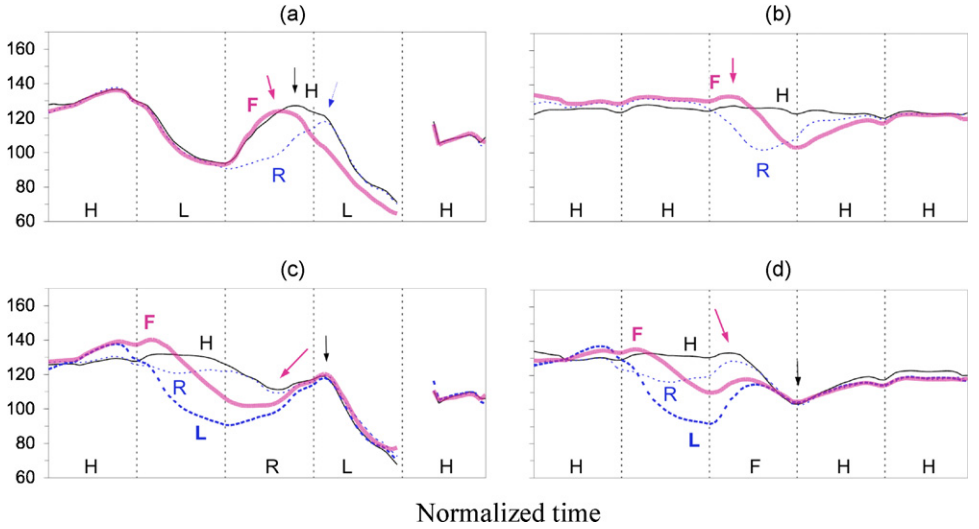

Fig. 10. Illustration of variability and stability of F_0 turning point alignment. Adapted from Xu (1999).

could have been due to differences in the underlying pitch targets, however. For example, the onset time of F_0 rise in the R tone in Fig. 10c varies depending on the preceding tone, but the rise onset in the F tone in Fig. 10d remains constant regardless of the preceding tone. The difference is due to the nature of the rise, which is the start of the target approaching movement in the F tone, but an intermediate point during target approximation in the R tone. In cases where the underlying targets are not yet as clear as in Mandarin, it is critical to perform minimal-pair comparisons by varying both the pitch of the syllable in question and that of the surrounding syllables.

For perception, it has been reported that listeners are sensitive to experimental manipulations of turning point locations (e.g., Kohler, 2005). As is apparent from the discussion so far, any change in the pitch target assignment to a particular syllable, or to a sequence of syllables, will necessarily lead to apparent F_0 alignment differences. This means that any manipulation of turning-point alignment is likely to also change other aspects of the F_0 trajectories, which could be used by listeners to infer the nature of the underlying targets rather than the turning point or its temporal alignment. In other words, the fact that alignment differences affect perception does not necessarily mean that information is coded in terms of timing.

With the lack of freedom in micro-controlling the temporal alignment of underlying pitch targets and the surface location of F_0 turning points, what is still available for information coding is duration. This is nonetheless a very large control space. For example, it has been shown that the duration ratio of the longest to the shortest rhythmic units, and presumably the syllables within them, is as much as 6 or 7 to 1 (Hill et al., 1992). And this does not include variations in pause duration, which are even more extensive. Thus there should be sufficient space to allow the kind of parallel encoding of multiple layers of information seen in the pitch space (Xu, 1999; Xu and Xu, 2005), as captured by PENTA (Xu, 2005).

3.2. Timing for encoding lexical contrast and focus

There are quite a few communicative functions that seem to be encoded predominantly or partially through timing. What is critical for understanding each timing-related function is to

explicate what specific communicative meaning is being coded and how it is encoded. Probably the clearest case is the encoding of lexical contrast. Many languages use duration differences to distinguish words. In some cases, duration seems to be the predominant property for such lexical contrast. This is the case in quantity languages like Japanese, Thai, Finnish, Icelandic and Estonian (Abramson and Ren, 1990; Hertrich and Ackermann, 1997; Hirata, 2004; Pind, 1999; Suomi, 2005). In these languages, vowels often carry a two-way duration contrast: short and long. In some languages there is even a three-way contrast: short, long and extra long, e.g., Estonian (Traunmüller and Krull, 2003 and references therein). A common characteristic shared by these quantity contrasts is that the duration ratio between the short and long vowels is rather high, usually at least 2:1 and sometimes even larger, as can be seen in the following list.

Duration ratio of long vs. short vowels:
Thai: 2: 1 (Abramson and Ren, 1990)
Japanese: 2.5: 1 (Hirata, 2004)
Finnish: 2.5: 1 (Suomi, 2005)
Icelandic: 1.95: 1 (Pind, 1999)
German: 1.7: 1 (calculated from Table II of Hertrich and Ackermann, 1997)

Duration is also known to help code lexical contrasts related to word stress. In English, for example, although word stress typically has acoustic correlates such as vowel quality, intensity and F_0, the stressed/unstressed duration ratio is still quite high: 2.18:1 according to Crystal and House (1988). In Mandarin, though there is no equivalent of word stress in English, the neutral tone bears some similarity to the English weak stress (Chen and Xu, 2006; Xu and Xu, 2005), and the full tone to neutral tone duration ratio is about 1.7:1 (Lin, 1985; Chen and Xu, 2006).

Duration is also known to participate in making focal contrast. Focus has been consistently found to lengthen the lexical item being focused (Turk and Shattuck-Hufnagel, 2000; Xu, 1999; Xu and Xu, 2005). However, the ratio of focused to non-focused duration is generally much lower than has been found for lexical stress, 1.17:1 in Mandarin (Xu, 1999), 1.25:1 (Turk and Shattuck-Hufnagel, 2000) or 1.14:1 (Xu and Xu, 2005) in English, and 1.09:1 in Dutch (Sluijter and van Heuven, 1996). The reason for these relatively low ratios is probably because the most effective cue of focus is F_0, which varies extensively not only in the focused item, but also in post-focus items (Rump and Collier, 1996; Xu, 1999; Xu and Xu, 2005). It is also possible that the duration increase under focus is to allocate sufficient time for the focally expanded pitch range to be articulatorily realized. In a recent modeling study using the TA model, focus generated with expanded underlying pitch range (by adjusting the target parameters) as well as properly increased duration (as learned from natural speech) resulted in larger surface pitch range expansion than focus generated with underlying pitch range expansion only (Prom-on et al., in preparation). And focus identification by human listeners was better for the former than the latter. The perception support for this duration-for-pitch-range-expansion hypothesis is only partial, however, as it is not yet known whether increased syllable duration alone can also improve focus perception.

3.3. Timing for timing's sake or timing as communicative information?

While durational contrasts for lexical distinctions and focus both seem to be unmistakably communicative, in some cases the communicative functions of the timing patterns are much less clear. This is particularly true for what is known as the rhythm class hypothesis, originally

proposed based on non-instrumental observations (Pike, 1945). According to the hypothesis, there is a universal tendency for certain units to become equal in duration, and that languages of the world can be divided into three rhythm classes depending on the kind of unit involved in manifesting the isochrony tendency: stress-timed, syllable-timed and mora-timed (Abercrombie, 1967; Bloch, 1950; Pike, 1945). Although later empirical research has shown that no true isochrony can be found (Hill et al., 1992; Lehiste, 1977; Nakatani et al., 1981; Warner and Arai, 2001), weak tendencies toward isochrony have nonetheless been demonstrated at least for stress-timing (Hill et al., 1992; Hirst and Bouzon, 2005). A more critical issue from the articulatory-functional perspective, however, is whether the rhythmic tendency is an articulatory mechanism or an encoding scheme for certain communicative information, and in either case, whether it is a basic mechanism or a secondary phenomenon.

Efforts to reveal the nature of speech rhythm can be seen in two recent proposals. The first is that rhythmic patterns help infants to distinguish between different languages in a multi-lingual environment (Ramus et al., 1999), and thus it is the functional pressure of language acquisition that forces each class of languages (often unrelated to each other) to develop a distinct rhythmic pattern. An obvious question about this proposal is whether the proposed pressure is strong enough for unrelated languages to automatically evolve themselves into one of the rhythm classes. The second proposal is that rhythm is generated by periodic pulses produced by *neurocognitive oscillators* in the brain which guides both speech production and perception (Port, 2003). Since this proposal is based on timing patterns observed when subjects are asked to articulate repetitive phrases, one may wonder about its relevance to normal speech which is rarely repetitive. But a common question for both of these proposals as well as the original rhythmic hypothesis is that, assuming that a rhythmic tendency does exist, can it be accounted for by mechanisms that are either functional or articulatory but are non-rhythmical by nature?

Indeed, the isochrony tendency can be linked to certain temporal patterns reported in duration research, which is not typically aimed at understanding rhythm (Lehiste, 1972; Turk and Shattuck-Hufnagel, 2000). In particular, it has been found that a stressed stem syllable is continuously shortened as more syllables are added to its right in the same word. For example, the English syllable *sleep* is longest when produced as a monosyllabic word, shorter in *sleepy*, and shorter still in *sleepiness* (Lehiste, 1972). Known as *polysyllabic shortening* (Lehiste, 1972; Turk and Shattuck-Hufnagel, 2000), this durational pattern in effect generates a tendency toward equal duration for words of different lengths.[2] Such a tendency is in the same direction as the isochrony tendency proposed for rhythm. Turk and Shattuck-Hufnagel (2000) have found that adding a function word to the right of a monosyllabic word also shortens it, suggesting that the interval of polysyllabic lengthening may be larger than a word, which makes the two tendencies even more alike.

To add one more dimension to the puzzle, both rhythm and duration patterning could be related to the prosodic hierarchy hypothesis, according to which there exists an organizational structure that helps to determine various aspects of the speech signal that are not lexically fully specified (Beckman, 1996; Shattuck-Hufnagel and Turk, 1996). This prosodic structure is proposed to consist of a hierarchy of constituents of different sizes: intonational phrase, prosodic phrases, prosodic words, clitic groups, metrical feet, etc. (Shattuck-Hufnagel and Turk, 1996). The link of prosodic structure to rhythm is that the definitions of the smallest constituents, namely, prosodic words, clitic groups and metrical feet, all refer to word stress, and word stress is

[2] Polysyllabic shortening has been questioned by Nakatani et al. (1981), who argue that there is only phrase-final lengthening, but no general tendency to shorten syllables as word size increases.

what is supposed to recur at near even time intervals in a stress-timed language according to the rhythm class hypothesis.

A recent study of Mandarin tonal patterns, however, offers a new insight into the issue of prosodic organization (Xu and Wang, 2005). Mandarin has no equivalent of word stress in English, although the neutral tone in Mandarin, which occurs only in a small number of words, bears some similarity to the English weak stress (Chen and Xu, 2006 ; Xu and Xu, 2005). With no word stress, Mandarin cannot be stress-timed, although I am not aware of any serious claim that it is syllable-timed. Would Mandarin, then, exhibit the kind of duration pattern found in English such as polysyllabic shortening? This is indeed found to be the case, i.e., phrase duration decreases as the number of syllables in the phrase increases. That polysyllabic shortening exists in a language that cannot be stress-timed suggests that the tendency is probably not really one of regular occurrence of stressed syllables.

Xu and Wang (2005) further find that a 4-syllable phrase in Mandarin exhibits a duration pattern of 3 1 2 4, where a larger number represents longer duration, and a 3-syllable phrase has the duration pattern of 2 1 3. Thus longer durations occur at the edges of a phrase. Interestingly, this is consistent with the duration patterns of initial and final lengthening in English (Turk and Shattuck-Hufnagel, 2000). That longer durations occur at the edges of a constituent seems to suggest that duration serves an edge/boundary marking function. This seems to be supported by the further finding by Xu and Wang that duration adjustment is sensitive to variations in morphosyntactic structure. When 4-syllable phrase changed from consisting of two disyllabic words to one monosyllabic word followed by a tri-syllabic word, the duration pattern changed from 3 1 2 4 to 3 2 1 4. With this change, the second syllable, which has become word initial, is lengthened, while the third syllable, which has become word medial, is shortened.

What, then, is the nature of these duration patterns? The edge-marking durational adjustment is consistent with the finding that pre-boundary duration is related to boundary strength (Beckman and Edwards, 1990; Lehiste et al., 1976; Shattuck-Hufnagel and Turk, 1996; Wightman et al., 1992). But it is also known that a strong boundary is likely to be also associated with a pause (Lea, 1980; O'Malley et al., 1973). Thus both pre-boundary duration and pause duration affect the distance between the onset of the pre-boundary constituent and the onset of the post-boundary constituent. In polysyllabic shortening, the close relations between the syllables being shortened are in contrast to the detachment signaled by pauses. It is thus possible that the duration of a pre-boundary syllable serves as an *affinity index* that signals how closely two adjacent constituents adheres to each other. If this affinity index hypothesis is valid, the non-lexical and non-focal durational patterns reported by previous research on rhythm, duration and prosodic hierarchy all could be driven by a single communicative function: to signal the closeness of adjacent constituents. Apparently simplistic and sweeping, this hypothesis certainly needs rigorous testing in future research.

Finally, recent research by Dellwo (2007) has shown further evidence against the cohesiveness of rhythm class hypothesis. What he has found is that stress-timed languages like German and English can be made to sound more syllable-timed simply by having speakers increase their speech rate. As is known, polysyllabic shortening mainly affects the duration of stressed syllables in English (Turk and Shattuck-Hufnagel, 2000). Hence, what makes fast English sound syllable-timed at high speed is likely the incompressibility of syllable duration due to the intrinsic duration of segments as discussed in section 2.2. If this is true, perceptual impression of syllable-timing is likely due to an articulatory constraint rather than a functional specification, and as such it is mechanistically different from stress-timing. As for mora-timing, an extensive review by Warner and Arai (2001) has shown that, despite decades of research, there is lack of

consistent evidence even in the sense of speaker compensation to make moras occur regularly in time. From the articulatory-functional point of view, given that the lexical function of mora is clear, true mora-timing has to be demonstrated as duration adjustments *in excess of* what is required by its functional demand.

Overall, then, the evidence considered so far suggests that speech rhythm, rather than being a basic mechanism, is likely an epiphenomenon derived from a number of independent articulatory and functional mechanisms that have little to do with isochrony.

3.4. Parallel encoding of information with timing

According to PENTA, multiple layers of information may be encoded in parallel through the articulatory process of target approximation. For speech timing, existing research has shown some evidence for such parallel encoding. For example, it has been argued that the reason for the high duration ratio in making lexical contrasts in quantity languages like Japanese and Finnish is to guarantee that the lexical contrast remains sufficiently maintained when the duration of the involved segments varies due to non-lexical factors such as speech rate, focus, grouping and phrasing (Hirata, 2004; Jacobsen, 2000; Pind, 1999; Suomi, 2005; Traunmüller and Krull, 2003). For example, Jacobsen (2000) has shown that prepausal lengthening also occurs in Greenlandic, a quantity language with a 2-way durational contrast. Suomi (2005) reported accentual lengthening in Finnish, also a 2-way quantity language. In Mandarin, the durational difference between full tone and neutral tone is maintained under focus, although a full tone is lengthened more than a neutral tone (Chen and Xu, 2006).

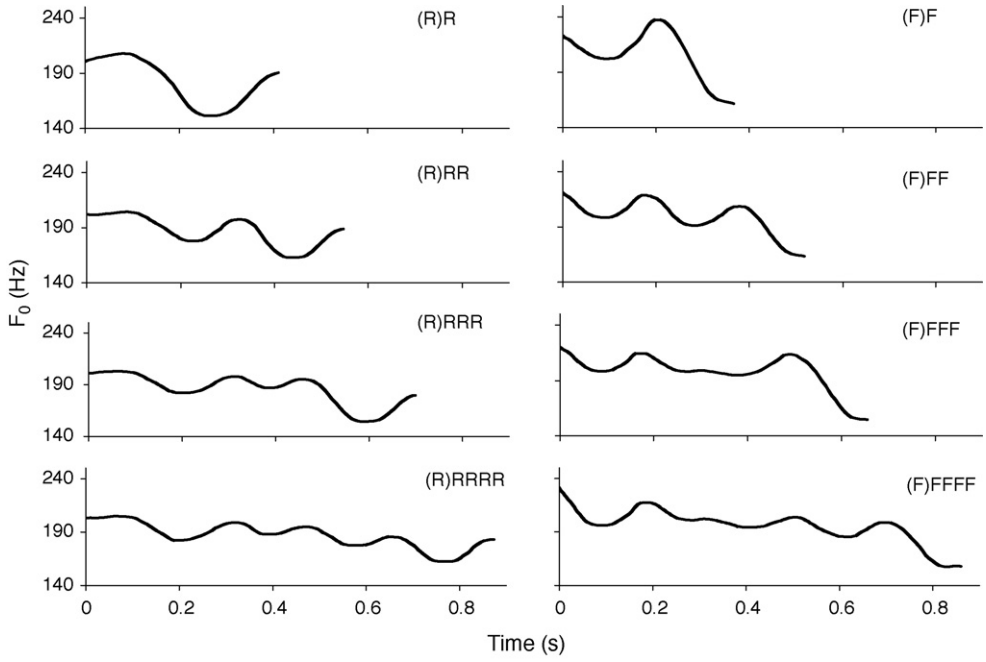

Fig. 11. Mean F_0 contours of (non-final) consecutive R and F sequences with varying numbers of syllables. Adapted from Xu and Wang (2005).

Similarly, in Arabic, English and Dutch (de Jong, 2004; de Jong and Zawaydeh, 2002; Sluijter and van Heuven, 1996), durational differences due to lexical stress are also retained under focus. So the encoding of tone and lexical stress with duration is done in parallel with that of focus in these languages.

Parallel encoding also happens when two or more different articulatory/acoustic dimensions need to be simultaneously controlled. Polysyllabic shortening and phrase-final lengthening in Mandarin discussed above are done together with the encoding of lexical tones with F_0, and they seem to have direct effects on the F_0 contours, as shown in Fig. 11. Here in a sequence of R or F tones, larger F_0 excursion sizes are associated with longer durations, and the 2 1 3 and 3 1 2 4 duration patterns in the 3-syllable and 4-syllable phrases are clearly matched by corresponding sizes of the F_0 contours (Xu and Wang, 2005). It is possible, of course, that a longer duration and larger F_0 excursion are both due to increased stress, hence greater articulatory strength and/or prominence. But this is found not to be the case, as longer syllables have neither greater stiffness (as indicated by the ratio of peak velocity of F_0 movement to amplitude of the movement, cf. Ostry and Munhall, 1985 for the nature of this parameter) nor higher F_0, once all the contributing factors are controlled. Thus the duration patterns seem to have directly affected the realization of the tonal targets without the mediation of stress or prominence.

4. Conclusion

While much more needs to be learned about speech timing, I hope to have demonstrated in this paper the benefit of understanding timing from an articulatory-functional perspective, i.e., separating as clearly as possible what is obligatory from what is informational. The understanding of obligatory timing can help us recognize aspects of it that cannot be controlled for information coding. The understanding of informational timing can help us identify true sources of timing control as opposed to proposed mechanisms that are neither articulatory nor functional.

References

Abercrombie, D., 1967. Elements of General Phonetics. Edinburgh University Press, Edinburgh.

Abramson, A.S., Ren, N., 1990. Distinctive vowel length: duration versus spectrum in Thai. Journal of Phonetics 18, 79–92.

Arvaniti, A., Garding, X., 2007. Dialectal variation in the rising accents of American English. In: Cole, J., Hualde, J. (Eds.), Papers in Laboratory Phonology IX: Change in Phonology. Mouton de Gruyter, The Hague, pp. 547–576.

Atterer, M., 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., 1996. The parsing of prosody. Language and Cognitive Processes 11, 17–67.

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

Bloch, B., 1950. Studies in colloquial Japanese IV: phonemics. Language 26, 86–125.

Bolinger, D., 1989. Intonation and Its Uses—Melody in Grammar and Discourse. Stanford University Press, Stanford, CA.

Chen, Y., Xu, Y., 2006. Production of weak elements in speech—evidence from f0 patterns of neutral tone in standard Chinese. Phonetica 63, 47–75.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York.

Crystal, T.H., House, A.S., 1988. Segmental durations in connected-speech signals: syllabic stress. Journal of the Acoustical Society of America 83, 1574–1585.

Cummins, F., Port, R., 1998. Rhythmic constraints on stress timing in English. Journal of Phonetics 26, 145–171.

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., Zawaydeh, B., 2002. Comparing stress, lexical focus, and segmental focus: patterns of variation in Arabic vowel duration. Journal of Phonetics 30, 53–75.

Dellwo, V., 2007. Influences of speech rate on the perception of speech rhythm. In: Presentation at CHC Workshop on Second Language Learning and Cross-Language Comparisons, University College London.

D'Imperio, M., Espesser, R., Lœvenbruck, H., Menezes, C., Nguyen, N., Welby, P., 2007. Are tones aligned with articulatory events? Evidence from Italian and French. In: Cole, J., Hualde, J. (Eds.), Papers in Laboratory Phonology IX: Change in Phonology. Mouton de Gruyter, The Hague, pp. 577–608.

Duanmu, S., 1994. Syllabic weight and syllable durations: a correlation between phonology and phonetics. Phonology 11, 1–24.

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

Gandour, J., Potisuk, S., Dechongkit, S., 1994. Tonal coarticulation in Thai. Journal of Phonetics 22, 477–492.

Gordon, M., 1999. Syllable weight: Phonetics, phonology, and typology. Ph.D. Dissertation. UCLA.

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

Gussenhoven, C., 2002. Intonation and interpretation: phonetics and Phonology. In: Proceedings of the 1st International Conference on Speech Prosody, Aix-en-Provence, France, pp. 47–57.

Haken, H., Kelso, J.A.S., Bunz, H., 1985. A theoretical model of phase transitions in human hand movements. Biological Cybernetics 51, 347–356.

Hertrich, I., Ackermann, H., 1997. Articulatory control of phonological vowel length contrasts: kinematic analysis of labial gestures. Journal of the Acoustical Society of America 102, 523–536.

Hill, D.R., Schock, C.-R., Manzara, L., 1992. Unrestricted text-to-speech revisited: rhythm and intonation. In: Proceedings of Second International Conference on Speech and Language Processing, Banff, Alberta, Canada, pp. 1219–1222.

Hirata, Y., 2004. Effects of speaking rate on the vowel length distinction in Japanese. Journal of Phonetics 32, 565–589.

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

Jacobsen, B., 2000. The question of 'stress' in West Greenlandic: an acoustic investigation of rhythmicization, intonation, and syllable weight. Phonetica 57, 40–67.

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

Jones, M.R., Boltz, M., 1989. Dynamic attending and responses to time. Psychological Review 96, 459–491.

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., 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., Goodman, D., 1979. On the nature of human interlimb coordination. Science 203, 1029–1031.

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.

Kohler, K., 2005. Timing and communicative functions of pitch contours. Phonetica 62, 88–105.

Kuo, Y.-C., Xu, Y., Yip, M., 2007. The phonetics and phonology of apparent cases of iterative tonal change in Standard Chinese. In: Gussenhoven, C., Riad, T. (Eds.), Tones and Tunes, vol. II, Phonetic and Behavioural Studies in Word and Sentence Prosody. Phonology and Phonetics Series. Mouton de Gruyter, Berlin, pp. 211–237.

Ladd, D.R., 1996. Intonational Phonology. Cambridge University Press, Cambridge.

Ladd, D.R., 2004. Segmental anchoring of pitch movements: autosegmental phonology or speech production? In: Quené, H., v. Heuven, V. (Eds.),On Speech and Language: Essays for Sieb B. Nooteboom. LOT, Utrecht, pp. 123–131.

Large, E.W., Jones, M.R., 1999. The dynamics of attending: how people track time-varying events. Psychological Review 106, 119–159.

Lea, W., 1980. Trends in Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Lehiste, I., 1972. The timing of utterances and linguistic boundaries. Journal of the Acoustical Society of America 51, 2018–2024.

Lehiste, I., 1977. Isochrony reconsidered. Journal of Phonetics 5, 253–263.

Lehiste, I., Olive, J.P., Streeter, L.A., 1976. Role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America 60, 1199–1202.

Lin, T., 1985. Preliminary experiments on the nature of Mandarin neutral tone. In: Lin, T., Wang, L. (Eds.), Working Papers in Experimental Phonetics. Beijing University Press, Beijing, pp. 1–26 (in Chinese).

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

Löfqvist, A., Gracco, L., 1999. Interarticulator programming in VCV sequences: lip and tongue movements. Journal of the Acoustical Society of America 105, 1864–1876.

Mechsner, F., Kerzel, D., Knoblich, G., Prinz, W., 2001. Perceptual basis of bimanual coordination. Nature 414, 69–73.

Nakatani, L.H., O'Connor, K.D., Aston, C.H., 1981. Prosodic aspects of American English speech rhythm. Phonetica 38, 84–106.

Nelson, W.L., 1983. Physical principles for economies of skilled movements. Biological Cybernetics 46, 135–147.

O'Malley, M.H., Kloker, D.R., Dara-Abrams, B., 1973. Recovering parentheses from spoken algebraic expressions. IEEE Transaction on Audio and Electroacoustics AU-21, pp. 217–220.

Ostry, D.J., Munhall, K.G., 1985. Control of rate and duration of speech movements. Journal of the Acoustical Society of America 77, 640–648.

Pierrehumbert, J., 1980. The phonology and phonetics of English intonation. Ph.D. Dissertation. MIT, Cambridge, MA.

Pierrehumbert, J., 1990. Phonological and phonetic representation. Journal of Phonetics 18, 375–394.

Pike, K.L., 1945. The Intonation of American English. University of Michigan Press, Ann Arbor.

Pind, J., 1999. Speech segment durations and quantity in Icelandic. Journal of the Acoustical Society of America 106, 1045–1053.

Port, R.F., 1981. Linguistic timing factors in combination. Journal of the Acoustical Society of America 69, 262–274.

Port, R.F., 2003. Meter and speech. Journal of Phonetics 31, 599–611.

Prom-on, S., Xu, Y., Thipakorn, B., 2006. Quantitative target approximation model: simulating underlying mechanisms of tones and intonations. In: Proceedings of the 31st International Conference on Acoustics, Speech, and Signal Processing. Toulouse, France I-749-752.

Ramus, F., Nesporb, M., Mehlera, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73, 265–292.

Rump, H.H., Collier, R., 1996. Focus conditions and the prominence of pitch-accented syllables. Language and Speech 39, 1–17.

Saltzman, E.L., Munhall, K.G., 1989. A dynamical approach to gestural patterning in speech production. Ecological Psychology 1, 333–382.

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

Shattuck-Hufnagel, S., Turk, A.E., 1996. A prosody tutorial for investigators of auditory sentence processing. Journal of Psycholinguistic Research 25 (2), 193–247.

Sluijter, A.M.C., van Heuven, V.J., 1996. Spectral balance as an acoustic correlate of linguistic stress. Journal of the Acoustical Society of America 100, 2471–2485.

Spoor, P.S., Swift, G.W., 2000. The Huygens entrainment phenomenon and thermoacoustic engines. Journal of the Acoustical Society of America 108, 588–599.

Suomi, K., 2005. Temporal conspiracies for a tonal end: segmental durations and accentual f0 movement in a quantity language. Journal of Phonetics 33, 291–309.

Tanaka, H., Krakauer, J.W., Qian, N., 2006. An optimization principle for determining movement duration. Journal of Neurophysiology 95, 3875–3886.

Traunmüller, H., Krull, D., 2003. The effect of local speaking rate on the perception of quantity in Estonian. Phonetica 60, 187–207.

Turk, A.E., Shattuck-Hufnagel, S., 2000. Word-boundary-related duration patterns in English. Journal of Phonetics 28, 397–440.

Warner, N., Arai, T., 2001. Japanese Mora-timing: a review. Phonetica 58, 1–25.

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

Wightman, C.W., Shattuck-Hufnagel, S., Ostendort, M., Price, P.J., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. Journal of the Acoustical Society of America 91 (3), 1707–1717.

Wong, Y.W., 2006. Contextual tonal variations and pitch targets in Cantonese. In: Proceedings of Speech Prosody. Dresden, Germany, 2006 PS3-13-199.

Wong, Y.W., Xu, Y., 2007. Consonantal perturbation of f0 contours of Cantonese tones. Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, pp. 1293–1296.

Xu, C.X., Xu, Y., 2003. Effects of consonant aspiration on Mandarin tones. Journal of the International Phonetic Association 33, 165–181.

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., 2004. Understanding tone from the perspective of production and perception. Language and Linguistics 5, 757–797.
Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. Speech Communication 46, 220–251.
Xu, Y., 2007a. Speech as articulatorily encoded communicative functions. Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrücken, pp. 25–30.
Xu, Y., 2007b. How often is maximum speed of articulation approached in speech? Journal of the Acoustical Society of America 121 (Pt 2), 3140–3199.
Xu, Y., Liu, F., 2007. Determining the temporal interval of segments with the help of F0 contours. Journal of Phonetics 35, 398–420.
Xu, Y., Liu, F., 2006. Tonal alignment, syllable structure and coarticulation: toward an integrated model. Italian Journal of Linguistics, 18, 125–159.
Xu, Y., Sun, X., 2002. Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111, 1399–1413.
Xu, Y., Wallace, A., 2004. Multiple effects of consonant manner of articulation and intonation type on F0 in English. Journal of the Acoustical Society of America 115 (Pt 2), 2397.
Xu, Y., Wang, Q.E., 2001. Pitch targets and their realization: evidence from Mandarin Chinese. Speech Communication 33, 319–337.
Xu, Y., Wang, M., 2005. Tonal and durational variations as phonetic coding for syllable grouping. Journal of the Acoustical Society of America 117, 2573.
Xu, Y., Xu, C.X., 2005. Phonetic realization of focus in English declarative intonation. Journal of Phonetics 33, 159–197.
Zhang, J., 2002. The Effects of Duration and Sonority on Contour Tone Distribution—A Typological Survey and Formal Analysis. Horn, Laurence, New York.