Multi-dimensional Information Coding in Speech

Yi Xu

Department of Speech, Hearing and Phonetic Sciences
University College London
yi.xu@ucl.ac.uk

Abstract

We speak to convey information to the listener, and we listen to decode information carried by the speech signal. How we are able to do so is the ultimate puzzle for speech research. Much of the existing research effort, however, is devoted not directly to this central puzzle, but to various what could be called epiphenomena: speech rhythm, prosodic hierarchy, intonational structure, naturalness of synthetic speech, etc. In this paper I argue that cracking the central puzzle of speech coding is not only the ultimate call for us as speech scientists, but also the key to understanding various epiphenomena in speech. I will demonstrate that speech involves multi-dimensional information coding due to the richness of information to be encoded and the complexity of the underlying neuro-physiological and biophysical mechanisms. Understanding this process may lead to better understanding of many of the epiphenomena as well.

1.   Introduction

We humans as curious and intelligent beings are fascinated by many things we observe in nature: Why does the sun rise from the east? Why do apples fall to the ground? Why are flowers so colorful? Why do humans speak? As our understanding of nature improves, we often come to recognize that many of the things that we observe are in fact epiphenomena of certain fundamental processes, as they have no reciprocal causal relations to the fundamental processes. For instance, the pleasure we humans experience when admiring wild flowers, as an epiphenomenon, cannot explain why flowers are so colorful (unless we are talking about garden flowers). But how the colors of the flowers are perceived by bees is indeed part of the explanation for the colorfulness, as flowers are likely to have co-evolved with bees and other pollinating insects.

In the case of speech, one of the first things we may have wondered about is, given that when we speak, we apparently pass information to each other, how are we able to do that? Also, what is it that is actually passed on from the speaker to the listener? These are not easy questions, of course, and we are still struggling with them. At the same time, as we try to answer the central questions like these, many interesting phenomena catch our attention. Frequently, an effort to explain one of these phenomena takes on a life of its own. In this paper, I would like to suggest that when working on a particular area of speech, it helps to never lose sight of the central issues, and to frequently ask questions like, is the phenomenon part of the central mechanisms of speech communication, or is it just an epiphenomenon that has no reciprocal causal relation to the main mechanism? It could be argued that, epiphenomena or not, developing a good understanding of all the observed patterns would ultimately contribute to the understanding of speech as a whole. What I will argue, however, is that treating an epiphenomenon as if it stands on its own may not be the best research strategy. Instead, understanding the core mechanisms of information coding in speech can not only help us address the central puzzles, but also improve our understanding of various epiphenomena.

2.   What might be the core mechanisms of information coding in speech?

It is quite firmly established by now that vowels, consonants and tones all have acoustic patterns that make them distinct from each other when produced in isolation or in isolated words [2, 9, 47]. What this means is that at the most rudimentary level, information is encoded in speech by associating function-specific categories with distinct patterns. But speech is mostly made up of connected utterances that require rapid shifts from one distinctive pattern to another. Thus one of the basic questions about connected speech is, are phonemes distinguished from each other in connected speech by distinct patterns similar to those said in isolation? To understand the issue, we may start from the Morse code, which encodes discrete symbolic information with a set of distinct long and short pulses, separated by distinct lengths of pauses (Fig. 1a). To make it more like speech, we could replace each dash-dot combination with a tone of a specific frequency, and remove the pauses in between, as shown in Fig. 1b. But here comes the critical problem. The articulatory system that produces speech is a biophysical device whose state can be changed only sluggishly [60]. Conceivably, there could be many different solutions to the problem. One of the simplest is to treat each distinct static tone as a target and to reach them one by one. What we will get, then, is a continuous output like the solid curve in Fig. 1c, where the target tones are shown as the dashed lines.
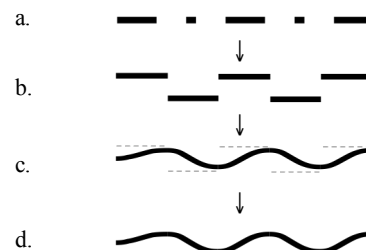

Figure 1: *From pseudo Morse code to continuous surface curves. See text for explanation (from [78]).*

Now, if the dashed lines in Fig. 1c are removed, we are left with only the solid curve in Fig. 1d, namely, the "surface" signal. Several problems arise immediately for our understanding of the signal. First, there are no obvious unit boundaries in the surface form, which simply keeps changing smoothly. Second, no matter where we imagine the boundaries

are, or even if we happen to know the real boundaries, no part of the signal seems to be exclusively attributable to a single static element. Thus it is easy to conclude from looking at Fig. 1d alone that neither discrete nor invariant units exist in the signal.

Suppose we know at least the identities of the coded elements and are able to manipulate them, say by making the third element identical to the two adjacent ones. We would then get the thin curve in Fig. 2a. Overlaying Fig. 2a with Fig. 1c we would get Fig. 2b, from which we could see that, a) the difference in the middle part of Fig. 2b is only due to the third element, b) the third element has extensive influence on the portion of the curve corresponding to the fourth element, but c) it has no influence on any of the preceding elements.
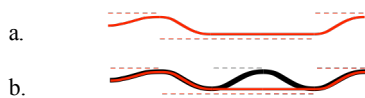

Figure 2: *a. Same as Fig. 1c except that the 3rd element is now identical to the surrounding elements. b. Overlay of a. and Fig. 1c.*

Interestingly, while this imaginary scenario may seem simplistic, it is very close to what has been found for lexical tones. Fig. 3 shows that in Mandarin, the tone of the second or third syllable in the 5-syllable utterances recorded in [75] has little influence on the preceding tone(s) but extensive influence on the following tone. Despite the influence, the F_0 curves of the third syllable in Fig. 3a gradually converge to a falling slope appropriate for the F tone. Likewise, the F_0 curves of the fourth syllable in Fig. 3b converge to a high-level shape appropriate for the H tone. Such convergence reveals a coding mechanism not unlike that seen in Fig. 1c. Several characteristics of the coding mechanism in Fig. 1-2 are worth noting:
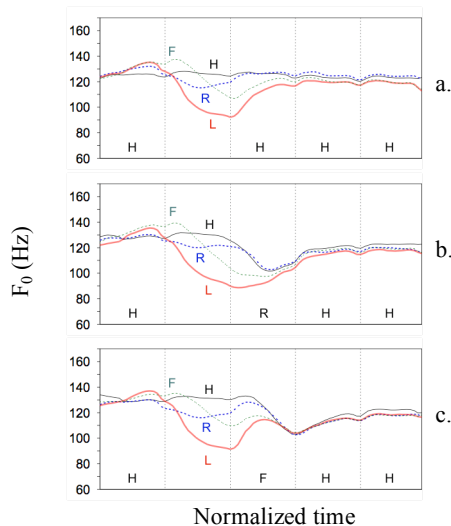

Figure 3: *Mean F_0 contours of Mandarin five-syllable utterances. H, R, L and F stand for High, Rising, Low and Falling tones, respectively. Adapted from [75].*

1. Unidirectionality — The surface signal is always moving monotonically toward one desired target or another.

2. Syllable-synchronization — The unidirectional movement largely coincide with the syllable to which the tonal target is associated

3. No anticipatory execution — The movement toward a target does not start until the movement toward the preceding one is over.[1]

4. No return to rest position — No portion of the curve is for the sake of returning to a non-target rest position after approach a target.

To capture this coding mechanism we have proposed the Target Approximation model (TA) [85], as shown in Fig. 4. According to TA each tone is associated with a distinct underlying target as indicated by the dashed lines in Fig. 4. The articulatory system tries to reach the targets one at a time, resulting in a smooth surface signal that asymptotically and successively approaches the targets, as indicated by the solid curve.
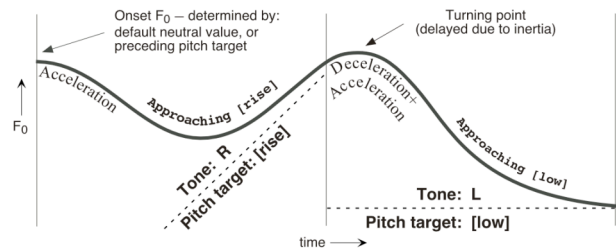

Figure 4: *Illustration of the TA model. The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from asymptotic approximation of the pitch targets.*

Note that TA shares some similarities with two other models, the Command Response model (CR) for intonation [17], and the Task Dynamic model (TD) for vowels and consonants [54]. Both of them, similar to TA, assume that surface signals result from asymptotic movements toward underlying goals. But TA differs from CR and TD in a number of nontrivial ways. First, TA assumes that all movements unidirectionally approach one target or another, with no obligatory return phases to a base line as assumed in CR, or optional return phases to a rest position as assumed in TD. Second, TA assumes full state transfer at the boundary between two targets, which includes the transfer of displacement, velocity and acceleration. CR and TD explicitly assume only the transfer of displacement across the boundaries. Third, TA assumes that underlying targets can be either static or dynamic. In Fig. 4, for example, the first target is a dynamic [rise], whose approximation results in a high velocity that is transferred across the boundary, causing the turning point to occur in the temporal interval of the second target. Neither CR nor TD have explicit assumptions about dynamic targets. Fourth, TA assumes synchronization of tonal targets with the syllable, so that each target-achieving articulatory effort starts at the syllable onset and ends at the syllable offset. CR has no internal assumptions about such synchronization. TD so far

---

[1] A tone in fact exerts some anticipatory effect on the preceding tone. But the effect is dissimilatory rather than assimilatory, as found in languages for which ansitipatory effects are systematically investigated. See summary in [75].

has had only limited concerns with laryngeal articulation [37]. Fifth, TA assumes that targets and their temporal intervals are separately controlled for information coding and are thus independent of each other. Such independence is not explicitly assumed in either CR or TD.

That there is independent control of targets and their temporal intervals is an important assumption of TA. In fact, independent control is assumed for other aspects of the target approximation process as well, as will be discussed next.

3. Expanded core mechanism for multi-dimensional information coding

Given TA as sketched in Fig. 4, it is not difficult to imagine that various aspects of the process can be differentially specified. These could include 1) target, 2) strength (with which a target is approached), 3) range (within which a target is approached), and 4) duration (or temporal domain of target approximation). Modification of any of these aspects may have an impact on the output signal of the system. Fig. 5 shows an illustration of the impacts simulated by a recent quantitative implementation of TA [51].
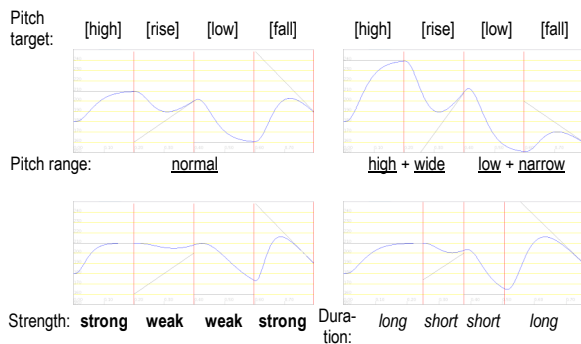

Figure 5: *Illustration of effects of pitch targets, pitch range, strength and duration, simulated by qTA [51].*

The impact of target can be seen in all four panels of Fig. 5. The targets are [high], [rise], [low] and [fall], of which the first and third are static with different heights, and the second and fourth are dynamic with opposite slopes. The asymptotic approximation of this target sequence produces similar up-and-down patterns in all panels. (The initial value is arbitrarily set at a middle level in all the plots.) In the qTA implementation, a target is specified by two parameters: height (y-intercept) and slope. In the upper left panel, all the other parameters are assumed to have normal values and the output signal there can therefore serve as a reference.

The effects of range adjustments can be seen in the upper right panel. The range for the first two targets is high and wide, while that for the second two targets is low and narrow. Note that the range adjustments are applied through changes of the height and slope of the underlying targets rather than the surface range.[2] As a result, the local shape of the contours remain the same but the movement magnitudes are changed.

The impact of strength is simulated in the lower left panel, where the strength for the first and last targets is strong while that for the middle two targets is weak. As a consequence, the surface curves in the first and last target domains actually reach the targets, but those in the middle domains fall far short of their targets, resulting in severe *undershoot*: the surface slope of the [rise] target is much shallower than the desired value even in the final portion of the temporal domain; and the surface minimum of the [low] target is much higher than the desired low value.

The effects of duration can be seen in the lower right panel. There the temporal domains of the first and last targets are long while those of the middle targets are short. Note that the impact of duration is similar to that of strength: long duration leads to better target realizations while short duration leads to greater undershoot.

When the effects of all the TA parameters are combined, the resulting surface signal can be quite complicated. Nevertheless, because their manipulations are all applied to the core mechanism of target approximation, the effects of the parameters are predictable and likely recoverable in perception. As I will discuss next, the TA parameters can be effectively used as encoding elements in transmitting multiple communicative meanings.

4. Prosodic speech as multi-dimensional information carrier

The above discussion has focused on the issue of how various aspects of the target approximation process can be separately controlled for information coding. Just as important, however, is why such multi-dimensional control is necessary. As has been found over the past decades, lexical identity, which has often been considered as the core of phonetic coding, constitutes only one layer of information to be transmitted by the speech signal, albeit a very important one. In the following, I will use our recent findings about English intonation to illustrate how the multiple layers of information may demand of the TA process.

First, English is known as a non-tonal language. However, what this means is only that there are no lexically determined fixed local pitch targets for individual syllables. As found in recent research on American English, syllables *are* assigned specific pitch targets once the modality of the sentence as well as the location of focus are given. In a statement, unstressed syllables are assigned a neutral pitch target whereas stressed syllables are assigned a [high] pitch target, unless it is word-final and on-focus, in which case it has a [fall] target [86]. In a declarative question, however, all the stressed syllables are given [rise] targets whether it is on-focus, pre-focus or post-focus [36]. Thus lexical stress in American English is partially encoded by syllabic pitch targets, although in a complex way interacting with focus and sentence modality. Second, there is some evidence that, similar to the neutral tone in Mandarin, unstressed syllables are assigned weak strength [10]. Third, similar to Mandarin [75], focus is encoded in American English by a tri-zone pitch range control — expanding pitch range of the on-focus syllable, compressing the pitch range of the post-focus syllables, and leaving the pitch range of pre-focus syllables largely neutral [86]. This is true of both statements [86] and questions [36]. Fourth, also similar to Mandarin, sentence modality is encoded in American English by raising pitch range continually toward the end of the sentence, starting from the focused word [36]. However,

---

[2] This bears the assumption that the adjustment is done before the neural commands are issued to the laryngeal muscles. This is different from CR [17], in which two continuous curves resulting from muscle responses to two streams of neural commands are generated first before being summed up to form surface contours.

unlike in Mandarin [35] where post-focus pitch range is lowered in both statements and questions, post-focus pitch range in American English dramatically increases in a question [36]. Finally, again similar to Mandarin, duration of the focused syllable is increased, whereas that of the rest of syllables in the sentence remain unchanged [86].

Thus all the four TA parameters are involved in the encoding of the three communicative functions in English that are likely to be among the most frequently used: lexical stress, focus and sentence modality.

Such multi-dimensional information coding is captured by the Parallel Encoding and Target Approximation model (PENTA) [76], as shown in Fig. 6. In PENTA the target approximation process (large square box) serves as the basic articulatory encoding mechanism that is controlled by multiple communicative functions (stacked boxes on the far left). These communicative functions are realized through distinct encoding schemes (second stack of boxes from left) that specify the values of the TA parameters (middle block). The parameters then control the TA process to generate surface acoustic output.

Several characteristics of PENTA make it distinct from existing models of speech generation. The first is its explicit representation of the communicative functions as the driving force of the system. This differs from models that assume that the primary driving force of speech comes from formal structures that are not directly defined in terms of communicative meanings [25, 48]. The second is that in PENTA, the communicative functions do not directly specify surface acoustic forms. Instead, they are implemented via encoding schemes that specify TA parameters which in turn control an articulation process. This differs from models that directly specify either the surface acoustic forms [57, 64], or component acoustic forms [4, 68].

PENTA was originally proposed for tone and intonation [76], but recent findings about the similarity between segmental and tonal aspects of speech [80, 81] have made it logical to extend it to other aspects of speech. Most important among these findings are those about timing and coordination in speech, as will be discussed next.

5. Timing and coordination

The most important theoretical basis for the PENTA model is the articulatory-functional view of speech [76]. What is critical to this view is the explication of what is articulatory and what is functional. That is, for any observed phenomenon or proposed mechanism, it is critical to ask, is it due to an articulatory constraint or is it for information coding? In the case of speech timing, it is thus necessary to distinguish between aspects of timing that are obligated by articulatory mechanisms, which can be referred to as *obligatory timing*,

and those that are part of the encoding schemes of communicative functions, which can be referred to as *informational timing* [79].

5.1. Obligatory timing

Speech is produced by manipulating the state of the human articulatory apparatus: changing the location of the articulators, reshaping them, or adjusting their physical properties such as stiffness. These state manipulations are dynamical processes that take time. Part of the timing patterns of these processes are directly determined by the properties of the articulatory system and its neural control mechanisms, and are hence obligatory. There are at least two kinds of obligatory timing: maximum speed of articulatory movements and synchronization of concurrent movements.

5.1.1. Maximum speed of articulation and the near-ceiling performance hypothesis (NCP)

The maximum speed of an articulatory 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 [40, 63]. The maximum speed is positively related to maximum muscle force and movement magnitude [82], but negatively related to precision of movement goal [63]. The importance of the maximum speed for speech depends on how much impact it has on the surface trajectories of the acoustic variation. In could be the case that it is so fast that any target can be reached within a negligible amount of time. But this is apparently not the case for pitch movements [61, 82]. According to [82], the mean minimum duration of a pitch rise or fall is quasi-linearly related to the size of pitch movement that are above 1 semitone, and can be estimated with the following equations:

$$t = 89.6 + 8.7\,d \qquad \text{(raising)} \qquad (1)$$
$$t = 100.4 + 5.8\,d \qquad \text{(lowering)} \qquad (2)$$

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

In Figure 3a, for example, the amount of pitch increase from the end of the L tone to the highest point of the H tone is about 4.2 semitones. From equation (1) it takes at least 126 ms for an average speaker to complete such a movement. Yet the mean duration of that syllable is only 181 ms in the study, which means that the transition would take up most of the syllable duration. Thus, the transitions in Fig. 3a are mostly obligatory, because speakers cannot make the pitch movements much faster.

The constraint of maximum speed of articulation has even greater impact in cases where the targets are dynamic. In Fig. 3c, for example, to approach the [fall] target of the Mandarin F tone after a L tone, F_0 needs to go up before
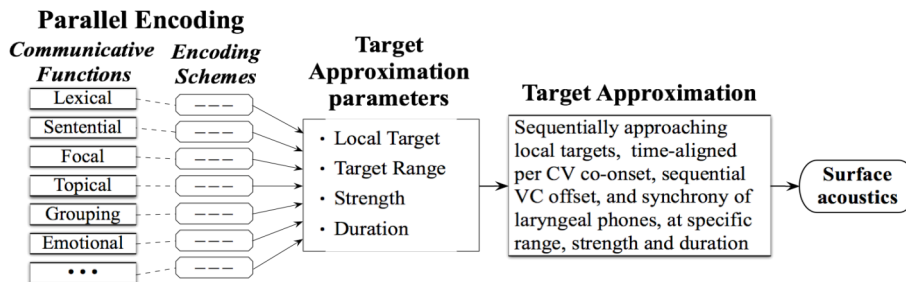

Parallel Encoding

Figure 6: *A schematic sketch of the general PENTA model. Modified from [2005].*

making a sharp fall. Thus two movements need to occur within the same syllable. Calculations with equations (1) and (2) indicate that there is hardly enough time to make the two movements. Furthermore, if we consider the highest point in the HF sequence in Fig. 3c to be the targeted value, undershoot of the F tone has apparently occurred in other sequences. According to calculations with equations (1) and (2), the undershoot is not due to speakers' laziness, but because they do not have much of a choice. Indeed, it is precisely during the dynamic tones in Mandarin that the maximum speed of pitch change is reached [82], indicating that speakers are already trying as hard as possible.

Note that these cases cannot be explained by the *economy of effort* hypothesis [34], according to which speakers often avoid applying full muscle forces in order to conserve energy, and undershoot is the result of doing so. The undershoot here occurs, rather, when full muscle force *is* applied, as indicated by the maximum speed of pitch change. So, undershoot has occurred *despite full articulatory effort*. Meanwhile, there is at least initial evidence that when full or "hyper-" articulation does occur, as in the case of stressed syllables, the articulatory effort as measured by peak velocity [40] is actually less than that during unstressed syllables, and that it is the latter that has approached the real maximum speed as measured from repetitive nonsense syllable strings [77].

These problems put into question the basic assumption of the economy of effort hypothesis, i.e., speakers stay comfortably away from their dynamic articulatory limits, and so can choose to be "economical" unless the demand for intelligibility is high. If even at normal speech rate the maximum speed of articulation is frequently approached, speakers are probably operating near their optimal performance level. So, an alternative to the economy of effort hypothesis is the *near-ceiling performance* hypothesis (NCP), which states that speech is maintained near an overall performance ceiling due to its vital importance for the survival and wellbeing of human individuals, and it is such near-ceiling performance that is responsible for many cases of undershoot [77].

5.1.2. Synchronization of concurrent target approximation movements

In addition to the maximum speed of articulation, there is another articulatory constraint that seems to be just as rigid. That is, there is a strong pressure for concurrent articulatory movements to be synchronized. The initial evidence comes from F_0 patterns of Mandarin tones as discussed earlier. That is, the transitions between adjacent tones take place during the targeted tone itself rather than during a dedicated transition interval [73, 75], as can be seen in Fig. 3 and is captured by TA. Also, as found in Mandarin, Cantonese and English, even in a syllable with a voiceless initial consonant, the pitch target approaching movement starts from the syllable onset rather than from the voice onset [71, 72, 83]. Furthermore, the interval of target approximation is not affected by coda consonants [74]. Thus the execution of the tone-approaching movement coincides, or is synchronized, with the entire syllable. Given that the speed of pitch change is often as fast as possible in a dynamic tone, 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 suggests that such onset timing is likely to be also obligatory, and so cannot be readjusted for the sake of information coding.

5.1.3. The syllable as a time structure

In a recent study in which we tried to determine the temporal intervals of glides and approximants, we found evidence that, just like tones, vowels and consonants are produced with unidirectional target approaching movements [81]. Based on this finding, we have argued that the conventional segmentation of syllables based on acoustic landmarks is problematic, because the onset of the articulatory movement toward an initial consonant is not at the conventional landmarks, such as the onset of frication, onset of nasal or lateral murmur, or onset of stop closure. Rather, the movement starts at the onset of the formant transition toward the consonant. In other words, the formant movements toward a segment should be viewed not as the *anticipation*, but as the *execution* of the segment. Using F_0 alignment as reference, the onset of the transition toward an initial consonant in English and Mandarin is estimated to be about 26-48 ms earlier than the conventional onset of nasal closure [81].
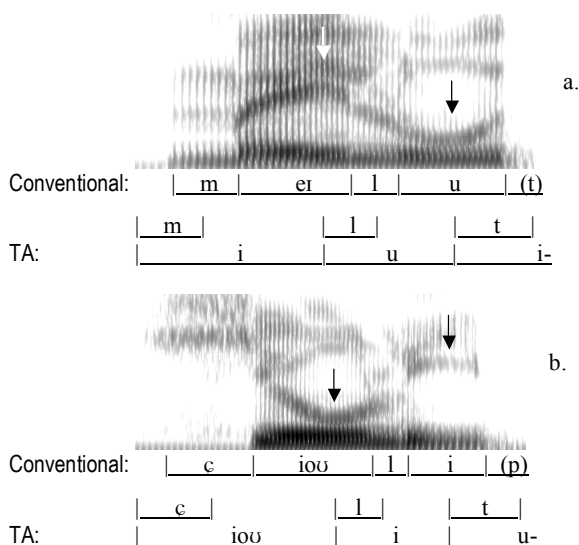
Figure 7: *Spectrograms and conventional vs. TA segmentations of Mandarin [l] + V sequences. a. [meɪ lu (tiɛn χuo)] (to light coal stove). b. [ciou li (bu t̢ʂou)] (repair procedure). The arrows mark F2 turning points.*

An illustration of the new segmentation as compared to the conventional one is shown in Fig. 7. At each arrow F2 changes movement directions. Before each change, F2 moves toward the most characteristic pattern of a segment: [ɪ] and [u] in Fig. 7a, and [ʊ] and [i] in Fig. 7b. Thus each turning point can be viewed as the offset of a previous segment and onset of the next segment. In Fig. 7 it can be also seen that the F2 movements after the first arrows are not only toward [l], but also toward the following vowels. It is downward toward [u] in Fig. 7a but upward toward [i] in Fig. 7b. This is consistent with the classic finding as early as in 1933 [38] that, in a CV syllable, the articulatory movement related to V actually starts at about the same time as that related to C [24, 38, 42]. In fact, the term coarticulation (originally in German) was coined to refer to this phenomenon [38].

The findings just discussed have led to the *time structure model* of the syllable [80], according to which the syllable serves as a framework that assigns the temporal intervals of

consonants, vowels, tones and phonation registers, as illustrated in Fig. 8. The temporal alignments are 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 consonant, and c) synchrony of tone and phonation register with the entire syllable.
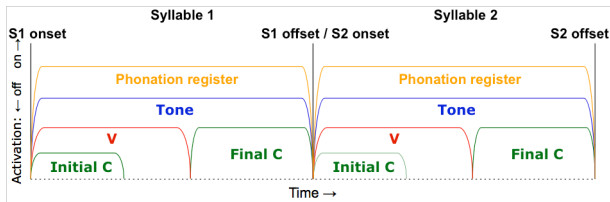

Figure 8: The time structure model of the syllable. Adapted from [80].

The time structure model itself may not represent the most fundamental mechanism of speech timing at the micro level. The timing characteristics of the syllable have been explained in terms of entrainment [18, 23]. As I have pointed out in [79], however, entrainment cannot account for timing stability of syllabic components in monosyllabic utterances, since it take many cycles for two independent oscillating systems to become synchronized [59]. Furthermore, the *co-onset* principle guarantees only the synchronization of the onset of initial consonant with the rest of the syllabic components. The consonantal offsets occur at a non-fixed time during the other movements, relatively early when the syllable is long, but late when the syllable is short.

5.1.4.   Syllable onsets as time markers

To probe deeper for the mechanism that underlies the behavior of the syllable, I have proposed, very tentatively, that syllable onsets probably serve as time markers in speech [79]. A time marker is an event, such as the tick of a clock, that serves as a reference for the measurement of time and timing [22]. To control timing in speech production and to detect timing in speech perception, a reference system is needed. What is desirable are recurrent 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 where the temporal components are not fixed, such as music [22, 26]. Syllable onsets, where the unidirectional movements toward the initial C, the first V, the tone, and possibly the phonation register, all start simultaneously, seem to serve this purpose well. According to this hypothesis, which has yet to be fully tested (but see [43] for initial evidence), it is the need to have recurrent and unambiguous events to serve as time markers that gives rise to the temporal organization of the syllable as captured by the time structure model.

5.2. Informational timing

With the lack of freedom in micro-controlling the temporal alignment of the target approaching movements toward consonantal and vocalic targets, as discussed above, what is still available for information coding is syllable duration. This is nonetheless a very large control space. And the space is even larger if pause duration is included. Thus there should be

sufficient space to allow the kind of parallel encoding of multiple layers of information seen in the pitch dimension [75, 85], as captured by PENTA [76].

5.2.1.   Lexical contrast

First, duration is used by so-called quantity languages to directly distinguish words. In these languages, vowels often carry a two-way (or three-way e.g., Estonian [66]) duration contrast. A common characteristic shared by these quantity contrasts is that the duration ratio between the short and long vowels is rather large, as can be seen in the following.

Duration ratio of long vs. short vowels:

Thai:	**2.0 : 1** [3]
Japanese:	**2.5 : 1** [20]
Finnish:	**2.5 : 1** [62]
Icelandic:	**1.95 : 1** [50]

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 [12]. In Mandarin, though there is no equivalent of word stress, the neutral tone bears some similarity to the English weak stress [10, 85], and the full tone to neutral tone duration ratio is about 1.7:1 [10, 33].

5.2.2.   Focus

Duration is also known to participate in making focal contrast. Focus has been consistently found to lengthen the lexical item being focused [67, 75, 86]. However, the ratio of focused to non-focused duration is generally much lower than that for lexical stress, 1.17:1 in Mandarin [75], 1.25:1 [67] or 1.14:1 [86] in English, and 1.09:1 in Dutch [58]. The reason for these relatively low ratios is probably because duration is not the predominant cue for focus. 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.

5.2.3.   Boundary marking and affinity indexing

A durational phenomenon that has been long noted is final lengthening, i.e., the last syllable of a phrase or sentence is much longer than the preceding syllables [28, 29]. This has been recognized as a cue for sentence or phrases boundaries [28, 29]. In addition, more subtle durational changes at smaller boundaries have also been found to have an effect of disambiguating ambiguous syntactic structures [28]. Such disambiguation is in essence done by marking the relative strengths of individual word boundaries [69]. It has been further demonstrated that the duration difference related to boundary strength is gradient rather than categorical [8, 11]. Boundary related durational changes have also been found in Mandarin in cases where no word stress is involved [84], which indicates that this kind of duration control is independent of stress.

In addition to final lengthening, pauses are also known to mark boundaries with even stronger strengths [27, 28, 44]. Interestingly, there is something in common between lengthening and pausing, i.e., both affect the distance between the onset of the pre-boundary syllable and the onset of the post-boundary syllable. Thus according to the time marker hypothesis, the amount of lengthening plus the length of the pause would directly determine the temporal distance between

the two adjacent syllables. In this way, the temporal distance is used *iconically* to encode relational distance of adjacent linguistic constituents. In other words, the inter-onset interval (IOI), which consists of both the duration of the pre-boundary syllable and the duration of the optional pause, serves as an *affinity index* that signals how closely two adjacent constituents are relationally associated with each other.

6. Epiphenomena and their possible articulatory-functional explanations

Now that the possible basic mechanisms of conveying information through multi-dimensional coding have been briefly outlined, we can take a fresh look at some of the well known phenomena to see if they are part of the central mechanisms discussed so far or byproducts of the central mechanisms.

6.1. The rhythm class hypothesis

In his 1945 book Pike noticed that languages like English and German seem to be spoken with a morse-code-like strong-weak rhythm, whereas languages like French and Spanish are spoken at a machinegun-like fast rate [49]. This initial observation has since evolved into a special field of research known as speech rhythm, centered largely around the rhythm class hypothesis, according to which there is a universal tendency for certain units to become equal in duration, and that languages of the world are divided into three rhythm classes depending on the kind of unit involved in manifesting the isochrony tendency: stress-timed, syllable-timed and mora-timed [1, 6, 49]. Later empirical research has shown, however, that no true isochrony can be found [19, 30, 39, 70]. Nevertheless, a weak tendency toward isochrony has been demonstrated at least for stress-timing [19, 21]. A more recent development in rhythm research is the proposal of the hypothesis that rhythmic patterns help infants to distinguish between languages in a multi-lingual environment [52], and it is thus the functional pressure of language acquisition that forces languages to evolve into pre-existing rhythm classes.

A paradoxical question arises from this hypothesis, however. If rhythm class exists to help infant distinguish languages, it must be the case that a) infants in bilingual environments have greater difficulty acquiring two languages of the same rhythm class than different classes (which is yet to be demonstrated), and b) as a result, two languages closely in contact with each other and so are spoken by many bilinguals would tend to diverge in their rhythm tendencies. This is apparently against known trend in language contact: the more closely two languages are in contact with each other, the more similar they will become over generations. So, the ease of infant discrimination of ambient languages is unlikely to have been the reason why languages differ or resemble each other in terms of rhythmical characteristics in the first place. Rather, infants' perceptual behavior in the laboratory is probably an epiphenomenon of the more basic facts, namely, languages differ in their phonologies, which may lead to temporal characteristics that are perceptually salient and acoustically measurable, as is suggested in [52]. So, unless the lab-observed infant behavior is shown to have a reciprocal causal relation to the temporal differences across languages, it cannot provide support for the rhythm class hypothesis.

Assuming that the reported rhythmic tendency does exist [19, 21], from an articulatory-functional perspective, we may still ask, is it an articulatory mechanism, or does it serve some

kind of communicative functions? In the preceding discussion we have seen that first, much of the obligatory timing can be explained by articulatory mechanisms that are not rhythmic in nature. Secondly, much of the durational control is done for the sake of encoding rather specific information, including lexical contrast, focus, and affinity indexing (via an iconic use of inter-onset interval). The last one is especially relevant if rhythm is about isochrony of some kind. In English, for example, the stressed syllable in a trochaic word is necessarily much shortened when compared to a word-final stressed syllable, because it is followed by an unstressed syllable that is by definition closely related to it. Such shortening would tend to even out the stress-to-stress intervals, making the language sound like stress-timed. Of course this is just one of the possible explanations. In general, although a rhythmic explanation of speech tempo cannot be totally ruled out, its necessity is compelling only when the explanatory power of the obligatory timing and informational timing discussed earlier is shown to be inadequate.

6.2. The prosodic hierarchy hypothesis

It is a widespread idea that there exists a prosodic structure in speech that consists of a hierarchy of constituents of different sizes: intonational phrase, prosodic phrases, prosodic words, clitic groups, metrical feet, etc. [5, 31, 32, 55, 56]. This prosodic hierarchy hypothesis conceptually overlaps with the rhythm class hypothesis, because the definitions of the smallest constituents, i.e., prosodic words, clitic groups and metrical feet, all refer to word stress, and word stress is what is supposed to recur at near even time intervals in a stress-timed language according to the rhythm class hypothesis. Thus an obvious question that has seldom been asked is, do the two types of theories offer alternative or complementary explanations of the commonly observed patterns?

Of course, yet another possibility is that both prosodic hierarchy and rhythm are epiphenomena of the central mechanisms of speech. As discussed in the previous section about rhythm, similar questions can be posed to the prosodic hierarchy hypothesis: Is it obligated by an articulatory process or does it serve a communicative function? Many of the existing arguments for the existence of a prosody hierarchy, however, assumes that it is an autonomous structure that stands on its own, and syntactic relations has to be parsed by this structure so as to be manifested in the phonetic implementation [5, 25, 56]. This has been argued to be especially true for marking the boundaries of the prosodic constituents whose numbers are fixed [5, 56]. But recently, it has been shown that recursive syntactic relations are directly reflected in gradient durational differences rather than only in terms of categorical boundary signals [69]. Thus as far as duration is concerned, there may not be categorical markings for the constituents of the hypothesized prosodic hierarchy.

Another argument for the existence of a prosodic hierarchy is that it is needed to assign prominence levels to the constituents of a sentence [5, 25, 56]. The problem with this argument is obvious from a functional point of view. That is, the functional sources of at least two of the prominence-related contrasts are clear: Lexical stress serves to distinguish words; and focus serves to mark pragmatic emphasis. Thus the two are functionally independent of each other, and neither dominates the other. There does seem to be a functional conflict between the two, however. The contrast of lexical stress in a language like English always occurs

between adjacent syllables and mostly within a word. It therefore requires only a small pitch difference, as found in both production and perception studies [16, 86]. But given the omnipresence of lexical stress, focus has to be encoded with a much larger prominence boost so as to be clearly different from stress. Furthermore, focus typically has a much larger operational domain than lexical stress, involving multiple words rather than just two adjacent words. Large acoustic differences are therefore needed to manifest a clear focal contrast. Thus both the occurrence and prominence levels of lexical stress and focus have plausible functional explanations. Is there still a need, then, for a prosodic hierarchy to assign prominence levels to them?

Overall, it seems that prosodic hierarchy, just like speech rhythm, is likely an epiphenomenon derived from a number of basic articulatory and functional mechanisms.

6.3. Intonational structure

The idea that there exists an intonational structure goes back to 1922, if not earlier. According to Palmer [46] English intonation consists of an obligatory nucleus and optional head, pre-head and tail. This tradition has continued even today in the theoretical framework of intonational phonology [25, 48]. That is, although the intonation nucleus is no longer treated as being special, the idea that intonation stands on its own as a structure that guides production and perception is still the essence of main intonation theories.

From an articulatory-functional perspective, again, a natural question about this assumed structure would be, is it due to articulatory mechanisms or communicative functions? Indeed the definition of the nucleus alludes to the fact that is usually corresponds to the emphasis of the sentence [41]. But the obligatory occurrence of such an emphasis in each and every sentence is puzzling from a functional point of view. As is stipulated by the intonation structure hypothesis, if a sentence does not contain a particular emphasis, by default the nucleus occurs at the sentence final position. But is such a nucleus different from final focus? If it is, as has been found in both production and perception studies [35, 53, 75, 86], what is the communicative function of such a default nucleus?

Still, a counter question could be asked: Why is it that a sentence with no narrow focus often sounds as if the final word is focused? A possible functional answer can be found in the findings of existing focus studies. That is, although a sentence with final focus is both acoustically and perceptual different from a sentence with no focus, the contrast is much less effective than between a non-final focus and no focus [7, 35, 53, 75, 86]. That is, a final focus is easily confused with no focus, and vise versa. As a result, when forced to give a structural answer, i.e., to identify focus from a no-focus sentence, one would most likely hear it in the sentence-final location. Thus the obligatory nucleus of a sentence is probably an epiphenomenon of the way final focus is encoded.

More interestingly, there is also a possible functional explanation for why final focus is not effectively encoded. As found by many studies, questions are produced with dramatic raising of final F_0 [15, 35, 36, 45, 65]. But that pattern would conflict with the F_0 raising by a final focus if the latter were as dramatic as that of a non-final focus. Indeed, questions and final focus are often confused with each other in perception by Mandarin listeners [35]. Thus it is likely that a *functional conflict* with question intonation has led to a compromised

final focus, which in turn has led to the perceptual illusion of an obligatory final focus as the default intonational nucleus.

Finally, as I have argued elsewhere [76], the notion of pitch accent as assumed in the main structuralist approaches is likely a confound between lexical stress and focus. The evidence comes from both production and perception findings. In production, word-stress-related F_0 patterns are found to be present in post-focus regions in English [36, 86], French [13], and Neopolitan Italian [14]. In perception, F_0 difference as small as 5 Hz is found to be sufficient for distinguishing stressed and unstressed syllables in English [16], which is consistent with the magnitude of stress-related F_0 difference found in production [86]. Overall, then, intonational structure, while seemingly obvious from a structuralist point of view, is likely an epiphenomenon of the encoding schemes of individual communicative functions.

6.4. Naturalness in synthetic speech

Unlike the notions discussed so far, naturalness is not really a theoretical hypothesis. But it is nevertheless a very important concept in speech research, as it is something well-sought after in speech technology. From an articulatory-functional view, however, some interesting questions may be asked about naturalness. For example, why is natural speech natural? Is it because people speak so as to sound natural? Apparently not. So, the unnaturalness in synthetic speech is probably not because we are unsuccessful in simulating speakers' effort to sound natural, but because we have not yet adequately modeled the articulatory encoding of communicative functions. So, even if we continue to treat naturalness as a desirable goal, the best way to achieve it, I would argue, is to improve the modeling of both the articulatory mechanisms and the encoding process of individual communicative functions.

7. Conclusions

Speech is first and foremost a communication system. A central question about speech is therefore how information is coded and transmitted in such a system. Much of the current research effort, however, has been spent on issues that do not seem to be directly related to this central question. In this paper I have argued that it is critical to address the central question even if our main interest is in certain non-central issues. I have demonstrated that *target approximation* is likely the core encoding mechanism in speech, and that not only the targets themselves, but also other aspects of the target approximation process, including range, strength and duration, can be separately controlled for information coding purposes. Such multi-dimensional coding strategy, as modeled by PENTA [76], makes it possible for multiple layers of communicative meanings to be encoded through the articulation process. In light of this articulatory-functional view of speech, some of the popular issues in speech research, such as speech rhythm, prosodic hierarchy, intonational structure and naturalness, are likely epiphenomena that can be explained by the multi-dimensional coding process. But as epiphenomena, they are unlikely to be part of the central mechanisms of speech communication.

8. References

[1] Abercrombie, D., 1967. *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

[2] Abramson, A. S., 1978. Static and dynamic acoustic cues in distinctive tones. *Lang. Speech* 21(4): 319-325.

[3] Abramson, A. S.; Ren, N., 1990. Distinctive vowel length: Duration versus spectrum in Thai. *Journal of Phonetics*: 18, 79-92.

[4] Bailly, G.; Holm, B., 2005. SFC: a trainable prosodic model. *Speech Communication* 46: 348-364.

[5] Beckman, M. E., 1996. The parsing of prosody. *Language and Cognitive Processes* 11: 17-67.

[6] Bloch, B., 1950. Studies in colloquial Japanese IV: phonemics. *Language* 26: 86-125.

[7] Botinis, A.; Fourakis, M.; Gawronska, B., 1999. Focus identification in English, Greek and Swedish. In Proceedings of The 14th International Congress of Phonetic Sciences, San Francisco. pp. 1557-1560.

[8] Byrd, D.; Saltzman, E., 1998. Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics* 26: 173-199.

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

[10] 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.

[11] Cho, T.; Keating, P. A., 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics* 29: 155-190.

[12] 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.

[13] Di Cristo, A.; Jankowski, J., 1999. Prosodic organisation and phrasing after focus in French. In Proceedings of The 14th International Congress of Phonetic Sciences, San Francisco. pp. 1565-1568.

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

[15] Eady, S. J.; Cooper, W. E., 1986. Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America* 80: 402-416.

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

[17] 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.

[18] Haken, H.; Kelso, J. A. S.; Bunz, H., 1985. A Theoretical Model of Phase Transitions in Human Hand Movements. *Biological Cybernetics* 51: 347-356.

[19] 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.

[20] Hirata, Y., 2004. Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics* 32: 565-589.

[21] 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 2005, Lisbon, Portugal. pp. 29-32.

[22] Jones, M. R.; Boltz, M., 1989. Dynamic attending and responses to time. *Psychological Review* 96: 459–491.

[23] 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.

[24] Kozhevnikov, V. A.; Chistovich, L. A., 1965. *Speech: Articulation and Perception.* Washington, DC: Joint Publications Research Service.

[25] Ladd, D. R., 1996. *Intonational phonology.* Cambridge: Cambridge University Press.

[26] Large, E. W.; Jones, M. R., 1999. The dynamics of attending: How people track time-varying events. *Psychological Review* 106: 119–159.

[27] Lea, W., 1980. *Trends in Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall.

[28] Lehiste, I., 1973. Phonetic disambigation of syntactic ambiguity. *Glossa* 7: 107–122.

[29] Lehiste, I., 1973. Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America* 54: 1228-1234.

[30] Lehiste, I., 1977. Isochrony reconsidered. *Journal of Phonetics* 5: 253–263.

[31] Liberman, M., 1975. *The intonational system of English.* Ph.D. Dissertation. M.I.T.

[32] Liberman, M.; Prince, A., 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249-336.

[33] Lin, T., 1985. Preliminary experiments on the nature of Mandarin neutral tone [in Chinese]. In *Working Papers in Experimental Phonetics.* T. Lin and L. Wang. (eds.) Beijing: Beijing University Press**:** 1-26.

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

[35] Liu, F.; Xu, Y., 2005. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62: 70-87.

[36] Liu, F.; Xu, Y., 2007. Question intonation as affected by word stress and focus in English. In Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrücken. pp. 1189-1192.

[37] McGowan, R. S.; Saltzman, E. L., 1995. Incorporating aerodynamic and laryngeal components into task dynamics. *Journal of Phonetics* 23: 255-269.

[38] Menzerath, P.; de Lacerda, A., 1933. *Koartikulation, Seuerung und Lautabgrenzung.* Berlin and Bonn: Fred. Dummlers.

[39] Nakatani, L. H.; O'Connor, K. D.; Aston, C. H., 1981. Prosodic aspects of American English speech rhythm. *Phonetica* 38: 84-106.

[40] Nelson, W. L., 1983. Physical principles for economies of skilled movements. *Biological Cybernetics* 46: 135-147.

[41] O'Connor, J. D.; Arnold, G. F., 1961. *Intonation of Colloquial English.* London: Longmans.

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

[43] Olsberg, M.; Xu, Y.; Green, G., 2007. Dependence of tone perception on syllable perception. In Proceedings of Interspeech 2007, Antwerp. pp. 2649-2652.

[44] 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: 217-220.

[45] O'Shaughnessy, D., 1979. Linguistic features in fundamental frequency patterns. *Journal of Phonetics* 7: 119-145.

[46] Palmer, H. E., 1922. English Intonation, with systematic exercises. Cambridge: Heffer.

[47] Peterson, G. E.; Barney, H. L., 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 175-184.

[48] Pierrehumbert, J., 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. dissertation. MIT, Cambridge, MA.

[49] Pike, K. L., 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

[50] Pind, J., 1999. Speech segment durations and quantity in Icelandic. *Journal of the Acoustical Society of America* 106: 1045-1053.

[51] 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. pp. I-749-752.

[52] Ramus, F.; Nesporb, M.; Mehlera, J., 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73: 265-292.

[53] Rump, H. H.; Collier, R., 1996. Focus conditions and the prominence of pitch-accented syllables. *Language and Speech* 39: 1-17.

[54] Saltzman, E. L.; Munhall, K. G., 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology* 1: 333-382.

[55] Selkirk, E., 1984. Phonology and syntax: the relation between sound and structure. Cambridge, Mass: MIT Press.

[56] Shattuck-Hufnagel, S.; Turk, A. E., 1996. A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research* 25(2): 193-247.

[57] Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pierrehumbert, J.; Hirschberg, J., 1992. *ToBI: A standard for labeling English prosody*. In *Proceedings of The 1992 International Conference on Spoken Language Processing*, Banff. pp. 867-870.

[58] 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.

[59] Spoor, P. S.; Swift, G. W., 2000. The Huygens entrainment phenomenon and thermoacoustic engines. *Journal of the Acoustical Society of America* 108: 588-599.

[60] Stevens, K. N., 1998. *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

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

[62] 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.

[63] Tanaka, H.; Krakauer, J. W.; Qian, N., 2006. An Optimization Principle for Determining Movement Duration. *Journal of Neurophysiology* 95: 3875–3886.

[64] Taylor, P., 2000. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America* 107: 1697-1714.

[65] Thorsen, N., 1978. An acoustical investigation of Danish intonation. *Journal of Phonetics* 6: 151-175.

[66] Traunmüller, H.; Krull, D., 2003. The Effect of Local Speaking Rate on the Perception of Quantity in Estonian. *Phonetica* 60: 187-207.

[67] Turk, A. E.; Shattuck-Hufnagel, S., 2000. Word-boundary-related duration patterns in English. *Journal of Phonetics* 28: 397-440.

[68] van Santen, J.; Kain, A.; Klabbers, E.; Mishra, T., 2005. Synthesis of prosody using multi-level unit sequences. *Speech Communication* 46: 365-375.

[69] Wagner, M. 2005. *Prosody and Recursion*. Ph.D. Dissertation. Massachusetts Institute of Techonology.

[70] Warner, N.; Arai, T., 2001. Japanese Mora-Timing: A Review. *Phonetica* 58: 1-25.

[71] Wong, Y. W.; Xu, Y., 2007. Consonantal perturbation of f0 contours of Cantonese tones. In Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrucken. pp. 1293-1296.

[72] Xu, C. X.; Xu, Y., 2003. Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association* 33: 165-181.

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

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

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

[76] Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46: 220-251.

[77] Xu, Y., 2007. How often is maximum speed of articulation approached in speech? *Journal of the Acoustical Society of America* 121, Pt. 2: 3199-3140.

[78] Xu, Y., 2007. Speech as articulatory encoding of communicative functions. In Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrucken. pp. 25-30.

[79] Xu, Y., in press. Timing and coordination in tone and intonation -- An articulatory-functional perspective. To appear in *Lingua*.

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

[81] Xu, Y.; Liu, F., 2007. Determining the temporal interval of segments with the help of F0 contours. *Journal of Phonetics* 35: 398-420.

[82] 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.

[83] 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.

[84] Xu, Y.; Wang, M., 2005. Tonal and durational variations as phonetic coding for syllable grouping. *Journal of the Acoustical Society of America* 117: 2573.

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

[86] Xu, Y.; Xu, C. X., 2005. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* 33: 159-197.