Syllable as a synchronization mechanism that makes human speech possible

Yi Xu

University College London, UK

Department of Speech, Hearing and Phonetic Sciences, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom, Email: yi.xu@ucl.ac.uk

Abstract:

Speech is a communication system that transmits information by using aerodynamics of articulatory movements to generate sequences of alternating sounds. For this system to work, the articulators need to shift their state frequently and rapidly in order to encode as much information as possible in a given period of time, and the many articulators involved need to be coordinated so as to make the central control of their rapid movements possible. These demands are met by the syllable, a temporal coordination mechanism whose main function is to synchronize multiple articulatory movements so as to make speaking possible. This coordination involves three basic mechanisms: *target approximation*, *edge synchronization* and *tactile anchoring*. With target approximation, articulatory movements successively approach non-overlapping underlying targets. With edge synchronization, multiple articulatory movements are aligned by their onsets and offsets. And with tactile anchoring, contact sensations during segments with relatively tight closure provide alignment references for the synchronization. This synchronization mechanism is the source of many speech phenomena, from segmental coarticulation to tonal alignment, and from normal adult production to child acquisition and speech disorders. It is also what links speech to motor movements in general.

Keywords: syllable, target approximation, edge synchronization, tactile anchoring, degrees of freedom, temporal coordination

1. Introduction

That the syllable is an important unit of speech may seem obvious. Most, if not all, early writing systems (Sumerian, Linear B, Akkadian cuneiform, Chinese, Mayan, etc.) started as syllabaries, in which the written symbols represent syllables (or sometimes morae) rather than consonants and vowels (DeFrancis, 1989; Gnanadesikan, 2010; Liberman et al., 1974). It is also much easier for anyone, including non-experts, to count the number of syllables in a word than the number of segments in a syllable (Fox & Routh, 1975; Liberman et al., 1974; Shattuck-Hufnagel, 2010). But as pointed out by Ladefoged (1982:220): "Although nearly everyone can identify syllables, almost nobody can define them." Indeed, clear and quantifiable evidence for the syllable has been hard to come by. After examining eight lines of traditional evidence in support of the syllable as a representation unit in speech production, Shattuck-Hufnagel (2010) found none of them unequivocal. The lack of clear evidence has led to skepticisms about the existence (Kohler, 1966) or universality (Labrune 2012) of the syllable. It is virtually ignored in the early work of generative phonology and the word *syllable* does not even appear as a separate entry in the subject index to the classic work of Chomsky and Halle (1968) on English phonology. Similar reservations have been expressed by Gimson (1970), Steriade (1999), and Blevins (2003). Some theories treat the syllable as by-product of other more basic units, such as gestures (Browman & Goldstein, 1989). Critically, some of the most fundamental issues about the syllable have remained unresolved:

1. Why are there syllables?
2. Are there clear phonetic boundaries between syllables?
3. Where should each segment belong in a syllable: onset, offset or both (ambisyllabic)?

The following is a brief review of how well each of these questions has been addressed in the literature.

1.1 Why are there syllables?

There is little doubt that the syllable plays many important roles in speech. It is the unit that carries stress and accent (Bolinger, 1961; de Jong, 2004; Pierrehumbert, 1980), rhythm (Barbosa & Bailly, 1994; Cummins & Port, 1998; Nolan & Asu, 2009) and tone (Abramson, 1978; Chao, 1968). It is the domain of applying many phonological rules (Blevins, 2001; Hooper, 1972). It is also critical for the perceptual segmentation of the speech signal (Bertoncini & Mehler, 1981; Content, Kearns & Frauenfelder, 2001; Cutler et al., 1986). These roles, however, entails that syllables are already in the speech signal, which does not tell us why they are there in the first place. Some theories take syllable as the basic unit of speech, e.g., Stetson's motor phonetics (Stetson, 1951) and Fujimura's (1994) C/D model. But they have offered no explicit proposal as to why syllables are obligatory at the articulatory level. In MacNeilage's (1998) frame/content theory, the syllable is suggested to have evolved from the oscillation of the jaw in such movements as chewing, sucking and licking. However, the ability to oscillate the jaw is shared by virtually all mammals, yet not even our closest relatives, i.e., chimpanzees and gorillas, have developed syllable-based speech (Fitch, 2010; Pinker, 1995). Thus being able to oscillate the jaw does not seem to inevitably lead to an ability to articulate syllables. Something extra must be involved.

It has also been proposed that the syllable is a unit of stored motor programs (Dell, 1988; Levelt, Roelofs & Meyer, 1999). But the proposal is questioned for its inability to explain cases of resyllabification or the lack thereof (Shattuck-Hufnagel, 2010). More importantly, even if stored syllable-sized motor programming is shown to exist, it cannot explain why the unit has to have the form of the syllable. What remains unclear is why the syllable, with its own unique characteristics, is indispensable, i.e., serving a function that is so vital that speech would be impossible without it.

1.2 Are there clear boundaries to the syllable?

Given an utterance like the one shown in Figure 1a, it may seem that some of the syllables are well separated by the alternation of consonants and vowels whose spectral patterns show clear boundaries (Jakobson, Fant & Halle, 1951). However, the syllable boundaries are much less clear-cut in the case of /wei/. Because it begins with a glide /w/, it is hard to determine when the preceding syllable ends and the current one starts. Even more difficult are cases where a word starts with a vowel, as in English words like *artist*, *article*, *articulate*, *arbitrary*. When they are preceded by word ending in a vowel, as in *new artist*, *my article*, *to articulate*, or *fairly arbitrary*, there would be continuous formant movements across the word (hence syllable) boundaries (unless when spoken very carefully so that the syllable would start with a glottal stop). The same problem would be seen in cases of word internal syllable boundaries, like *hiatus*, *appreciate*, *mediocre*, etc., where there should presumably be a syllable boundary between /i/ and the following vowel or diphthong, yet all we can see in the spectrogram in most cases are continuous formants between the preceding and following consonants.

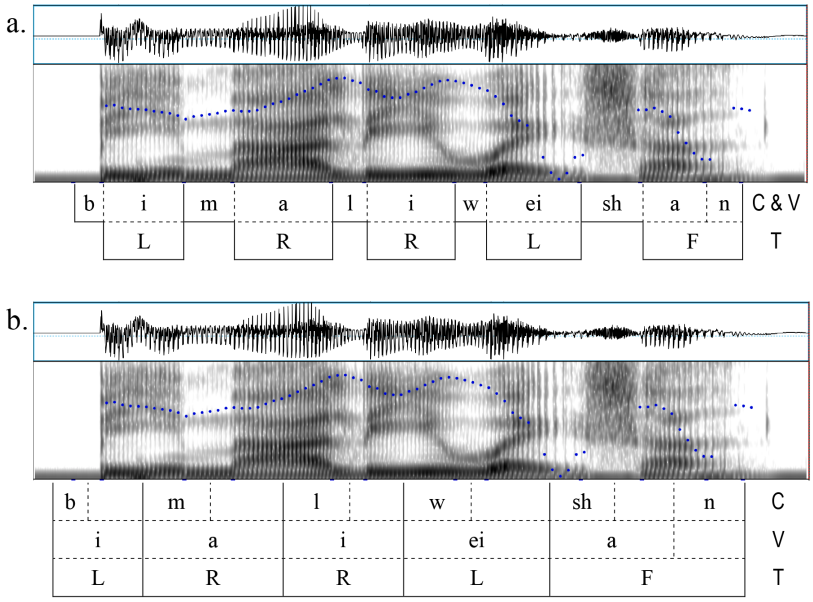

Figure 1. Spectrogram of the Mandarin phrase “比麻黎偽善” [more hypocritical than Ma Li], with broad phonetic transcriptions. In both panels, C, V and T stands for consonant, vowel and tone. In a. the segmentation is conventional (Jakobson et al., 1951; Turk et al., 2006). The segmentation of /w/ is based on Peterson and Lehiste (1960). In b. the segmentation is based on the synchronization hypothesis.

The difficulty of syllable boundary identification has led to the view that it is simply futile to look for clear-cut boundaries in the speech signal, as argued by Hockett (1955), who likens segments as colored raw Easter eggs lined up on a belt. After being crushed by a wringer, the heavy smearing makes the edges of the individual eggs unrecognizable. But boundaries are critical when duration measurements are needed for speech segments (Klatt, 1967; Turk et al., 2006; van Santen & Shih, 2000). Also, if we don't know where the boundaries are, how can we be so sure about the heavy temporal overlap of segments (Kühnert & Nolan, 1999)? More importantly, if speakers do not know when a segment starts and when it ends, how do they articulate them when speaking?

Much of the fussiness of the syllable boundaries, however, could be removed by simply moving them all leftward, more so for the vowels than for the consonants, as shown Figure 1b. As will become clear, the revision of segmentation in Figure 1b, which is based on the theory to be proposed in the subsequent discussion, lie at the core of not only the difficulty of syllable segmentation, but also many other issues related to the syllable.

1.3 Do segments have definitive syllable affiliations?

The identification of syllable boundaries involves not only recognizing the acoustic cues, but also determining the syllabic affiliation of every segment. That is, for each segment, there is a need to decide which syllable it belongs to and where exactly it should belong: onset, offset or ambisyllabic, i.e., belonging to two adjacent syllables at once. There have been many theories of syllabification, including the law of initials and law of finals (Vennemann, 1988), the maximal onset theory (Pulgram, 1970; Steriade, 1982), the theory that stressed syllables are maximized (Hoard, 1971; Wells, 1990) and the weight-stress principle (Fudge, 1969; Selkirk, 1982; Wells, 1990). But so far no consensus has been reached, and syllabification of even some of the simplest cases may have vastly different solutions. For the word *happy*, for example, at least four ways of syllabification are possible according to the summary of Duanmu (2008): /hæ.pi/, /hæp.i/, /hæ<u>p</u>i/ and /hæp.pi/ (where a period stands for syllable boundary and an underscore indicates the segment is ambisyllabic).

The abovementioned syllabification theories, however, are based on the authors' intuition or non-experimental observations. There are also experimental investigations of naïve subjects' syllabification intuitions (Chiosáin, Welby & Espesser, 2012; Content, Kearns & Frauenfelder, 2001; Goslin & Frauenfelder, 2001; Schiller, Meyer & Levelt, 1997). None of these syllabification findings, however, has directly addressed the issue of what syllable boundaries look like in the acoustic signal or in terms of articulatory movements. As it will be shown next, phonetic details are vital for determining syllable boundaries, and boundary marking is directly about the nature of the syllable.

The theory that offers the most explicit articulatory characterization of the syllable so far is articulatory phonology, which regards the syllable as an organizational unit in speech production (Browman & Goldstein, 1992a:165): "[s]yllable-sized organizations are defined by phasing (oral) consonant and vowel gestures with respect to one another. The basic relationship is that initial consonants are coordinated with vowel gesture onset, and final consonants with vowel gesture offset (the specific points being coordinated also differ in the two cases). This results in organizations in which there is substantial temporal overlap

between movements associated with vowel and consonant gestures". However, articulatory phonology focuses mostly on articulatory movements themselves, with only limited description of the acoustics signals they generate.

2. Proposition: Syllable as a synchronization mechanism

The brief review above has shown that the three fundamental questions about the syllable remain unanswered to this day. In the following, I will offer a *synchronization hypothesis* to address all three questions. The overarching proposal is that the syllable is a temporal coordination mechanism whose main function is to synchronize multiple articulatory movements so as to make speaking possible. This coordination involves three basic mechanisms: *target approximation*, *edge synchronization* and *tactile anchoring*.

The basic logic of the synchronization hypothesis is as follows. Speech encodes information by variation of phonetic (segmental, tonal and phonational) properties. The generation of these properties in quick succession during speaking requires multiple articulatory movements toward specific targets (*target approximation*). The central nervous control of such multiple concurrent movements is made possible by the critical reduction of degrees of freedom (DOF) through synchronization of movement onset and offset (*edge synchronization*). And tactile sensation during the closed phase of each syllable provides alignment references for the synchronization of movements (*tactile anchoring*).

The specifics of the timing assignment have mostly been proposed in the time structure model of the syllable (Xu & Liu, 2006), as shown in Figure 2. Here, consonant (C), vowel (V), tone (T) and phonation register (P) are phones defined as a collection of unidirectional articulatory movements towards acoustically defined targets that are either simple or composite, or either static or dynamic (*target approximation*: Figure 3). As shown in Figure 2, the onset of the syllable is where the targets of the first consonant, the vowel, the tone, and the phonation register all start their respective approximations (co-onset). The end of the approximation varies depending on the phone, however. The movement toward the initial C ends the earliest; that toward V ends later, but before the onset of the final C, if there is any (sequential V-C offset). The approximations of tone and phonation register are the slowest, both ending at the offset of the syllable. Finally, at the boundary between two syllables, if there is more than one consonant, the adjacent consonants are split across the boundary, serving as coda and onset, respectively (*tactile anchoring*).

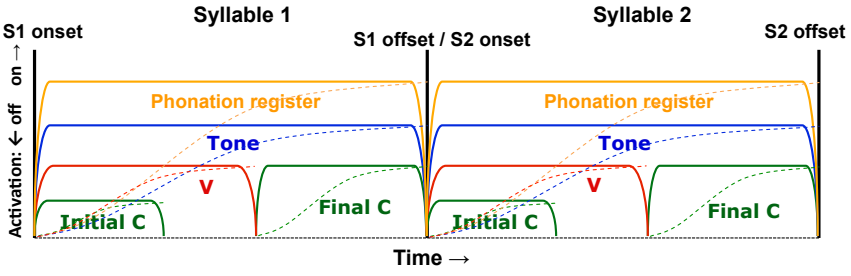

Figure 2. The time structure model of the syllable. Adapted from Xu and Liu (2006).

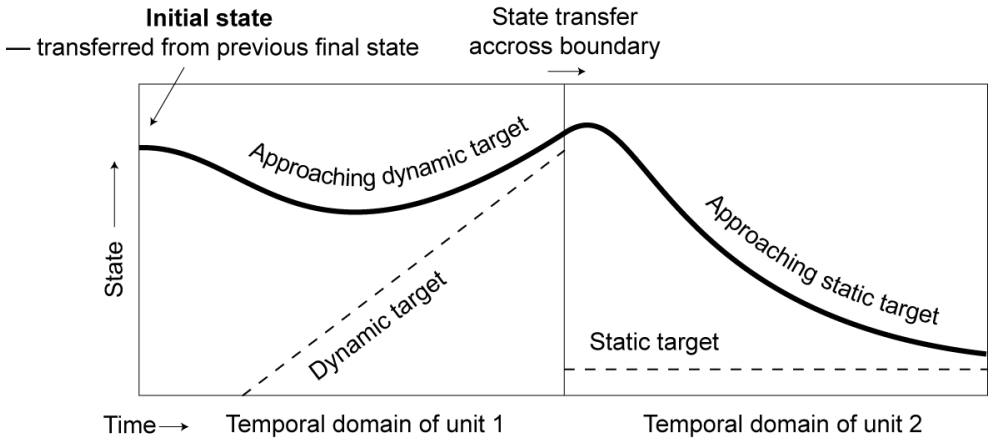

Figure 3. The target approximation (TA) model. A schematic illustration of hypothetical pitch targets (dashed lines) and their surface realization (solid curve). The three vertical lines represent the boundaries of the two consecutive target-approximation intervals. The level dashed line on the right represents a static target, and the oblique dashed line on the left represents a dynamic target. In both intervals, the targets are asymptotically approximated. Adapted from Xu and Wang (2001).

An illustration of how the synchronization hypothesis is applied to the segmentation of acoustic signals of an utterance is shown in Figure 1b, as mentioned earlier. There the *offset* of a segment is where the spectral pattern exhibits its most prototypical configuration (rather than its steady-state portion). For example, for /i/, it is the peak of F2 and F3; for /a/ it is the peak of F1 and valley of F2; and for /w/ it is the valley of F2. For the obstruent consonants, the offset is not at the end of its prototypical spectral pattern (e.g., closure gap in /b/, nasal or lateral formants in /m/ and /l/, and the frication in /sh/), but around the middle of these intervals. The onset of a segment is at a time when the spectral pattern *starts to move toward* its prototypical configuration. For the first /a/, it is in the middle of in the conventional /i/ interval where F2 starts to drop, and for the second /i/, it is in the middle of the conventional /a/ interval where F2 starts to rise. In this new segmentation scheme, the onset of a vowel is fully aligned to the onset of the initial consonant. This has moved each vowel onset leftward from its conventional onset by at least about 100 ms.

In the following, I will first provide a brief elaboration of the synchronization hypothesis, focusing mainly on the similarity and difference between this hypothesis and some major existing models.

2.1 Phones

A phone is redefined as *a collection of unidirectional articulatory movements towards an audible target that encodes a piece of communicative information*. Here the term phone is borrowed from the traditional division between phoneme and phone, with the former referring to all the allophones that make up a phoneme, while the latter to each individual occurrence of an allophone. This usage of the term has become quite rare nowadays, and so it is possible to reuse it with a new enrichment. Consistent with the idea that the notion of phoneme can be

extended to include lexical tone (Abramson, 1978; Clements, 1976; Stagray et al., 1992), phonation (Kreiman & Gerratt, 2010; Esposito, 2010) or even intonational tone (Pierrehumbert, 1980), here phones are extended to include both tonal and phonational targets.

Note, however, that intensity is not included as a phone. This is because short-term (i.e., shorter than a syllable) intensity variation in speech is heavily dependent on other contrastive phones, which makes it hard to be used as an independent short-term dimension to manipulate. Nonetheless, the possibility of having intensity targets does not need to be fully ruled out, as it is ultimately an empirical matter.

2.2 Targets and target approximation (TA)

The notion of target approximation can go back as far as Lindblom (1963), who suggested that underlying phonetic targets are often only partially realized due to time constraint:

"Articulators respond to control signals not in a stepwise fashion but smoothly and fairly slowly, owing to intrinsic physiological constraints. Since the speed of articulatory movement is thus limited, the extent to which articulators reach their target positions depends on the relative timing of the excitation signals. If these signals are far apart in time, the response may become stationary at individual targets. If, on the other hand, instructions occur in close temporal succession, the system may be responding to several signals simultaneously and the result is coarticulation." (p. 1778)

The only major deviation of our notion of target approximation from this view is that ours is strictly sequential. That is, for any articulatory movement, only one target approximation process is executed at any given moment. Those not given enough time to reach their targets are simply *truncated*, i.e., terminated prematurelly, by the commencement of the following movement. Thus there is generally no overlap or simultaneous execution of adjacent target approximation movements by the same articulator (with possible exception of velar consonants, see discussion in 4.5).

Our notion of target approximation (Xu & Wang, 2001), as schematized in Figure 3, was developed based on empirical data on contextual tonal variations (Xu, 1997, 1998, 1999, 2001). In this model, each movement is a process of approaching an underlying target which has both position and velocity specifications, within a designated time interval. Each target approximation movement is therefore controlled by three parameters: target height, target slope (intended velocity) and target strength. Furthermore, adjacent target approximations are contiguous without overlap, shifting from one to the next abruptly at each boundary. The resulting surface contour is nevertheless smooth and continuous, with no clear trace of the underlying boundaries.

The target approximation model, though developed based on tone production, share similarities with a number of other models proposed since Lindblom (1963), in particular, the Fujisaki model for intonation (Fujisaki, 1983), and the task dynamic model for segmental articulation (Saltzman & Munhall, 1989) as the computational implementation of articulatory

phonology (Browman & Goldstein, 1992a). It differs from them, however, in a number of nontrivial ways:

1. Target approximation is strictly sequential, i.e., there is neither overlap of adjacent approximation movements so far as the same articulatory dimension is concerned, nor any gap between target approximation movements, unless there is a genuine pause. The lack of gaps also means that there are no temporal intervals (other than a pause) without specified targets. In contrast, both the task dynamics-articulatory phonology paradigm and the Fujisaki model allow temporal overlap of articulatory gestures or tonal/accentual commands for the same articulatory dimension (Browman & Goldstein, 1992a; Fujisaki, 1983) as well as gaps between gestures or commands.

2. Every target approximation movement is assumed to have an independent strength specification, which determines the rate at which the target is approached. This has allowed the model to account for the F_0 undershoot in both Mandarin neutral tone and English unstressed syllables in terms of weak strength (Chen & Xu, 2006; Xu & Xu, 2005). In contrast, the time constant parameter in the Fujisaki model, which is equivalent to the strength parameter in TA, is typically fixed (Fujisaki, 1983). In the task dynamic model, stiffness, also equivalent to strength in TA, is assumed to take only a few fixed values, e.g., a high constant value for consonants and a low constant value for vowels (Browman & Goldstein, 1992a; Nam et al., 2012; Saltzman & Munhall, 1989). In a more recent development, flexible stiffness has been used to account for gestural lengthening at the end of a phrase or sentence in the form of π-gestures (Byrd & Saltzman, 2003). But in the π-gesture model stiffness is a means of controlling gestural duration rather than completeness of target approximation.

3. In TA, velocity is a fully specified target property, so that targets can be intrinsically dynamic. This allows tones like Rising and Falling to have unitary dynamic targets. In contrast, the task dynamic model and Fujisaki model do not allow dynamic targets, and so they need to use multiple targets, gestural overlap or variable alignment to model dynamic tones (Fujisaki et al., 2005; Gao, 2008).

The target approximation model has been quantified in the form of qTA (Prom-on et al., 2009). The present paper will not focus on the quantitative aspect of the model, but some of the graphics (Figures 7, 13) to be presented later are generated with qTA.

Also, although the target approximation model was developed for tones at first, its relevance for the segmental aspect of speech has also been demonstrated (Cheng & Xu, 2013). Birkholz et al. (2011) have developed a higher-order version of the target approximation model for an articulatory synthesizer, and it has shown to be effective in simulating the articulation of Thai and German vowels in connected speech (Prom-on, Birkholz & Xu, 2013, 2014).

2.3 Edge synchronization

Edge synchronization means that a) the beginning of the syllable is the onset of the target approximation movements of most of the syllabic components, including the initial

consonant, the first vowel, the lexical tone and the phonation register; b) the end of the syllable is the offset of all the final movements. The mechanism therefore entails full synchrony at both the onset and offset of the syllable. The synchrony is asymmetrical at the two edges, however. At the left edge there is synchronous onset of all the phones involved, while at the right edge, there is synchronous offset of only supralaryngeal phones with that of either a C or V, but not both. As shown in Figure 2, at the left edge, the initial consonant is fully overlapped with the vowel. At the right edge, in contrast, the nuclear vowel and the coda consonant are sequentially aligned, with no overlap. This is also illustrated in Figure 1b, where the last syllable ends with a nasal coda, which sequentially follows the nuclear vowel /a/ rather than overlapping with it. This asymmetrical synchronization is due to the asymmetrical nature of target approximation plus the intrinsic open oral cavity of the vowel. That is, because each target is approached asymptotically, it is tolerable for a vowel to have a relatively small oral opening near the syllable onset while an intrinsically closed consonant is being executed. Near the offset, the initiation of the closing gesture of a consonant necessarily truncates the movement toward the preceding vowel.

Edge synchronization is a strong claim, as it asserts that there is full synchrony of multiple articulatory movements at both the onset and offset of the syllable. Previous accounts, notably articulatory phonology, go only so far as saying that initial consonants are more tightly coupled with the vowel than coda consonants are, because in general gestures are flexibly timed (Browman & Goldstein, 1992a; Saltzman & Munhall, 1989). Also, neither articulatory phonology nor the Fujisaki model specifies how laryngeal movements as related to pitch or phonation are aligned relative to vowels and consonants. Likewise, the frame/content theory (MacNeilage, 1998) does not say anything about laryngeal and supralaryngeal alignment.

2.4 Tactile anchoring

Tactile anchoring means that the achievement of edge synchronization relies on sensory anchors that serve as reliable alignment references, and that tactile sensations generated by the articulation process best serve as such anchors. This hypothetic mechanism is the final piece that completes the syllable puzzle, because it is key to determining the precise locations of syllable boundaries, and to explaining a number of remaining issues.

Tactile anchoring explains why the points of synchronization are at the edges rather than the center of the syllable. Most previous theories of the syllable regard the center, where sonority is the highest, as the core of the syllable (see detailed review in Ohala, 1992). Tactile anchoring predicts, in contrast, that the center of the syllable, where contact sensation is likely weak, would be the least reliable as anchors.

Precise predictions, however, are not yet easy to make based on tactile anchoring, particularly in the case of consonant clusters. This is because specific predictions will need to take the relative amount of sensory feedback a consonant generates into consideration. Short of such information, predictions and explanations can be made only for cases where tactile sensations robustly differ between adjacent segments. This will be discussed in the evidence section.

3. Evidence

Many pieces of evidence already exist in the literature for all three proposed mechanisms. Some were reported many decades ago, but have mostly been dismissed or ignored. Others have emerged more recently. Most of them are scattered in the literature, however, and their relevance for the syllable is often not obvious in the original reports. Also the evidence provided will be of two kinds, rational and empirical. Rational arguments are particularly important because they help to maximize the coherence of the theory.

3.1 Evidence for target approximation

Target approximation is a critical component of the current proposal because it is the basis for defining the onset and offset of each movement. There are three key ideas to target approximation, as illustrated in Figure 3. First, surface trajectory is the result of asymptotic approximation of successive underlying targets. Second, there is neither overlap of adjacent targets nor gaps between them. Third, targets have velocity specifications, i.e., they can be either flat or with a slope of various degrees. The following sections provide evidence for each of the ideas.

3.1.1 Asymptotic approximation

The nature of asymptotic approximation is probably most clearly seen when speech is compared with singing, both highly sophisticated human vocalization skills. Figure 4 shows a sequence of notes in a musical score overlaid by the pitch tracks of an amateur singer. The figure shows that the sung melody consists of virtually flat F_0 plateaus connected by quick transitions, where the plateaus correspond to individual notes. The transition at the beginning of each note is so rapid that there is often a momentary overshoot of the note followed by a quick return to its proper level. F_0 then stays at that level for the rest of the note's duration. Such an overshoot is one of the key characteristics of the singing voice as found by Saitou, Unoki & Akagi. (2005). In speech, in contrast, a target is typically only approached, as can be seen in Figure 5 for Mandarin tones (Xu, 1999). In each of the three panels, F_0 contours are plotted for tonal sequences that start with a High (H) tone followed by a High, Rising (R) or Low tone and ending with a low followed by a High tone. The panels differ in the tone of the third syllable, which is High, Falling (F) and Rising, respectively. As can be seen, the tonal target of the third syllable is gradually approached (hence, target approximation). The same imperfect target attainment is also observed for vowels by Lindblom (1963), Moon and Lindblom (1994)[1], and Cheng and Xu (2013).

[1] Although many counterexamples have been presented after Lindblom's (1963) seminal study, the reported complete vowel attainment in those studies are mostly observed when the adjacent consonants are labials that do not compete with the vowels for tongue body control, as pointed out by Moon and Lindblom (2012).

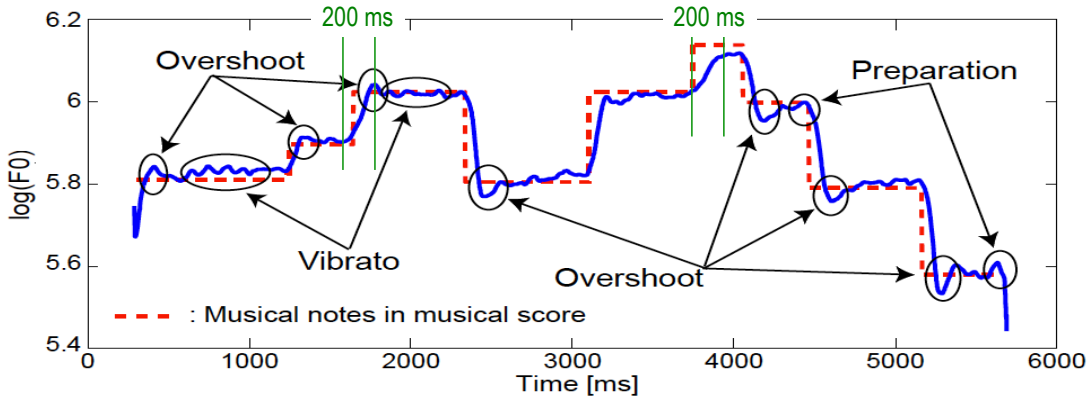

Figure 4. Examples of F_0 fluctuations in the singing voice of an amateur singer (Saito et al., 2009; courtesy of Saitou and Goto).

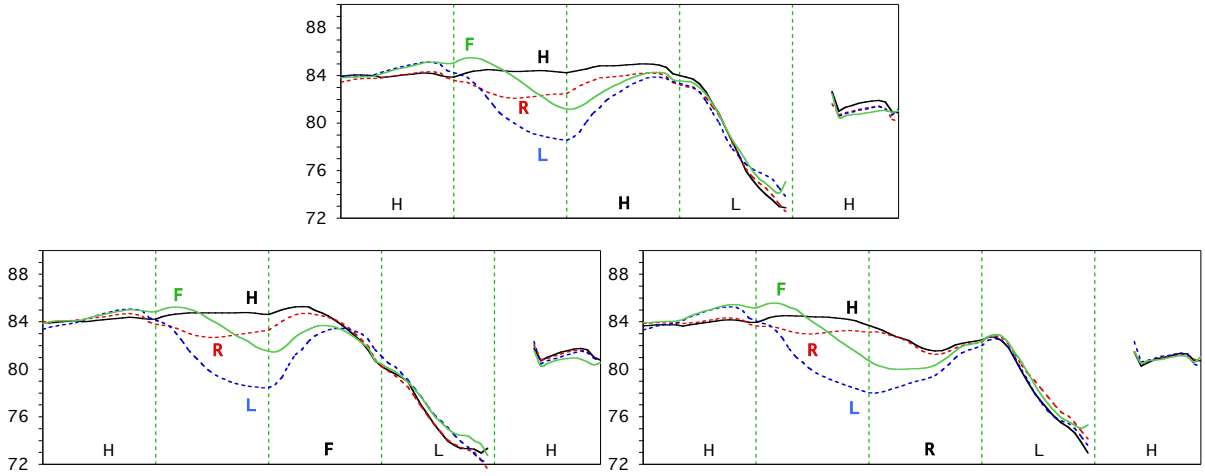

Figure 5. Mean time-normalized F_0 contours of Mandarin tones in 5-syllable sentences, where all syllables are in the form of nasal+vowel. In each plot, the tones of all the syllables remain constant except those of the second syllable, which alternate from High (H) to Rising (R), Low (L) and Falling (F). Data from Xu (1999).

Interestingly, however, it is not the case that the pitch movement in singing is exceptionally fast when compared to the maximum speed of pitch change in speech as found in Xu and Sun (2002). Rather, the transitions between notes seem to take about 200 ms as can be estimated from the plot in Figure 4. Such overshoot-and-return transition between notes seems to be a special strategy to guarantee maximal sustention of a flat notes within the designated duration. Speech, in contrast, does not seem to employ such a strategy. A spoken target is often given just enough time for an approximation, as can be seen in Figure 5.

3.1.2 Unidirectional target approximation

Another comparison that can help highlight the nature of target approximation is between speech and other skilled actions such as those in sports. Figure 6 illustrates a

sequence of movements that make up a smash in badminton. Here it is from frame 4 that the racket starts to move in the direction of hitting the shuttlecock. From frame 3 to frame 4, however, the racket is actually moving in the opposite direction of the smash. This *preparatory* movement is to ensure a maximum travel distance for the racket so as to achieve a high velocity at the end of the upcoming unidirectional movement. Thus, the preparatory movement is not part of the unidirectional smashing movement. Such preparatory movements have been seen in both singing and speech. For singing, preparatory movement in the opposite direction toward the target note is proposed to be one of the three core properties of singing voice (Saitou et al., 2009), as is illustrated in Figure 4. For speech, pre-L raising, which raises the pitch of a non-low tone before a low-pitched tone, has been reported for a number of languages (Gandour, Potisuk & Dechongkit, 1994; Gu & Lee, 2007; Laniran & Clements, 2003; Lee, Prom-on & Xu, in press; Xu, 1997, 1999). The exact reasons for these preparatory movements are not clear, but both could be due to an extra velocity needed to reach a low pitch (Lee et al., in press).

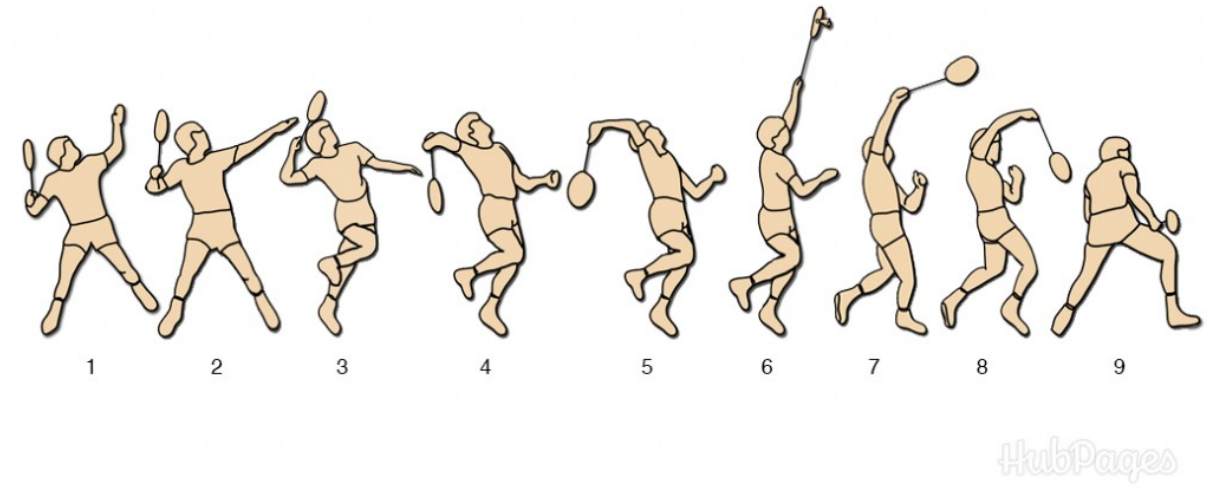

Figure 6. Frames 1-4 are the preparation phase, while frames 4-6 are the unidirectional approximation phase. The goal is not only to reach the position of the racket-shuttle contact, but also to achieve a high velocity at the point of contact. (hubpages.com)

Target approximation is not the only conceivable way of producing a sequence of F_0 contours, of course. A widely assumed alternative is target-and-interpolation (Pierrehumbert, 1981). Here a target refers to an actually realized turning point such as a peak, valley or elbow, as opposed to something aimed at as in the target approximation model (Xu & Wang, 2001). Interpolation then connects adjacent targets with straight or curved lines. There is at least one major issue with an interpolation model, however. By definition, the exact values of the points to be connected have to be known before the application of interpolation. This implies that the amount of undershoot is known before the interpretation if a target is not fully realized. But this further means that a separate (and yet unknown) mechanism has to be executed to predict the amount of undershoot before the interpolation takes place.

3.1.3 Sequential target approximation

For sequential target approximation, there are two major alternatives. One is overlapping approximation and the other is intermittent approximation, as mentioned earlier. Overlapping approximation is seen in articulatory phonology (Browman & Goldstein, 1992a), which assumes that gestures can be temporally overlapped even for a single articulatory dimension. Gestural overlap is used to explain anticipatory coarticulation as well as undershoot (Browman & Goldstein, 1992a, 1992b). The execution of gestural overlap is implemented in task dynamics as weighted average of overlapping gestures (Saltzman & Munhall, 1989). There have been arguments, however, that the movement of any single articulatory dimension consists of sequential rather overlapping execution of successive targets. This has been shown for tongue body (Wood, 1996), velum and lips (Bell-Berti & Krakow 1991; Boyce, Krakow & Bell-Berti, 1992), and F_0 (Chen & Xu, 2006). Wood (1996:161) concluded, for example, that "[s]peech motor control avoids conflicts by sequencing conflicting demands on an articulator, rather than blending them by peripheral competition, which speaks in favor of the gesture-queuing paradigm and against the tug-of-war paradigm." Also Ostry, Gribble and Gracco (1996) demonstrate that, a model based on the equilibrium point (EP) hypothesis of motor control (Laboissiere, Ostry & Feldman, 1996) is able to generate kinematic movements that show coarticulatory overlap with non-overlapping underlying control signals. Nevertheless, these findings have so far not been able to offset the appeal of gestural overlap. An important reason is the non-unique relations between the observed articulatory/acoustic trajectories and the hypothetical underlying control parameters, as illustrated in Figure 7.

Figure 7 shows trajectories generated with the qTA model (Prom-on et al., 2009). In panel a there are three successive target approximation movements, and each target is largely attained by the movement offset. These movements are strictly sequential, as indicated by the alternating line patterns. Panel b also shows three target approximation movements, but the first one is much shortened relative to that in panel a, resulting in an undershoot, i.e., incomplete attainment of the target. From the graph it is clear that the undershoot is due to a premature termination of the first target approximation movement by the early onset of the second first movement, which effectively *truncates* the first movement. But the truncation also makes the offset of the first movement appear "assimilated" to the second target, as indicated by the arrow. When the time reference (vertical line) remains unchanged from panel a, the first movement also appears to "anticipate" the second one, although there is no true anticipation given the clearly marked movement boundary. In panel c, instead of truncation, the final portion of movement 1 and the initial portion of movement 2 are overlapped. The overlap is implemented by inserting a new target which is the average of the first and second targets (There are also other, more sophisticated ways of blending, cf. Saltzman & Munhall, 1989). This *blending* thus explicitly models an "anticipatory assimilation." The resulting trajectory, however, is not that different from the one in panel b if the boundaries are ignored[2]. Thus truncation can generate trajectories very similar to those generated by

[2] Compared to panels a and b, the second target is less fully attained in panel c. This is because the blending also shortens the target approximation movement of the second target. Thus there is less than enough time to reach the target even though the onset of the movement is actually higher than in the other two panels.

blending. So, apparent assimilation do not necessarily result from gestural overlap.

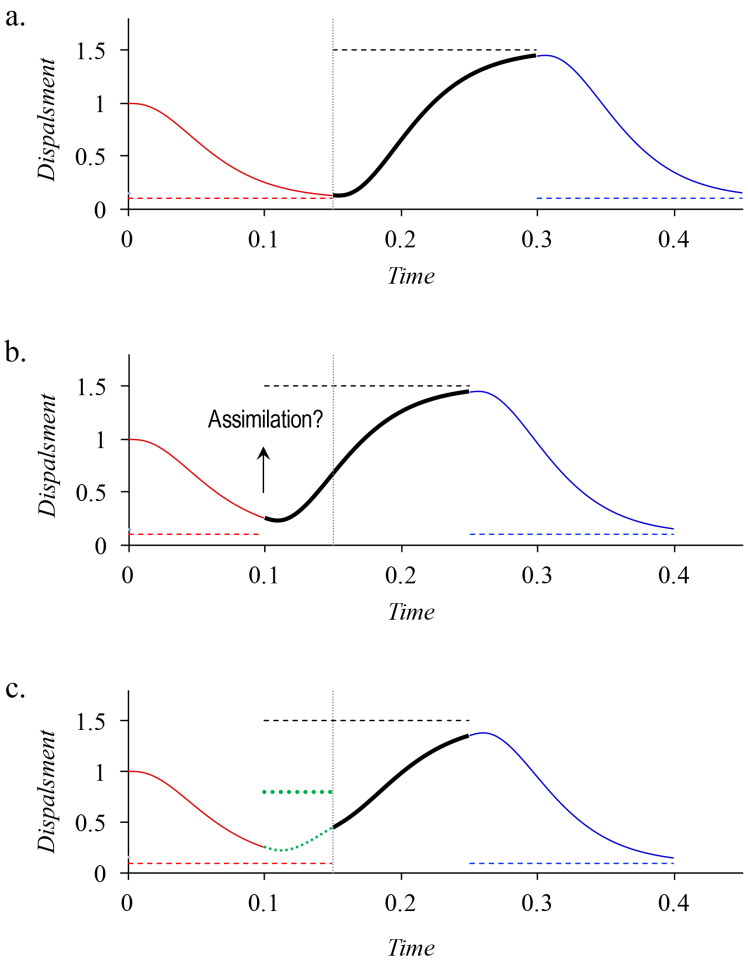

Figure 7. Sequential and overlapping target approximation processes generated with the qTA model. The units of both axes are arbitrary. In a, the three target approximation movements are strictly sequential, and the vertical line is the boundary between the first two movements. In b, the vertical reference remains at time 0.15, but the first movement is shortened by 0.05 unit. All the movements remain sequential. In c, the first and second movements overlap with each other by 0.05 units. The overlap is implemented by applying a blended target (horizontal green dotted line), which is the average of the first two targets.

Another major alternative to sequential target approximation is intermittent target approximation, which assumes that there are temporal intervals that are targetless. A weaker version is seen in articulatory phonology, where intervals not given any gestural scores can have default schwa-like targets, which do not need special specifications (Saltzman & Munhall, 1989). A stronger version is seen in the Fujisaki model, in which F_0 of intervals without commands are attributed either to the return phase of commands, which gradually move toward a slowly declining baseline (Fujisaki et al., 2005). Thus in both alternatives, intervals where full target attainment is not clearly observed are treated as governed by rather

different mechanisms. This contrasts the target approximation model which assumes that movements in all non-pausal intervals involve the same mechanism, with differences only in the target parameters. Findings from our modeling experiments so far has shown that such a single-mechanism approach is sufficient to simulate complex F_0 contours (Prom-on et al., 2009; Xu & Prom-on, 2014) and formant movements (Prom-on et al., 2013, 2014).

3.1.4 Full vs. underspecification of targets

The idea of intermittent approximations is closely related to an observation of severe variability in the form of undershoot or a seeming lack of fixed targets. The observation has led to the popular idea of underspecification (Arvaniti & Ladd, 2015; Keating, 1988; Myers, 1998; Steriade, 1995), i.e., certain units do not have fully specified phonetic values, and their surface pattern comes from interpolation between adjacent, specified units. But as pointed out by Boyce et al. (1992), support for this idea is weakened when highly systematic observations are made. Their argument was based on cases of lip rounding and nasalization. A similar finding was made in our study of the neutral tone in Mandarin (Chen & Xu, 2006). In Figure 8a, the F_0 contours of the Falling (F) tone in the second syllable converge quickly to a falling slope after four different tones in the first syllable. In contrast, the F_0 contours of the neutral tone sequence do not show convergence by the end of the syllable, but they nevertheless gradually converge to a mid-level F_0 by the end of the third neutral tone. The convergence indicates that the neutral tone has its own target, which is at a middle level between the Falling tone and the Low tone as shown in Figure 8c. But the slow convergence, as compared to the quick convergence in Figure 8a indicates a weak articulatory strength during the approximation of the neutral tone target.

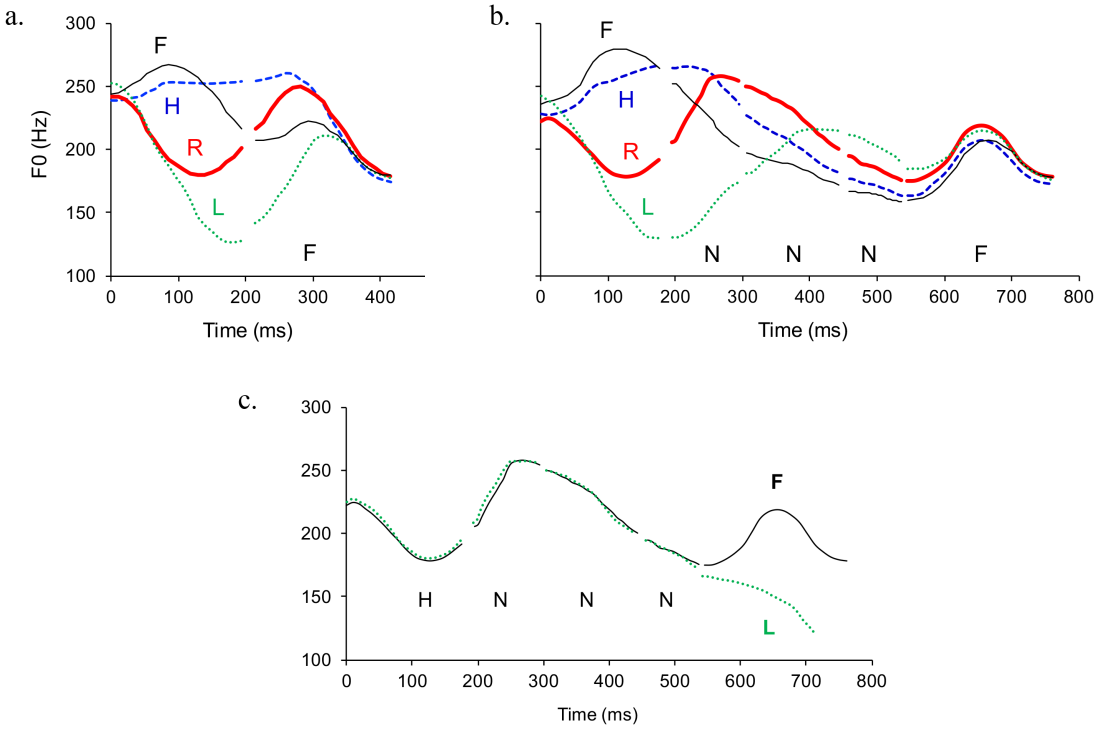

Figure 8. F_0 contours obtained in Chen & Xu (2006). a. Four Mandarin tones followed by a Falling tone. b. The same four Mandarin tones followed by a sequence of neutral tones. c. A sequences of neutral tones followed by either a Falling tone or a Low tone.

From Figure 8c we can also see that there is no anticipatory effect of the F_0 differences due to tone of the final syllable upon the preceding neutral tones. This suggests that there is no need to assume leftward overlap of the full tone target with that of the preceding neutral tone, a solution adopted by Browman & Goldstein (1992a, 1992b) for the schwa in terms of gestural scores.

3.1.5 Dynamic targets

From Figure 6 we can also see that the goal of the smash is not only to reach the position where the racket comes into contact with the shuttlecock, but also to achieve a high velocity at the moment of contact. Thus the target of the smash consists of specifications for position as well as velocity. Likewise, we can clearly see in Figure 5 that the F_0 contours of the Rising tone after four different tones all converge to a linear rise with high final velocity. The final velocity of the Rising tone is so high in Figure 8a that the F_0 rise continues for over half of the syllable in the following neutral tone. It has also been shown that both in dynamic tones of Mandarin (Xu, 2001) and Cantonese (Wong, 2006), and in diphthongs in American English (Gay, 1968), the final velocity of F_0 and formants is what remains the most constant when speech rate varies from normal to slow. Thus speakers seem to aim at a highly specific velocity as part of the phonetic target associated with those linguistic units.

To summarize, given that an articulator has to constantly alternate its state to produce a series of linguistically meaningful sounds, it has to engage in a sequence of movements toward quickly shifting targets. While it is conceivable, and in fact widely assumed, that the movements of the same articulator can be temporally overlapped with each other or intervened by targetless gaps, it is also possible for an articulatory dimension to be engaged only in sequential target approximation movements with neither overlap nor gaps. The evidence just discussed shows that simple, successive target approximation movements can also generate highly variable phonetic patterns without overlapped, intermittent, underspecified or targetless movements. The evidence discussed in this section mainly comes from movements of a particular articulator—the larynx. Evidence involving the other articulators will be discussed later, as it will be more directly related to the other two mechanisms. What is critical is that sequential target approximation makes it possible to clearly define temporal domains of movements of any particular articulatory dimension, which is key to the following discussion of edge synchronization.

3.2 Evidence for edge synchronization

Synchronization is the core of the current proposal, as it is the very mechanism that makes motor control possible. In the following, I will first discuss why edge synchronization is vital for speech, and then draw evidence from different lines of empirical studies.

3.2.1 Degrees of freedom and edge synchronization

In the field of motor control, one of the central problems is motor redundancy, as first recognized by the Russian physiologist Nikolai Bernstein (Bernstein, 1967). That is, most motor movements involve multiple body structures: joints, muscles, articulators, nerves, etc. As described by Huys (2010:70): "*The human motor apparatus … comprises more than 200 bones, 110 joints and over 600 muscles, each one of which either spans one, two or even three joints. While the degrees of freedom are already vast on the biomechanical level of description, their number becomes dazzling when going into neural space. Functional goal-directed behavior requires that a certain order arises in this multi-degree of freedom system. From a control-theoretical perspective, this poses a seemingly unsolvable problem*." Bernstein's proposal is that during action, the motor redundancy is minimized by freezing or reducing many degrees of freedom. He suggests that this is done by temporally organizing the motor parts into a functional unit, referred to as a synergy or coordinative structure. The idea of coordinative structure is also adopted in theories of speech production, e.g., Fowler et al. (1980); Saltzman & Munhall (1989). However, these theories assume that articulatory gestures involved in the production of adjacent phonetic sounds, or even the same sound, are overlapped with each other in time by various amounts. This may actually increase the degrees of freedom, as in addition to the gestures, their amount of temporal overlap would also need to be controlled.

The solution being considered in this paper is to *eliminate most of the temporal degrees of freedom by synchronizing multiple motor movements* involved in an action or even multiple actions that form a single, complex action, so that their temporal alignments relative to each other are fully fixed. In this scenario, there is no a priori independent timing control for individual gestures. Rather, a central timing mechanism collectively controls multiple motor movements at the same time. Although this solution may seem radical, many sources of evidence can be found in the literature.

3.2.2 Synchronazation of nonskilled movements

First, it is already known that there exists a strong synchronization tendency in performing concurrent voluntary movements, even if the actions are not highly practiced. Kelso, Southard and Goodman (1979) showed that, when asked to point to two targets of different sizes (hence different levels of difficulty), each with one hand, subjects automatically performed the two movements synchronously, starting and ending both at the same time, without being told to do so. They even reached the peak velocities of the two movements in synchrony.

In a long series of studies, it is shown that when trying to perform two cyclic movements, such as wriggling two fingers, tapping on a table with two fingers, etc., only two phase relations are possible between the two movements: fully synchronous and 180º out of phase. As the movements speed up, those that are 180º out of phase abruptly shift to full synchrony, typically in one or two cycles (Kelso, 1984; Kelso, Tuller & Harris, 1983; Mechsner et al., 2001). The synchronization inclination is so strong that the same steady 180º and 0 phase modes as well as the abrupt shifts from the first to the second occur even when two people are asked to swing their legs side by side (Schmidt, Carello & Turvey, 1990).

Notably, the motor movements studied in those experiments, finger wiggling, tapping, leg swinging, etc., are not highly practiced, especially in terms of their temporal coordination. But Kelso et al. (1986) show that similar abrupt phase shifts also occur in reiterant speech. When speakers are asked to say syllable sequences like ip, ip, ip… at an accelerating rate, they inevitably shift the articulation to pi, pi, pi… when the rate reaches a critical level. While reiterance syllables are not what usually occur in natural speech, there is also evidence of synchronization in more realistic utterances.

3.2.3 Synchronazation of tone and syllable

The idea that lexical tones are synchronized with syllables, as depicted in Figure 2, is by no means immediately obvious. The widely accepted assumption is that tones are carried only by the syllable rhyme (Duanmu, 1993; Howie, 1974), thus excluding all initial consonants from the tone bearing unit (TBU), as illustrated in Figure 1a. According to Howie, the F_0 patterns that best reflect the underlying tonal features occur in the rhymes, while the patterns in the voiced initial consonants exhibit the greatest variation. The variation is interpreted as evidence that initial consonants, even if voiced, should not be considered as part of the tone carrier. Some theories assume that tones are carried only by the voiced part of a syllable (Chao, 1968; Wang, 1967), thus excluding voiceless consonants as part of TBU. Lin (1995) further suggests that coda nasals are also not part of the tonal domain in Mandarin. In the following, I will show four lines of evidence that tones are fully synchronized with entire syllables, based on tonal dynamics revealed by systematic examination of continuous F_0 contours and their alignment to the segmental units of the syllable.

3.2.3.1 *The convergence to the canonical form of tone starts from syllable onset and ends at syllable offset*

In Figure 5 that we saw earlier, the F_0 contours of each tone in the second syllable start at various heights depending on the preceding tone. But they then all gradually converge to a linear shape appropriate for the respective tone. As a result, it is indeed the case that the rhyme, which is in the later part of a syllable, bears the F_0 contours with the greatest resemblance to the underlying tonal shapes. But as can be also seen in the figure, the movements toward those tonal shapes are continuous across the syllable. Thus the *tone-approaching* movement seems to be executed throughout the length of the syllable. In other words, the target approximation (as opposed to steady-state) of a tone is synchronized with the entire syllable rather than restricted to the rhyme. Similar full correspondence of tonal target approximation movement with the syllable is also found for Cantonese (Wong, 2006) and Shanghai Chinese (Ling & Liang, 2015).

3.2.3.2 *Tonal target approximation starts from syllable onset even if initial consonant is voiceless*

If tone articulation fully coincides with the syllable, tonal target approximation should start from the onset of the syllable even when the initial consonant is voiceless. This is indeed the case for Mandarin (Xu & Xu, 2003) and Cantonese (Wong, 2006). Figure 9 displays the F_0 contours of Mandarin syllables /ma/, /da/, /ta/ and /sha/ with the Rising tone (left) and the Falling tone (right). In the case of /ma/, the transition toward the tonal target is clearly visible.

In the other cases, F_0 is absent at the beginning of the syllable because of lack of voice. At the onset of F_0 in these cases, there is an upward perturbation of F_0, an effect that is mostly aerodynamic (Hanson & Stevens, 2002). Despite this local perturbation, however, the rest of the F_0 curves in /da/, /ta/ and /sha/ look very similar to those of /ma/. That is, by the time the apparent local perturbation is over, F_0 is already quite low in R but quite high in F, and the subsequent F_0 contours are fully overlapped with those of /ma/. Apparently, the movement toward the initial value of a contour tone has started at the syllable onset rather than at the voice onset. Thus target approximation is executed throughout the syllable whether or not the vocal folds are vibrating during the initial consonant.

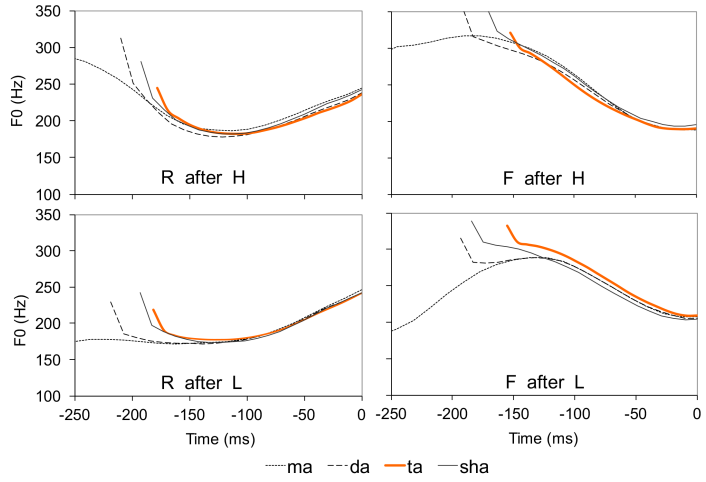

Figure 9. Effects of voiceless consonants on the F_0 contours of Mandarin R and F produced after H and L. Each curve is an average across 5 repetitions, 2 carrier sentences and 7 female speakers. All curves are aligned to the syllable offset.

Also, Gu, Hirose & Fujisak (2007) found in their modeling of Cantonese tones with the Fujisaki model that the most optimal onset of each tone command is 50–100 ms before the onset of the rhyme. This earlier onset is consistent with the idea that the F_0 movement toward a tonal target starts with the syllable-initial consonant rather than with the rhyme. Also, there is preliminary evidence that target-approximation-like F_0 movements are likewise fully overlapped with the syllable in English (Xu & Wallace, 2003).

3.2.3.3 *Tonal target approximation ends at syllable offset even if there is a coda consonant*

Furthermore, the entire syllable as the interval of target approximation does not seem to be affected by the presence or absence of coda consonants (Xu, 1998, 2001). Figure 10 shows comparisons of mean F_0 contours of disyllabic words with or without a nasal coda in the first syllable (with data from Xu, 1998), both carrying a sequence of Rising and Low tones. In each plot the two F_0 contours are plotted with their peaks aligned. As can be seen, the F_0 peak occurs in the middle of the nasal segment when the nasal is the initial consonant of the second syllable: *haomiao* in Figure 10a and *leiniao* in Figure 10b. This *delay* of the F_0 peak associated with the Rising tone into the following syllable has been found to be a consistent feature of Mandarin (Xu, 1997, 1999, 1998, 2001). When the first syllable ends

with a nasal coda (*hongmi* in Figure 10a and *lingming* in Figure 10b), which forms a nasal geminate (two nasals in a row without obvious break) with the following initial nasal, the early part of the nasal geminate shows a sharp rise comparable to the final portion of the vowel in the CV words. Thus the final part of a tonal target approximation is executed in the early portion of the nasal geminate (where the coda nasal belongs), which is indication that the entire syllable is the domain of tonal target approximation whether or not there is a nasal coda. The virtually parallel F_0 contours in each plot indicate a lack of major deviation of F_0 contours once the two are aligned by their F_0 peaks, a phenomenon that will be visited again in 3.2.4.

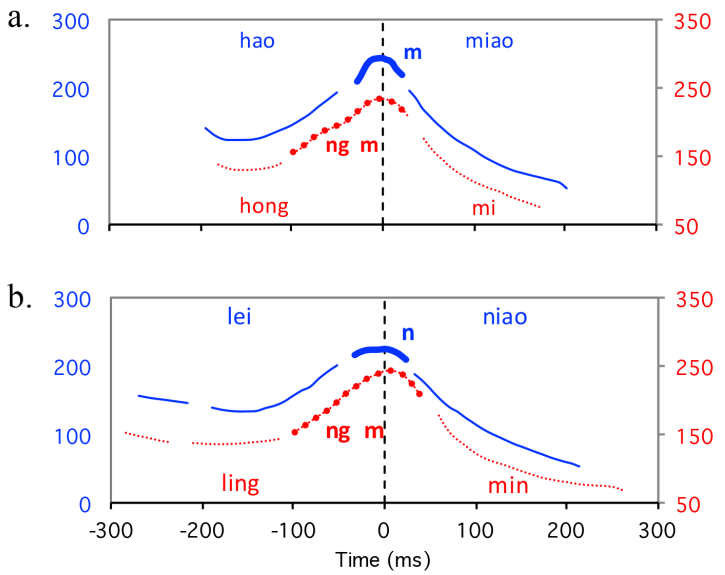

Figure 10. Mean F_0 contours of four Mandarin disyllabic words with nasal codas in the first syllable: *hongmi* [red rice] and *lingmin* [agile] versus phonetically similar words without coda: *haomiao* [millisecond] and *leiniao* [thunderbird]. In each plot the curves are aligned by peak F_0 as indicated by the vertical dash, but the second curve is lowered by 50 Hz (using axis labels on the right) to avoid visual crowding. The thick curves in the middle of each word are where the nasal murmurs are. The breaks in each F_0 contour are inserted to indicate segmental landmarks in the otherwise continuous F_0 movement. Data from Xu (1998).

3.2.3.4 *Tone-syllable synchrony does not change under time pressure that leads to undershoot*

The F_0 contours in Figure 5 also reveal another line of evidence for tone-syllable synchronization. In the bottom left plot, the longest cross-syllable F_0 transition occurs when the tone of the second syllable is Low and that of the next syllable is Falling. But this is also where the F_0 peak in the third syllable is the lowest among all the tones. In other words, there seems to be an undershoot of the F_0 peak when the time pressure is the highest. What is

remarkable is that there does not seem to be an adjustment of the timing of target approximation, as the rising movements toward the tone of the third syllable always start at the same time (around the vertical line) regardless of the tone of the second syllable. This is further indication that the timing of the onset of tonal target approximation is fixed rather than flexible.

3.2.4 Co-production of consonant and vowel at syllable onset

The evidence shown so far is for the synchronization of tones with syllables. Here the syllable refers to its segmental structure made of consonants and vowels. But the synchronization hypothesis also says that the initial consonant and the first vowel of a syllable are aligned to each other at the onset. There are actually multiple sources of evidence for C-V synchronization at the syllable onset. But they are scattered, and often in need of careful interpretation. Interestingly, most of these sources are intimately related to the phenomenon of coarticulation. In fact, the original proposal of the German term "Koartikulation" by Menzerath and de Lacerda (1933) was based on their observation that "the articulatory movements for the vowel in tokens such as /ma/ or /pu/ began at the same time as the movements for the initial consonant" (Kühnert & Nolan, 1999:14). Even earlier than that, there were observations that there is "preparatory" activity for the vowel during the preceding consonant at the beginning of a syllable (Jones, 1932; Scripture, 1902). According to review coarticulation by Kühnert & Nolan (1999), such "preparatory" activity can sometimes go as far ahead as during the vowel of the preceding syllable.

Even more relevant is the notion of *articulatory syllable* proposed by Kozhevnikov and Chistovich (1965), based on the observation that in Russian the lip protrusion movement of /u/ always begins at the same time as the first consonant when there is either one or two consonants preceding the vowel. According to this proposal the articulatory syllable is the domain of coarticulation. That is, as long as the consonants does not involve contradictory movements, all the articulatory actions connected with one articulatory syllable, including the vowel, start at its beginning (see review by Kühnert & Nolan, 1999). But the idea of the articulatory syllable was later disputed, as it is inconsistent with findings of acoustic or articulatory properties of a segment that occur well before its acoustic onset (Carney and Moll, 1971; Fowler 1981; Kent & Moll, 1972; Moll & Daniloff, l971; Öhman, 1966), which is taken as evidence of preparatory activities in anticipation of the segment (Kent & Minifie, 1977; Kühnert & Nolan, 1999).

What is critical here, however, is the concept of *preparation*. Recall that in Figure 6 frames 1-4 constitute the preparation phase whose function is to create sufficient travel distance for the racket to reach a high velocity at the moment of contact with the shuttlecock. The movement between frames 4-5, however, is already in the direction of making the ultimate contact, and so should not be considered as preparatory. Similarly, as explained in 3.1.2, the entire unidirectional movement toward the target constitutes a target approximation process. Thus the initial portion of a unidirectional movement should not be considered as a preparatory activity separate from the rest of the same target approximation process.

With target approximation as the basic framework, the findings of *anticipatory* activity of the vowel during the preceding consonant can be reinterpreted as evidence that the

consonantal and vocalic target approximation occur at similar times near the syllable onset. This interpretation also applies to the classic finding of Öhman (1966:165) that in a V_1CV_2 sequence, the final portion of the V_1 shows formant movements not only toward C, but also toward V_2: "[A] motion toward the final vowel starts not much later than, or perhaps even simultaneously with, the onset of the stop-consonant gesture". Ohala & Kawasaki (1984:116) went a step further, suggesting that "given the well-known fact of anticipatory coarticulation, the true 'beginning' of one syllable actually occurs in the middle of the preceding syllable."

Xu and Liu (2007) made acoustic comparisons of disyllabic words for English and Mandarin, where the onset consonant of the second syllable is either a glide /w/ or a nasal /m/. As can be seen in Figure 11a, when the onset consonant is a glide, as in *my wheel*, the continuous formant movements around /w/ are fully visible since there is no oral closure to cause spectral interruptions. This makes it easy to apply the principle that *the onset of the movement toward a target is the onset of the execution of that target*. So the onset of /w/ would be at the F2 peak (second arrow from left) where the movement toward /aɪ/ terminates and that toward /w/ begins. The same principle would suggest that the point where F2 reaches the minimum is where the execution of /w/ ends, because the subsequent rising movement is toward /i/. In other words, the articulatory interval of /w/ should be the interval between the first and second arrows in Figure 11a.

Thus the case of /w/, with its *transparent* formant trajectories, can serve as a segmentation reference for other consonants whose formant trajectories are much less transparent. To be able to do that, Xu and Liu (2007) employed F_0 contours as an additional temporal alignment reference. This is based on findings that certain F_0 events, such as turning points, are consistently aligned to segmental landmarks in Mandarin (Xu, 1999), English (Ladd et al., 1999; Xu & Xu, 2005), Dutch (Caspers & van Heuven, 1993; Ladd, Mennen & Schepman, 2000), Greek (Arvaniti, Ladd, & Mennen, 1998), Italian (D'Imperio, 2001), Portuguese (Frota, 2002) and German (Atterer & Ladd, 2004). Using F_0 turning points to temporally align the formant trajectories, Xu & Liu (2007) showed that similar division as in *my wheel* could be found in *my meal* as illustrated in Figure 11b. More precisely, first, the onset of the transition in the final portion of the vowel of the first syllable (the first arrow in Figure 11b) should be the true onset of the second syllable. Second, the initial consonant of the second syllable should end before the offset of the nasal murmur, as indicated by the second arrow in Figure 11b.

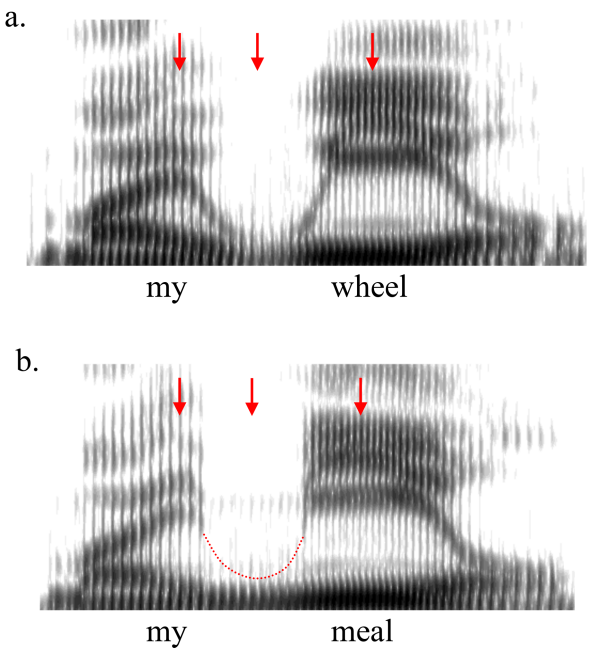

Figure 11. Spectrograms of "my wheel" and "my wheel". The arrows point to where the execution of one segment terminates and the next begins. The dotted curve in panel b portrays a hidden F2 movement related to the bilabial articulation, based on the finding of Löfqvist & Gracco (1999).

The first conclusion is particularly significant, as it says that a) the true onset of a syllable is well ahead (at least 50 ms according to Xu & Liu, 2007) of the conventional syllable onset, i.e., the start of the initial obstruent closure; b) this onset is actually *acoustically transparent*, as the final formant transition in the vowel before a consonant is usually visible rather than hidden in the spectrogram, and c) any activity related to the vowel of the second syllable during the final transition in the first syllable, as observed by Öhman (1966), would be within the C_2V_2 syllable in the $C_1V_1C_2V_2$ sequence rather than in anticipation of V_2 across the syllable boundary. Onset at 50 ms before consonant closure may not be quite right, however. This is because the vowels in the first syllable were all diphthongs in the study. As mentioned in 3.1.5, diphthongs are likely to have dynamic vocal tract targets, which may end with a high-velocity movement whose direction cannot be reversed, as illustrated in Figure 3. This means that the actual onset of the second syllable could be even earlier. But this will be further discussed in 4.7.

Also it is still possible that vowel-related activity starts even before the onset of activity related to the initial consonant. There is preliminary evidence that this is not the case at least for Mandarin (Gao & Xu, 2013) and Japanese (Chiu et al., 2015). Figure 12 provides an illustration. Both plots show formant trajectories (average of F2 & F3 by 7 speakers with 8 repetitions each) of a series of four syllables that differ segmentally in the second, third or fourth syllable. All series have the tone sequence Low, Rising, Rising, Low. The formant

trajectories in each plot are aligned by the two F_0 peaks associated with the two Rising tones, as indicated by the two vertical lines in the middle. Those F_0 peaks, as discussed earlier, would consistently occur within the /l/ murmur. In both plots, the difference in the fourth syllable extends back only to the middle of the third syllable. This would eliminate the contribution of the fourth syllable to the real contrasts in question.

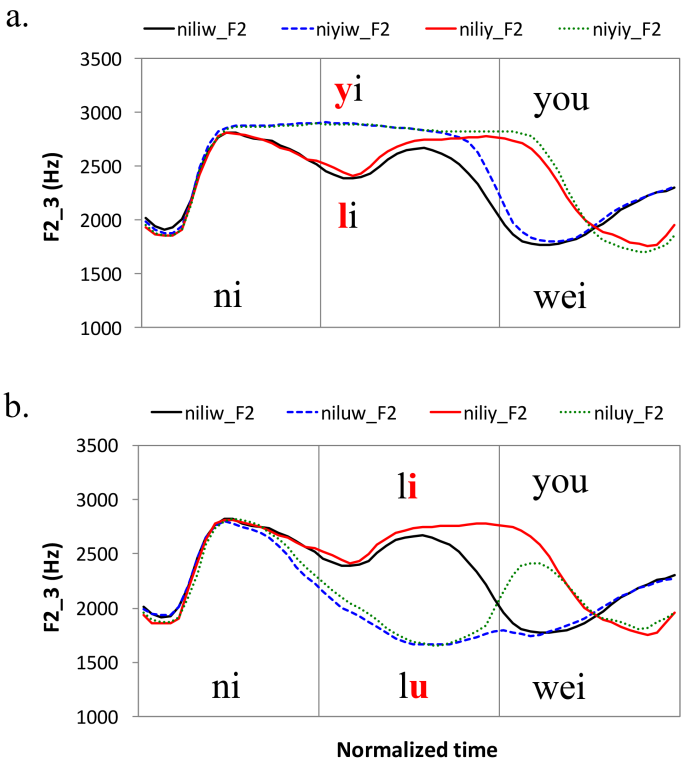

Figure 12. F2-3 movements of Mandarin phrases (*bi*) *ni* X *you*/*wei* (*shan*) [more friendly/hypocritical than X, where X is personal name *Yi*, *Li* or *Lu*. All phrases have the tone sequence L R R L F. The mean formant trajectories are all aligned to the F_0 peaks in the two R tones, at the location of the two vertical lines in the middle, which are used as references to conventional syllable boundaries (Xu & Liu, 2007). Panel a shows a consonant contrast between /y/ and /l/ in the middle syllable, and panel b shows a vowel contrast between /i/ and /u/.

The trajectories are plotted in such a way as to answer the question: when exactly does the execution of an initial consonant or a vowel start? The starting point is defined as the moment when the contrasting trajectories begin to diverge, which can be seen clearly in the early portion of each plot. In Figure 12a the contrast is between two initial consonants: /l/ vs. /j/ in the second syllable, while in Figure 12b the contrast is between the nuclear vowels in the third syllable: /i/ vs. /u/. The points of divergence in the two panels are at roughly the same temporal location whether the contrast is between two initial consonants or two nuclear vowels. This is so far the clearest evidence we have seen that the execution of initial consonant and the first vowel of a syllable starts virtually simultaneously.

The co-production of consonant and vowel at syllable onset, once in light of target approximation, can be quite obvious in the speech signal. Take the sentence in Figure 1 again as an example. The rise of F1 and F2 toward the high extremes of /i/ in /li/ starts not from the onset of /i/ as marked in panel a, but from the middle of the preceding /a/. And, because the middle of /a/ is also where F1 starts to drop toward the low extreme in /l/, that is where both /l/ and /i/ start, as is marked by the segmentation in panel b.

3.2.5 Benefit of synchronization in motor learning as revealed by modeling

A major benefit of reducing degrees of freedom suggested by Bernstein (1967) is for motor learning, given the insurmountable challenge for learners to master simultaneous controls of all the dimensions involved in a motor skill. The core idea of the synchronization hypothesis is that motor synchrony removes a large and critical number of degrees of freedom. To illustrate the problem with a relatively simple case in Figure 13. The solid curve in all three panels is the F_0 contour of a five syllable Mandarin sentence with the tone sequence R N N F H N (where N is the neutral tone). The dashed curve in each panel is the curve generated by PENTAtrainer1, a modeling program that learns underlying pitch targets by optimized local fitting (Prom-on et al., 2009). The three panels differ in their assumed timing when learning the target of the F tone. In panel a, the onset time is the same as in the original utterance. In panel b the assumed onset is 40 ms earlier than the original, and in panel b it is 40 ms later than the original. Despite the differences in the assumed timing of target onset, the fitted curves generated by the learned targets are all very close to the original. Thus the timing dimension seems to be *highly redundant* for the purpose of tone learning, assuming that the F_0 production mechanism is target approximation, and the task of a learner is to estimate the underlying pitch target of each tone (Prom-on et al., 2009).

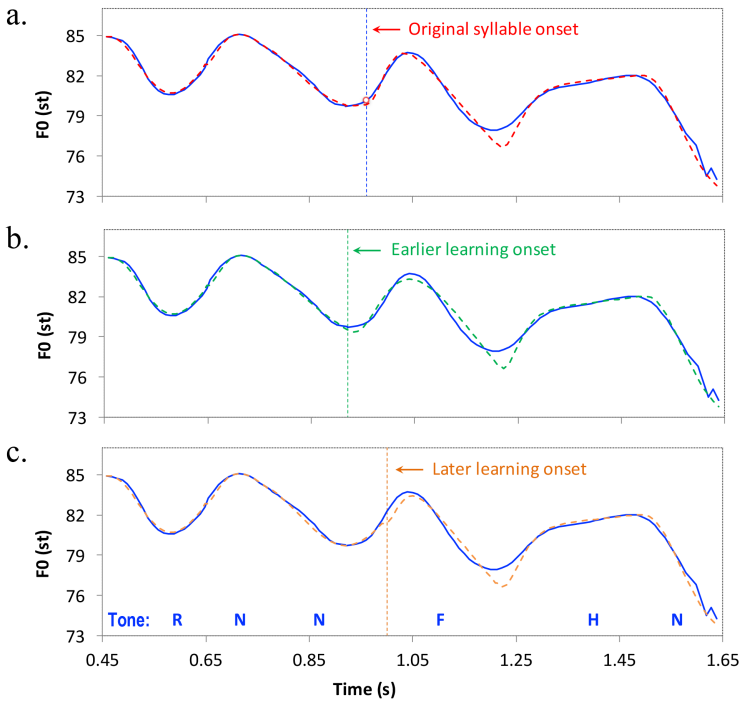

Figure 13. F_0 curve fitting by PENTAtrainer1 for a 6-syllable Mandarin utterance with a tone sequence of R N N F H N. The fitting was done by finding an optimal underlying pitch target in the qTA model for each tone. The three panels differ in the assumed timing of F tone onset during target learning: Original in panel a, 40 ms earlier in panel b, and 40 later in panel c.

But there is nevertheless still potential benefit to learn precise timing at the same time of learning the targets. This possibility is tested in Xu & Prom-on (2015) using a modified version of PENTAtrainer1. Three learning conditions were compared: Fixed timing, b) 50-ms timing flexibility, and c) and 100-ms timing flexibility. The time resolution (which is yet another degree of freedom) for the flexible timing conditions was limited to 20 steps regardless of the degree of flexibility. The learning simulations showed significant differences across the three learning conditions, but it was the 0 ms condition that produced better performance than both the 50 and 100 ms conditions, as indicated by root means square error (RMSE) and Pearson correlation, curve fitting widely used in speech synthesis. Interestingly, the learned alignments in the flexible conditions were still centered around the original syllable onset, with an average of –2.3 ms deviation from the original syllable boundaries in the 50-ms condition and -5.1 ms in the 100-ms condition (where the negative values mean that the optimized onset is earlier than the syllable boundary). Just as importantly, the increased timing flexibility led to 20 times more hypothetical timing possibilities to be tested, a massive increase in learning load.

Also, as mentioned earlier, in a simulation study (through manually performed analysis by synthesis) with the Fujisaki model which assumes full flexibility of timing of both accent and phrase commands, Fujisaki et al. (2005) and Gu et al. (2007) found that the optimized commands for Mandarin and Cantonese tones are largely fixed at 50–100 ms before the onset of the rhyme despite the high variability in syllable duration.

Although it is not fully comparable to the modeling of tone learning, the finding of Nam et al. (2012) also provides evidence of the difficulty in learning articulatory timing based on a flexibility assumption. The study tries to develop an effective procedure for estimating gestural scores in the task dynamic model as implemented in the Haskins Laboratories Task Dynamics and Application (TADA) system. The focus is on efficacy of learning the boundary times of gestural scores when the position parameters of the scores are all known. The results show, however, that the correlation of the TADA-generated vocal tract variables TVs and those derived from the original flesh-point data is only $r = .56$. This is quite low, given in particular that only the timings of the gesture scores are optimized while their amplitudes are already known.

In contrast, in two studies that also used an articulatory synthesizer to test the learning of articulatory parameters through acoustic optimization (Prom-on, Birkholz & Xu, 2013, 2014), the articulatory target approximation intervals of all the gestures involved in a vowel or glide were fixed to the manually annotated onset and offset, while only the spatial and strength parameters were optimized. The resulting parameters generated continuous speech that not only had close acoustic resemblance to natural speech, but also articulatory trajectories that are highly correlated with those of EMA data, with the *r* values ranging from 0.81 to 0.88.

The findings of these modeling studies show that there is no demonstrable benefit in assuming flexible timing during the learning of control parameters. The F_0 modeling experiment shows that flexible timing would increase the learning load due to added degrees of freedom, reduce the performance level of the learning, and eventually end up with a timing pattern that is not very different from the synchronous timing. The modeling with articulatory synthesizers show that at least there is no clear advantage in having to estimate gestural timing over fixing the timing while learning only the spatial and strength parameters of gestures. Although the modeling experiments do not necessarily represent reality, it is not hard to imagine that human learners, too, have to find optimal control parameters for their own articulation.

3.2.6 Motor synchrony is not entrainment

At this point it is necessary to address an issue highly relevant for the discussion of synchronization. That is, it has become increasingly popular that many speech phenomena can be explained in terms of entrainment, a well-established physical phenomenon whereby two oscillating systems with similar natural frequencies, e.g., two pendulum clocks, gradually fall into synchrony when they are connected through some mechanical link, such as being hung on the same beam (Huygens, 1665). Entrainment has been used to explain the abrupt VC to CV shift to be discussed in section 3.3. Haken, Kelso and Bunz (1985) used a system of coupled oscillators to simulate the shift phenomenon. However, there are a number of differences between motor synchrony and entrainment that make them unlikely to be of the same mechanism, as listed in Table 1. First, in entrainment, it takes many cycles for two oscillators to reach synchrony. In motor synchrony, the shift from 180º to 0º occurs in only 1-2 cycles (Kelso, 1984; Kelso, Tuller & Harris, 1983; Mechsner et al., 2001; Schmidt et al., 1990), thus virtually instantaneous. When using a system of coupled oscillators to simulate entrainment, the fastest phase shift Haken, Kelso and Bunz (1985) achieved was in 5-6 cycles. A gradual shift across 5-6 cycles also means that in some of those cycles a phase relation is maintained at neither 180º nor 0º, which is exactly what has been repeatedly shown to be impossible by studies on motor synchrony (Kelso et al., 1986; Mechsner et al., 2001; Schmidt et al., 1990).

Secondly, as shown in the third and fourth rows of Table 1, entrainment requires that the synchronized oscillators are highly similar in their natural frequencies, and even after reaching synchrony, they may go out of phase again (Adler, 1946; Bennet et al., 2002). Neither high similarity nor phase instability, however, is characteristics of motor synchrony (Kelso et al., 1979; Mechsner et al., 2001).

Thirdly, as pointed out by van Santen and Shih (2000:1025), “articulatory actions in speech are largely nonrepetitive (i.e., in nonreiterant speech the articulatory path hardly ever passes through the same subpath twice in articulatory space), there is no reason to suspect that articulatory actions involve pendulumlike muscle behavior such as in rhythmic music, sawing, or nodding one’s head.” Indeed, Kelso et al. (1979) shows that motor synchrony occurs in a bi-manual action with no repeating cycles. Such immediate synchrony, by definition, would be irrelevant to entrainment. But it is highly relevant to the immediate synchrony at every syllable onset regardless of how dissimilar adjacent articulatory movements are across syllable boundaries. Likewise, Cummins, Li & Wang (2013) have

shown that in speaking in unison—a skill surprisingly natural to most people without much practice, speakers can easily synchronize their reading aloud of the same text. As argued by Cummins (2011), because articulatory movements in speech are non-periodic, speaking in unison cannot be accounted for by theories that use periodicity as the basis of explaining synchronization. Interestingly, the alternative he suggested is a sensorimotor coordination account which is in harmony with the tactile anchoring mechanism to be discussed in 3.3.

Finally, probably the most fundamental difference is that in entrainment, the systems being synchronized are independent of each other, with no central control. Motor synchrony, in contrast, occurs between movements that are under a single central control, or in the case of synchrony between two individuals, under a shared control maintained by sensorial monitoring (Schmidt et al., 1990). Such a central control allows direct determination of both movement timing and movement velocity, as demonstrated by Kelso et al. (1979). It is therefore possible for the central control system to issue, for each syllable, a neural signal that initiates a group of movements at the same time, and to specify the velocity of each movement so that it ends at a particular moment in time. In contrast, to simulate VC to CV shift, Haken et al. (1985) had to introduce an initial phase condition for each movement to be ultimately synchronized. One would legitimately wonder, where do those initial phase conditions come from in the first place?

Table 1. Motor synchrony versus entrainment.

Property	**Motor synchrony**	**Entrainment**
Speed of achieving synchrony	Immediate (1-2 cycles)	Many cycles
Similarity in natural frequency	No	Yes
In-synch out-synch undulation	No	Yes
Synchrony in a nonrepeating cycle	Yes	N/A
Under central or shared control	Yes	No

3.2.7 No need for syllable internal synchronization

Edge synchronization, however, does not mean that synchrony has to be applied throughout the syllable. Rather, within the syllable, there seems to be flexible timing at various places, e.g., the offset of the initial consonant and the boundary between a nuclear vowel and the coda consonant. There may also be room for temporal flexibility within a consonant cluster, as long as the first consonant is synchronized with first vowel and the tone. In the case of lexical tone, it is also possible to have two tonal targets within one syllable, as

for the Low tone in Mandarin, which likely consists of two consecutive targets when said in isolation (Xu, 2004). The boundary between the two targets is probably partially free, as it does not affect synchrony at the syllable edges. Also, synchronization does not mean that it is impossible to have fine control of relative timing of different articulatory movements. In fact, VOT (Lisker & Abramson, 1964) is a fine example of using relative timing of different articulatory movements to encode phonological contrasts. But it is probably exactly because reliable time anchors are provided by synchronization that the achievement of precise relative timing in terms of VOT is possible. Finally, the synchronization proposed here is that of onset and offset of target approximation rather than that of landmarks, whether articulatory or acoustic. Nor is it at the kinematic level, which is a point also emphasized by Ostry et al. (1996).

In summary, the discussion in this section has shown multiple lines of evidence that there is a strong tendency for consonant, vowel and tone to be synchronized by their onset at the beginning of a syllable. There is also a synchronization tendency at the offset of the syllable, although the evidence is only in regard to tone-syllable alignment. More discussion of syllable offset will be done in light of tactile anchoring in the next section. Also the key idea that synchronization benefits vocal learning by eliminating temporal degrees of freedom is supported by preliminary evidence from modeling simulation of target learning. Finally, an argument is made that motor synchronization, including that in syllable production, is fundamentally different from physical entrainment. This difference will be further highlighted in the following discussion of evidence for tactile anchoring.

3.3 Evidence for tactile anchoring

Tactile anchoring, as mentioned in 2.4, is the final piece that completes the syllable puzzle, because it is key not only to determining the precise locations of syllable boundaries, but also to understanding how synchronization is achieved. The need for tactile anchoring is already partially implicated in the preceding discussion of motor synchronization versus entrainment. The last row of Table 1 shows that motor synchrony happens when there is either a single central control or shared control. Here the shared control is the most intriguing, as it brings out the importance of perceptual guidance in motor synchrony. Its finding foreshadows an emerging consensus in the area of motor synchrony research. With a series of experiments, Mechsner et al. (2001) show that the propensity for as well as the ability to achieve bimanual synchrony is perceptual in nature. That is, naïve subjects are able to perform bimanual oscillations in a 4:3 frequency ratio that are virtually impossible based purely on body-oriented strategies, *provided, that they are given visual input of their actions that are simplified to a 1:1 frequency ratio*. This critical role of perceptual guidance is confirmed by later studies (Bingham, 2004; Ivry et al., 2004; Kovacs et al., 2010; Swinnen & Wenderoth, 2004; Wilson, Collins & Bingham, 2005). It is further shown that tactile (Buchanan & Ryu, 2005; Johansson & Flanagan, 2009; Kelso et al., 2001; Koh et al. 2015) and proprioceptive (Baldissera et al., 1991; Mechsner et al., 2007; Ridderikhoff et al., 2007; Spencer et al., 2005; Wilson, Bingham & Craig, 2003) information also plays a critical role in stabilizing in-phase coordination in bimanual tasks. Thus the perceptual guidance needed for achieving motor synchrony includes any sensory feedback, and the contribution of each perceptual channel is a function of the clarity of the information it provides to the central control system.

The similarity of the phase shift in speech to that of bimanual movements (de Jong, 2001; Kelso et al., 1986), mentioned in 3.2.2, suggests that synchronization in the syllable also can be achieved only when there is sufficient perceptual guidance. Given that the clarity of sensory information is important as found for bimanual tasks, the clearest sensory feedback during articulation would be the most useful for synchrony in syllable articulation. In speech production, visual feedback is unlikely to be very useful, as the speakers mostly cannot see their own articulation. Proprioceptive feedback should be available virtually all the time, but the information it provides is likely to be spread out evenly in time, hence not very useful. Tactile information, in contrast, would fluctuate the most with the oral opening and closing movements, and would be the most abundant when the articulators involved in making the contact happen to be sensor-rich, such as tongue tip, tongue blade and the lips (Ringel & Ewanowski, 1965). This would point to consonants, especially obstruents, as the most likely candidates for tactile anchoring.

The critical role of consonants is best seen in the onset-coda asymmetry, a widely recognized phenomenon (Hooper, 1972; Levelt et al., 1999; Ohala & Kawasaki, 1984; Selkirk, 1982; Vennemann, 1988). That is, there is more consistent alignment of consonant and vowel at syllable onset than at syllable offset, though short of full synchrony, as mentioned earlier (Browman & Goldsten, 2000; Krakow, 1999). Also, coda consonants are more vulnerable than onset consonants. For example, there is a cross-linguistic preference for CV syllables than for CVC syllables (Locke, 1983). Even in languages that allow CVC syllables, coda consonants are subject to variation in the form of reduction, deletion or resyllabification (Dell, 1988; Schiller et al., 1997). When resyllabified, the coda of a syllable becomes like the onset of the next syllable that starts with a vowel, either within a word, e.g., *ending, producing* (which becomes *en-ding, pro-du-cing*), or across word boundaries, e.g., *let us, fine art* (which become *le-tus, fi-nart*). In language teaching, resyllabification is believed to be so common as to be considered a marker of fluent speech for languages like English (Hieke, 1984). As argued in Xu & Liu (2006), the vulnerability of coda consonant is a direct consequence of sequential articulation of coda as opposed to the fully overlapped CV articulation at the syllable onset. As the duration of a syllable shortens while speech rate increases, there is less and less time to allow sequential execution of multiple segments within the same syllable. This vulnerability means that syllable onset is the only temporal location for generating reliable tactile input.

The propensity for resyllabification is further seen in a phenomenon first observed by Stetson (1951). He found that when a CVC sequence such as *pup, pup, pup…* is spoken at an increasing speech rate, it will change abruptly at one point to a CV sequence *pu, pu, pu….* Kelso et al. (1986), in a more formal experiment, find that a sequence like *ip, ip, ip* ... changes abruptly to *pi, pi, pi* … when the speaking rate is increased up to about 4/s. What is striking is that, compared to the normal speech rate of about 4.23-4.97 syllables/s (Miller, Grosjean & Lomanto, 1984; Tsao & Weismer, 1997), an abrupt shift at 4 syllables/s means that resyllabification is virtually inevitable in normal speech given the right syllable sequence.

There is doubt, however, as to whether resyllabification actually occurs, especially across word boundaries. Most studies of resyllabification rely on listener's judgment (Chiosáin et al., 2012; Goslin & Frauenfelder, 2001; Schiller et al., 1997; Treiman & Danis, 1988), but the findings are diverse. Gao and Xu (2010) used a more objective method of

determining syllable affiliation of intervocalic consonants at word boundaries in Southern British English. by comparing word initial and word final nasals and nasal geminates made of final and initial nasals. Like in Xu and Liu (2007), F_0 contours were used as an independent reference, as shown in Figure 14. As can be seen, the initial nasal murmur occurs mostly to the right of the F_0 valley. The nasal geminate seems to be split in the middle by the F_0 valley, with the right portion largely equivalent to the initial nasal, while the left portion extending much further back into the first syllable than the initial nasal into the second syllable. Compared to these two cases, the coda nasal is aligned more like the initial nasal, rather than the left portion of the nasal geminate. Compared to both the initial nasal and the right portion of the nasal geminate, however, the duration of the coda nasal is much shorter, a phenomenon Lehiste (1960) found many years ago by comparing cases like *a nice man* vs. *an iceman*.

The shortened duration of the coda nasal seems to reflect speaker's knowledge of its underlying morphological association. Thus the coda nasal in these cases could be said to be ambisyllabic, since its F_0 alignment is characteristic of initial nasal while its duration shows a sign of effort to retain its coda identity. But in terms of how the onset of laryngeal and supralaryngeal articulations are aligned to each other, the coda nasal behaves like an initial nasal.

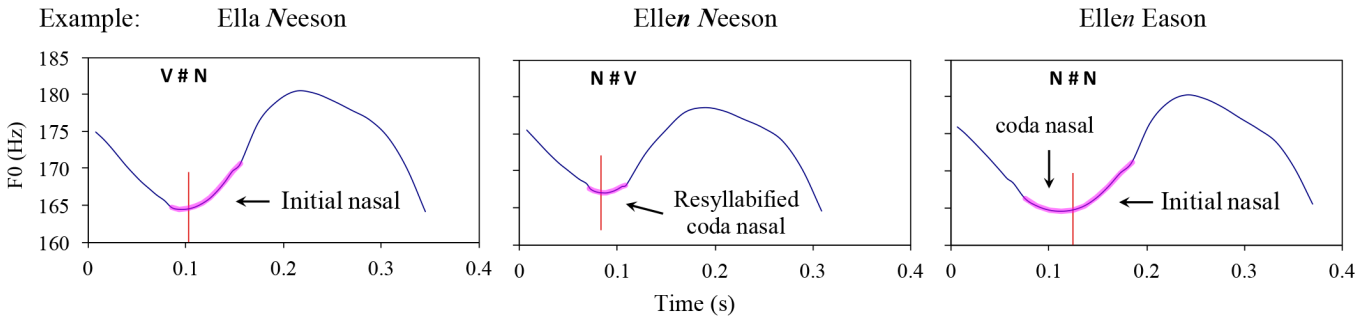

Figure 14. Grand mean F_0 contours and location and duration of the intervocalic nasals (thick section of each curve) in mean real time, averaged across all repetitions of all seven sentences, and across all eight subjects. The symbol # indicate word boundary. One set of the sentences are used as representatives of all sentence sets, as shown on top of the graphs. The italicized letter indicates the target nasal. The short vertical lines show the average locations of F_0 valley.

A further line of evidence for tactile anchoring is the finding that even within the same consonant, it is the gestures that involve greater oral contact that are attained closer to syllable edges. Sproat and Fujimura (1993) show that in English, the more consonantal (apical) gesture of /l/ reaches its extreme nearly the syllable margin, whereas the more vocalic (dorsal) component reaches its extreme closer to the nuclear vowel, whether the /l/ is a coda (hence the dark variant) or an onset (hence the light variant). The apical gesture of /l/ involves a contact with the alveolar ridge, hence the rich tactile sensation at the tongue tip would provide much more sensory feedback than the more vowel-like tongue body gesture (Ringel & Ewanowski, 1965). Another finding of similar kind is by Gick (2003) that in /w/, the labial gesture is more peripheral than the tongue gesture. Not only does the labial gesture of /w/

involve more tissue contact than the tongue body gesture, but also the lips have a rich sensory representation (Ringel & Ewanowski, 1965).

To summarize, the need for tactile anchoring is evidenced from the finding that the quality of bimanual synchrony of cyclic movements is contingent on the quality of perceptual guidance during the execution of the synchronization task (Mechsner et al., 2001 and many others cited above). Assuming that motor synchronization is the essence of the syllable as currently hypothesized, it also requires clear perceptual guidance. Of all the sensory channels available during speech production, the intermittently recurring tactile feedback would provide the best perceptual guidance. And given the vulnerability of codas, the only temporal location for tactile feedback is syllable onset. The preference for syllable onset as the synchronization site is supported by the finding that there is a strong tendency for resyllabification of coda consonants to onset of the next syllable (Gao & Xu, 2010; Kelso et al., 1986; Stetson, 1951), and that even for the same consonant, the gesture that would generate rich tactile feedback is realized near the syllable edges (Gick, 2003; Sproat & Fujimura, 1993).[4]

4. Implications and further issues

The syllable as a fundamental mechanism of speech is closely related to many of the perplexing phenomena about speech. In fact, it could be the key to some of the major unsolved puzzles, because it could provide coherent links between many of them. On the other hand, there are also questions that cannot be fully answered, for lack of relevant data. For these issues, predictions can be made based on the synchronization hypothesis that can be tested by future studies. The following discussion will start with one of the most classical findings that is at the core of the syllable puzzle.

4.1 Locus theory

Locus (Delattre et al., 1955; Liberman et al., 1967) is the phenomenon that there seems to be a virtual starting point, hence locus, for the second formant (F2) transition of a consonant, in particular /d/, such that it is consistently perceived if the first 50 ms of the transition from that point to the steady-state vowel F2 is missing from the acoustic signal, as illustrated in Figure 15b. The presence of the entire transition would, in contrast, result in the perception of different consonants (Figure 15a). As found in that study, the silent interval has to be 50 ms for all the stimuli to be perceived as /d/. As for the source of the phenomenon, it is explained that "the second-formant locus of a consonant presumably reflects the articulatory place of production, and the transition can be assumed to show the movement from that place to the articulatory position appropriate for the following vowel. The fact that

[4] It could be argued that the asynchronous alignment of the two gestures involved in /l/ and /w/ found by Sproat & Fujimura (1993) and Gick (2003) is counterevidence for the synchronization hypothesis. But the alignments reported in those studies were based on measurements of turning points in the trajectories of the articulators involved. As such they are not the onsets of the target approximation movements. Further studies are needed to examine whether the onsets of the different gestural components in those consonants are also asynchronously aligned.

the transition serves best if it does not begin at the locus might be taken as an indication that no appreciable sound is produced until at least part of the articulatory movement has been completed" (p. 772). This explanation is consistent with the discussion of C-V co-onset in 3.2.4, because during the silent closure of /d/ the vocal tract is already approximating the following vowel.

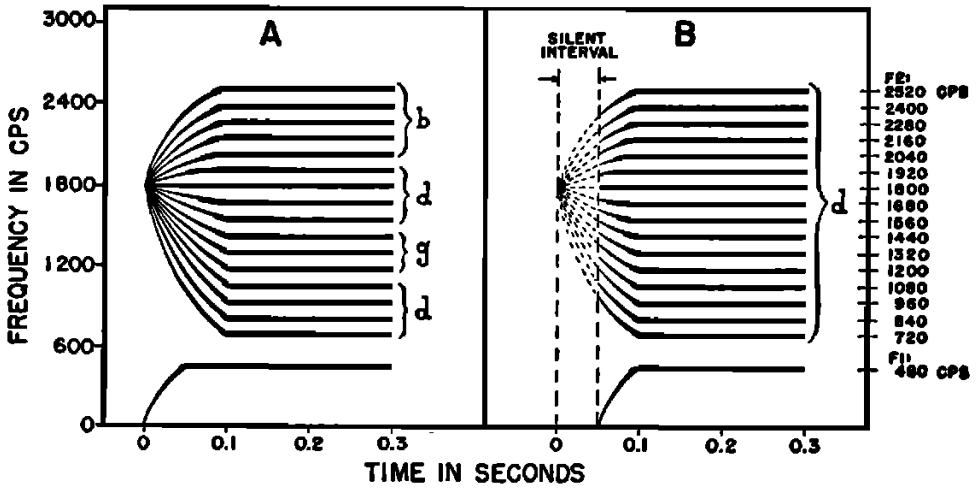

Figure 15. An illustration of the locus theory (Delattre et al., 1955. Reproduced with permission from Acoustical Society of America). The curves are F1 (the curves near the bottom in both plots), and F2 hand-painted for the pattern playback speech synthesizer (Cooper, Liberman & Borst, 1951). b, d and g mark how listeners identified the consonants.

The locus theory has been applied in synthesis systems and found to be very effective in reducing the number of consonantal allophones needed to achieve high intelligibility (e.g., Klatt, 1980; Yang & Xu, 1988). The synchronization hypothesis may explain why. As shown in Figure 1, the common starting point of the onset consonant and the first vowel of a syllable is near the start of the final formant transition before the consonant closure. Just like the hidden F_0 movement through a voiceless consonant shown in Figure 9, the formant transitions would have also continued during the consonant closure. This is in fact the essence of Öhman's (1966) finding. As can be seen in Figure 16a, starting from the beginning of the final transition in the first syllable, F2 is in a unidirectional movement toward the prototypical value of /y/, which is interrupted only momentarily by the stop closure. The transition from one syllable to the next is also influenced by the initial consonant, whose articulation coincides entirely with the first vowel of the second syllable, as discussed in 3.2.3. A replot of the locus graph to represent this idea is shown in Figure 16b, where the new locus is where the final formant transition in syllable 1 starts. Here the *warping* of the transitions is due to the intervocalic consonant, and the amount of warping would depend on the consonant, thanks to another classical phenomenon: *coarticulation resistance*, i.e., the amount of coarticulatory variability of a particular segment (Recasens, 1984a, 1984b), to be discussed in 4.5.

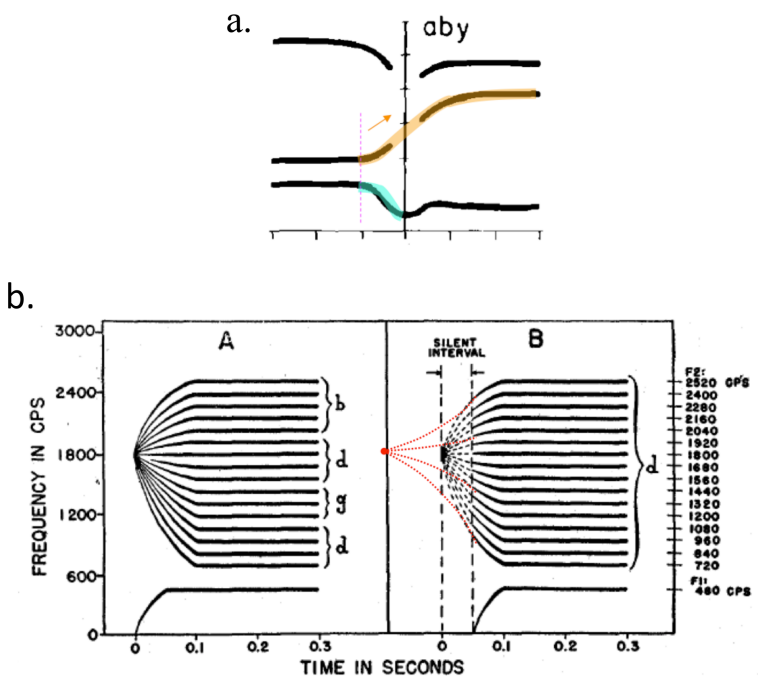

Figure 16. a) A reinterpretation of Ohman (1966). Reproduced with permission from Acoustical Society of America. Partially modified. b) A reinterpretation of Dellatre et al. (1955). Reproduced with permission from Acoustical Society of America. Par B is modified.

4.2 Locus equations

Locus equations (Lindblom, 1963; Lindblom & Sussman, 2012) are used to characterize the phenomenon that the onset of vowel F2 transition after a given stop consonant is linearly related to F2 at the "center" of the vowel across different vowels. There have been various accounts of locus equation. Sussman et al. (1998) proposes that it is due to an evolutionary adaptation that maximizes correlation between the onset and steady-state values of F2 for the ease of perception of the consonant. Fowler (1994) and Iskarous, Fowler & Whalen (2010) argue, however, that the linearity in the locus equation is related to the invariance in coarticulation resistance across different vowels. Lindblom and Sussman (2012) linked locus equation back to the classical locus phenomenon, proposing that the critical articulation of a stop consonant is the target: lips for /b/, tongue blade for /d/, and tongue body for /g/, but the rest of the articulators have no specified target and so are allowed to be coarticulated with the vowel. This account comes very close to the coarticulation resistance account by Fowler and colleagues. What is common to both accounts, however, is that they are concerned only with the articulatory or formant movement from the *onset* of the stop release to the center of the vowel.

As reiterated in the previous section, the synchronization hypothesis asserts that the movement toward the vowel in a CV syllable starts not from the voice onset after the consonant release, but from the onset of the final transition in the preceding syllable, as

illustrated in Figures 1, 2 and 12. Even if the syllable is utterance initial, the onset of the vowel target approximation would also have started *before* the consonant closure: at the same time as the formation of the consonant closure, which should be tens of milliseconds before the actual attainment of the closure. By this account, in a CV syllable where C is a stop, the onset of the formant transition is at least 100 ms *after* the onset of vowel articulation. There would naturally be a high correlation of this mid-point with the end point of the vowel articulation, because the two F2 measurements are taken simply from two locations along the same movement toward the vowel target, which originates about 50 ms before the consonant closure and is warped by of the consonant to approach its own target. So the linearity of locus equation is because it is largely a part-whole correlation, which inflates the magnitude of the correlation (Benoit, 1986; Löfqvist, 1991; Munhall, 1985). Furthermore, the slope of the locus equation would naturally be related to the amount of warping by the consonant, which is related to the amount of its coarticulation resistance.

4.3 Coarticulation

As mentioned in 3.2.4, the meaning of *Koartikulation* was very specific when it was proposed by Menzerath and de Lacerda (1933). It referred to the observation that the articulation of the vowel in a CV syllable seemed to start at about the same time as the consonant. By now, however, the term has been broadened to a very vague metaphor, referring to any influence of one segment upon another (Kühnert & Nolan, 1999). Yet it begs the question: how sure are we about the temporal domain of each segment to know for certain that properties characteristic of it in a particular location are its influence upon other segments rather than part of its own articulation? To answer this question, we need to know where a segment starts and where it ends. But that seems to make coarticulation and the temporal domain of segments a chicken-and-egg problem. The synchronization hypothesis offers a solution by a) identifying target approximation as the essence of segment articulation, which delimits the temporal domains of each segment, b) identifying motor synchronization as the essence of the syllable, which delimits the temporal domain of each syllable, and c) identifying tactile anchoring as the mechanism that makes synchronization possible, which explains why syllable onset, rather than offset, attracts the most consonants. These conceptual demarcations make most of the phenomena known as coarticulation less mysterious. In fact, many of those phenomena, according to the original definition of the term, should not even be called coarticulation. Some of the key points are summarized as follows.

1. There is no such thing as anticipatory coarticulation, in the sense of preparatory movement, *in the same direction of a phone*, before its temporal domain. This is because any movement, whether articulatory or acoustic, in the direction of a phone and is related to it, is part of its articulation. Thus the initial movement toward a phone's target is in fact the beginning part of a unidirectional target approximation; as such, it is not a preparatory act. Neither should long-distance assimilations, such as vowel harmony, be viewed as anticipatory coarticulation, because target approximation, by definition, is sequential and so simply cannot extend across a preceding consonant or vowel. They should therefore be treated as something entirely different from coarticulation, as will be discussed in the next section.

2. There is no such thing as carryover coarticulation either, i.e., lingering movement that continues the articulation of a phone *after* its temporal domain. This is because any movement in the direction of the next target is already part of its execution, i.e., articulatory approximation of its target. Target approximation, of course, happens precisely because it takes time to overcome the influence of one segment upon the next due to inertia. In physics, such influence is known as the initial condition of an event set in part by the final state of the preceding event, which also applies to motor movements (Schmidt, 1975). Initial conditions, by definition, are not part of the prior event.

3. True coarticulation happens only in the case of co-production, where two (or more) *independent* targets are approached at the same time. This occurs in the articulation of initial consonant(s) and the first vowel of a syllable. That is, due to synchronized C-V co-onset, the articulation of initial consonant(s) is executed entirely within the time domain of the first vowel, with the onsets of the two fully synchronized, as depicted in Figure 2.

4. Despite co-production of C and V, each specific articulatory dimension is only engaged in sequential target approximation (Wood, 1996; Boyce et al., 1992), and there is no temporal overlap between adjacent target approximation movements, as detailed in 3.1.3. Any incomplete target approximation is because it is *truncated* by the onset of the next one, resulting in target undershoot.

Thus the anticipatory V-to-V coarticulation reported by Öhman (1966) and reaffirmed in many later studies is actually co-production of initial consonant and the first vowel inside the same syllable. As such it is the vowel proper rather than its anticipation. By the same account, any property in an initial consonant that is characteristic of the first vowel of syllable is not in anticipation of the vowel, but part of the vowel articulation itself. Likewise, the formant transition after the release of consonant closure is not carryover coarticulation, but continued vowel articulation accelerating away from the previously co-produced consonant. This is can be seen clearly by revisiting Figure 12. In Figure 12b, although the /i/-approaching movement starts to divert from the /u/-approaching movement quite early on, in fact at about the same time as the /l/-approaching movement seen in Figure 12a, F2-3 nevertheless took a dip due to /l/, which end quite early in the second syllable. Thus both the co-production of C and V and sequential articulation of their respective key properties can be seen at the same time.

Two phenomena, however, add further complications that make the above account look overly simplistic: vowel harmony and coarticulation resistance. The following sections will address them specifically.

4.4 Vowel harmony

Vowel harmony is the phenomenon that within a certain temporal domain, e.g., word or phrase in some languages, there is a tendency, sometimes very strong, for vowels to share a particular property along a phonetic dimension, such as height, or front/back of the tongue root (Clements, 1976). When the sharing is in the leftward direction, the phenomenon is often

viewed as either a form of anticipatory coarticulation, or its fossilized remnant due to listeners' misperception (Ohala, 1994; Ohala & Kawaski, 1984).

What needs to be clarified first is that an apparent assimilation does not directly implicate coarticulation. As discussed in 3.1.3, a vowel whose articulation is truncated by the onset of the following vowel would necessarily appear *assimilated to* it, if the initial portion of the second vowel is taken as part of the first one, as shown in Figure 7b. Note that the truncation account contrasts with the overlap account shown in Figure 7c, but the two seem indistinguishable from each other when there is a lack of information about the temporal domains of the gestures involved. But just importantly, boundaries of target approximation cannot be determined by examining any trajectory on its own. Rather, an independent reference needs be used. Examples of independent references are F_0 contours (Gao & Xu, 2010; Xu & Liu, 2007), and minimally contrasting utterances (Bell-Berti & Krakow, 1991; Boyce et al., 1992; Gao & Xu, 2013; Gelfer, Bell-Berti & Harris 1989).

Secondly, there is a need to separate long-distance assimilation (i.e., that across more than one syllable boundary) from that between adjacent syllables. Long-distant assimilation is unlikely to be a form of coarticulation because it occurs across multiple target approximation movements with intervening consonants and vowels. The amount of articulatory overlap involved would not be acceptable even by a theory that do allow them (Browman & Goldstein, 1992a). Instead, it is likely to involve a change of phonetic targets before their articulatory execution. Such *reassignment of targets* is actually frequently seen in the tonal domain in the form of tone sandhi (Chen, 2000). Because the reassigned tonal targets often become very different from the triggering tone, coarticulation is unlikely involved (Xu, 2004). In both tone sandhi and vowel harmony, the reassignment of the targets may have various sources. In the latter case, it could have originated historically from surface assimilation (Gafos & Stephan, 2006), due to listeners' misperception (Ohala, 1994). What is critical, however, is that the triggering assimilation does not have to be due to anticipatory V-V coarticulation, but could well be due to truncation of a vowel by the vowel of the next syllable.

4.5 Coarticulation resistance

Coarticulation resistance is the phenomenon that phonetic segments differ in their ability to resist the coarticulatory influence of adjacent segments (Bladon & Al-Bamerni 1976; Recasens, 1984a, 1984b; Fowler & Saltzman, 1993). It has been shown that a major source of variation in coarticulation resistance is the amount of constraint that gestures of a consonant or vowel place on the tongue body (Recasens, 1984a, b). Those with intrinsically stronger tongue body constraints show greater resistance to coarticulatory influence than those with weaker constraints. The phenomenon poses a challenge to accounts that assume that there are invariant, context-free underlying articulatory targets, and contextual variation occurs due to the dynamics of articulation (Brownman & Goldstein, 1989; Xu & Liu, 2006). One solution is intergestural blending (Fowler & Saltzman, 1993), which involves blending of temporally overlapped consonant and vowel gestures.

An alternative, based on sequential target approximation as discussed in 4.3, is that, again, no blending is involved. Rather, each phone may have different levels of demand not

only for different articulators, but also for different dimensions of the same articulator. For example, a consonant may have a high demand for the vertical position of the tongue dorsum, but a low demand for the horizontal position of the tongue dorsum. Just like the case of F_0 production, any particular dimension of an articulator can only approach one target at a time. Thus for any consonant, if a dimensional target is essential to its core characteristic, the articulator involved has to approach it before starting to move toward the vowel target. For /d/, for example, the tongue tip as well as the edges of the tongue must first manage to completely seal the oral cavity at, and around, the alveolar ridge for the closure. Only after the maximal closure is achieved can they start to move toward the tongue shape of the vowel. In contrast, the back of the tongue can start to approach the vowel shape as soon as its movement toward the vowel of the preceding syllable is over. For /k/, the vertical target of the tongue dorsum is critical for its articulation, and therefore has to be executed before approaching a lower position require by any vowel. But the horizontal target is not critical, and so can start to move toward the vowel position from the syllable onset. This would explain the finding that during the /k/ closure, the point of contact between tongue body and the palate varies gradiently with the co-produced vowel, more advanced for the front vowels, and more retracted for the back vowels (Dembowski, Lindstrom & Westbury, 1998).

4.6 C-center and P-center

C-center is the phenomenon that, in a C…CVC syllable, where C…C is a cluster consisting of varying number of consonants, the most consistent temporal distance from the vowel offset is to the center, rather than the onset or offset, of the C-cluster (Browman & Goldstein 1988). On the other hand, it could be argued that the effect is simply one of duration compensation as a result of a tendency for syllables to have equal duration when the number of segments in a syllable varies while other things remain constant (Campbell & Isard, 1991). On the other hand, the phenomenon brings up a critical issue for the synchronization hypothesis: when there is more than one consonant in the onset of a syllable, how would they be aligned to each other and to the vowel? The solution offered by Browman and Goldstein (1988, 1990) is that all of the Cs in an onset compete to align with the vowel. The solution based on the synchronization hypothesis would not be fundamentally different, but the empirical evidence it will look for will be more based on the assessment of the onset target approximation movements rather than only the offset. Furthermore, it predicts that there is a possibility of a coda consonant also becoming synchronized with the onset of the following syllable even if the latter starts with a consonant. For example, in *nurse rhyme* and *second language*, /s/ and /d/ may join the following initial consonants to form onset clusters /sr/ and /dl/.

P-center or perceptual center is the phenomenon that subjects are able to align an audible click or their own finger tap to some point in a syllable (Morton et al., 1976). This lineup point typically occurs near the CV transitions of syllables, though the particular point varies with the length of the sequence. There has been much debate as to whether the phenomenon is articulatory (Fowler, 1979; Fowler, Whalen & Cooper, 1988) or perceptual (de Jong, 1994; Howell, 1988; Ohala & Kawasaki, 1984) in nature. Given the evidence that synchronization of motor movements depends heavily on sensory input, the observed p-center could be based on a combination of various sensory input channels. If a particular sensor channel happens to dominate, it would more likely to provide the major anchor points. This

sensory-based account is similar to the salience account of Ohala and Kawasaki (1984), but it does not rule out possible contribution of articulatorily based sensory channels such as tactile or proprioceptive input. Various experiments could be designed to test this hypothesis. For example, one could compare p-center obtained from normally phonated speech to whispered speech to silent speech (mouthing only). It is also possible to compare conditions where the finger taps are only either visible or audible. It would be interesting to see if the p-center would be attracted toward the most salient cues, be it acoustic or articulatory.

4.7 Delayed turning point

One issue that has been mentioned only in passing (3.2.3.3, 3.2.4) is the phenomenon of delayed turning point. That is, a trajectory that approaches two successive targets often change directions *after* the end of the first target interval. This can be seen in Figure 7b, where the low turning point occurs in the second interval where the target is high rather than in the first interval where the target is low. The exact amount of delay is a combined result of the final velocity of the first target approximation movement and the articulatory strength of the next movement. The final velocity of a movement is determined not only by the property of the underlying target, but also by the amount of time available to the target. In Figure 3, for example, the delayed peak into the second syllable is due to the dynamic rising target of the first syllable. In Figure 7b, however, the delayed low turning point is because the first target, which is static, is given insufficient time, so that the trajectory is still fast approaching the target when it is truncated by the second target approximation movement. The effect of articulatory strength of the second target on the delay of the turning point can be clearly seen in Figure 8b. After the Rising tone syllable, the ascending F_0 movement due to the preceding Rising tone is not reversed until near the end of the first neutral-tone syllable. This is because, as explained in 3.1.4, the neutral tone presumably has a weak strength (Chen & Xu, 2006), which would not generate sufficient articulatory force to reverse the final velocity of the preceding Rising tone.

Note that the delay of a turning point may not only give the impression of prolonged carryover coarticulation, but also an impression of an anticipatory coarticulation. That is, if a delayed turning point is taken as the boundary between the two targets, anything happening before it would be taken as belonging to the first target. And if the first target is truncated, as in Figure 7b, the measurement taken before the turning point could be interpreted as an indication of assimilation to the following target. Note also that the issue is relevant not only for tones, but also for segments. For example, given the finding of Gay (1968), diphthongs probably have dynamic targets just like Rising and Falling tones. As a result, their execution will achieve a high final articulatory velocity at the movement offset, which may result in a delayed turning point. So, the finding of Xu & Liu (2007), which was based on diphthongs in both Mandarin and English, could be too conservative in terms of the estimation of the onset of articulatory movement toward the next vowel (less than 50 ms ahead of the onset of consonant closure) when compared to the formant trajectories in Figure 12, where the vowel of the first syllable is a monophthongal /i/. The exact difference between monophthongs and diphthongs need be determined by future studies.

4.8 Feed-forward, feedback, stuttering, speech acquisition and evolution of speech

The early discussion of the evidence for tactile anchoring has established that its effectiveness is contingent on the quality of sensory feedback during articulation, which in turn entails that synchronization is achieved with feedback control. There is a prerequisite for feedback control, however, i.e., the error detection and correction should not take longer than the execution of the movement, as otherwise the correction cannot take effect before the execution is over (Kawato, 1999). This seems to be the case with segmental articulation in speech (Perkell et al., 1997). As argued by Perkell (2012), the articulation of segments is too fast for feedback control to be effective, and so speakers have to mainly rely on feed-forward control, which may explain why some post-lingually deaf individuals are able to speak largely normally decades after their hearing loss (Cowie & Douglas-Cowie, 1992; Lane & Webster, 1991). Indeed, target approximation as discussed in 2.2 is a feed-forward model, as it has no built-in feedback mechanism for monitoring target attainment (Prom-on et al., 2009). Once a target is selected, the model blindly executes it until the designated time is over, whether or not any or all its specifications are met. The accuracy of target approximation is achieved, not through online feedback control, but through extended offline training during learning; and even this learning is not based on direct feedback correction, but by random (Xu & Prom-on, 2014) or exhaustive (Prom-on et al., 2009) trial and error.

Thus tactile anchoring, assuming it works as claimed in the synchronization hypothesis, implies that not only is there sufficiently rich sensory feedback, but also the feedback is sufficiently fast, so that the central nervous system is able to plan, initiate, execute, monitor and rectify the timing of all articulatory movements involved in a syllable to achieve motor synchrony. This may not be easy from an evolutionary perspective, as there are only a small number of species groups that show vocal learning of syllable-like sound sequences (mammals: humans, bats, and cetaceans; birds: parrots, hummingbirds, and songbirds, according to Jarvis, 2004; but also see Chakraborty & Jarvis, 2015 for evidence of partially developed syllable-like systems in mice and non-human primates). Based on this knowledge, it is also not hard to imagine that in cases where there are some defects in the neural system of an individual that make the planning, initiation, execution or monitoring of synchronization ineffective, disorders like stuttering may occur. Indeed, one team was able to induce stuttering in Zebra finch by modifying the gene critical for timing control (Kubikova et al., 2014; Tanakaa et al., 2016). Interestingly, the induced changes nevertheless leave the structure of individual syllables in the bird songs intact. This is consistent with our hypothesis that tactile anchoring and target approximation involve different mechanisms, the first relying crucially on feedback control, while the second relying only on feed-forward control.

Chakraborty and Jarvis (2015) recently hypothesized that animals first evolved the neural circuit for motor learning, and then some species developed a second copy of the entire circuit just for vocal learning, by duplicating the brain pathways of the surrounding motor learning circuit. Based on our current discussion, this hypothesis would imply that motor synchrony is what underlies not only speech, but also motor movements in general, as reducing degrees of freedom is an essential requirement as envisioned by Bernstein (1967). Linking this back to McNeilage's frame/content hypothesis, mandibular oscillation, as a highly developed motor skill, must already have its own synchronization neural pathways in place before the emergence of speech. So the new invention in the evolution of speech should

be neural pathways for controlling synchronization of laryngeal and supralaryngeal movements, which presumably have need to be finely coordinated for other biological functions. This is consistent with the Ohala and Kawasaki (1984) hypothesis that the syllable is for the sake of synchronizing segmental and suprasegmental articulations, although the need is likely articulatory rather than perceptual.

It could be further speculated that even with the right genetic disposition, either the pathway to the brainstem needs time to develop after birth, or it requires first acquiring preliminary ability to control the key articulators involved (larynx, jaw, lips, tongue body, etc.) before attempting to synchronize their movements. This could be why canonical babbling, and with it the ability to produce syllables, start to emerge not right at birth, but around 6 months later (Kuhl & Meltzoff, 1996).

Finally, if synchronization has to rely on feedback control, planning would necessarily be involved. Whalen (1990) shows that if a speaker is told about the identity of the V in /aCV/ sequence only after the onset of vocalization, the usually observed anticipatory influence of the vowel on the consonant is absent or reduced. This is consistent with the necessity of planning for synchronization. Thus coarticulation of C and V at the syllable onset may indeed be planned as concluded Whalen, as the unseen vowel at the onset of the vocalization probably could not be activated soon enough to be fully synchronized with the consonant. But the planning here is that of synchronization rather than target approximation. Also the precision of the synchronizing all the articulatory movements in different syllables may not be easily achieved. It would thus take repeated practice during acquisition to get it right. It would then not be surprising to see evidence of stored motor commands for different CV structures (Levelt, Roelofs & Meyer, 1999). And, in the case of L2 learning, synchronization strategies like resyllabification as discussed in 3.2.5 may not be easy for the learners to figure out on their own and so may well have to be taught explicitly.

4.9 Intonation

While there are many potential implications of the proposed synchronization hypothesis for intonation research, here I will only discuss an issue that has been under debate for some time. That is, is it necessary to posit that every syllable has a pitch target? As illustrated in Figure 8b, it seems reasonable to assume that only sparse tonal specifications are needed. For example, only the F_0 values at the turning points need to be specified, while the quasi-linear portion in the middle can simply result from linking up the turning points. Such sparse tonal specification is one of the key assumptions of the Autosegmental-Metrical phonology of intonation (Pierrehumbert, 1980; Pierrehumbert & Beckman, 1988). In their account of Japanese intonation, Pierrehumbert and Beckman (1988) argue that there is no need to specify the tone of every syllable in unaccented words, because F_0 of those words can be obtained by interpolation between syllables where tones have to be specified. This would limit the degrees of freedom in tonal representation and will capture common phonological patterns shared by words and sentences of different lengths. As mentioned earlier, many other intonation theories have also adopted strategies that use underspecifications for part of the F_0 contours (Fujisaki, 1983; Taylor, 2000; ‘t Hart et al., 1990).

As discussed when motivating the synchronization hypothesis, degrees of freedom is indeed a central concern in motor control. But the concern is not about how patterns in any single acoustic dimension can be economically described. Rather, it is about how the central nervous system is able to coordinate multiple movements that affect all acoustic dimensions involved in speech. Assuming that the syllable is a mechanism of synchronizing laryngeal and supralaryngeal movements, as has been argued here, each syllable has to have a specific tonal target, as otherwise the laryngeal target approximation for some syllables would be aimless. It could be argued that it is still conceivable that a separate string of F_0 values can be computed before being imposed onto the segmental string. But that would entail a) the system is able to generate pitchless supersegmental movement trajectories and segmentless F_0 contours before combining the two, and b) both types of trajectories already contain transitional movements due to inertia. Given that inertia is a physical property, carrying out central commands that already contain inertia-driven transitions means that the effect of inertia would be applied twice, which is highly unlikely. It is therefore more likely for the articulatory system to performs simultaneous movements toward both laryngeal and supralaryngeal targets in each and every syllable.

A lesson to be drawn from this kind of consideration is that economy of representation or degrees of freedom in an intonation theory should not be assessed only in terms of adequacy of describing F_0 contours alone. Rather, it should be considered together with all the other articulatory dimensions that are also involved, and also in terms of the control mechanism that makes the learning and execution of the concurrent articulation of all the dimensions possible.

4.10 Speech technology

Though there are many potential implications of the synchronization hypothesis on speech technology, here I will focus only on one of the most enduring issues in speech synthesis, namely, whether the system should be deterministic or data-driven (Taylor, 2009). Specifically, should there be any system-internal rules for generating phonetic specifications such as those of spectral trajectories, F_0 contours, etc., or should all or nearly all phonetic specifications be statistically summarized from natural speed data? Early systems were almost invariably rule-based, hence deterministic (Klatt, 1987). Those systems, especially the best among them, e.g., DeckTalk, have achieved high intelligibility but lacked naturalness (Taylor, 2009). In fact, naturalness has become such a barrier for rule-based systems that they are virtually abandoned by now. Currently the two state of the art systems are unit selection and hidden-Markov-model (HMM) synthesis (Taylor, 2009), both heavily data driven.

Unit selection is one of the most successful concatenative synthesis systems (Hunt & Black, 1996). In this technology, recorded natural speech is segmented into fragments based on the diphone principle (Peterson, Wang & Shearme, 1958). They are then rearranged and concatenated into new sentences during synthesis. A major key to the success of unit selection is the principle of diphone for segmenting speech fragments (Taylor, 2009). A diphone is an acoustic chunk that extends from the middle of one phone to the middle of the next. Here the "middle" of a phone means the middle of the steady state of a vowel, middle of a fricative noise or middle of a stop closure (Peterson, Wang & Shearme, 1958).

Interestingly, the diphone is reminiscent of the temporal domain of a consonant as illustrated in Figure 1, according to the synchronization hypothesis. That is, a consonant would start from where the formants start to turn in the direction of the consonantal manner and place of articulation, which is often in the middle of a vowel, and end in the middle of the consonantal closure. The diphone, however, does not fit the temporal domain of a vowel based on the synchronization hypothesis, as it should have extended from the middle of one syllable (per conventioinal acoustic definition) to the middle of the next syllable, as illustrated in Figure 1, in cases where both syllables have a simple CV structure. Thus the true middle of a vowel, according to the synchronization hypothesis, should be about half of the length of the consonant closure more to the left than that of diphone. This means that a VC diphone may have left out about 100 ms of the initial part of the current vowel, while a VC diphone may have included 100 ms of the following vowel. Such a mismatch, unsurprisingly, is a major source of variability that has to be taken care of by including many context-specific diphones, making the number exceed several folds from that of mere CV, VC combinations.

HMM-based synthesis is a statistical parametric synthesis method that uses hidden Markov models to learn and generate frequency spectrum, F_0 and duration based on maximum likelihood criterion, which are then used to generate speech waveforms (Tokuda et al., 2013) through a vocoder like Straight (Kawahara, Masuda-Katsuse & Cheveigné, 1999). HMM is an algorithm for representing transition probability between adjacent states, and its adoption in synthesis is for the sake of capturing the dynamics of speech. This adoption, however, is based on the assumption that the acoustic states in speech from one moment to the next are largely independent of each other, so that the only way to capture the dynamics is to estimate the moment-to-moment transition probability. But articulation is a physical process and so follows basic Newtonian laws, which means that the moment-to-moment states are not random. Indeed, the major advances in the development of HMM-based synthesis are all in the direction of finding better ways to *implicitly* simulate the lawful dynamics of speech articulation, these include the use of maximum likelihood criterion to reduce random variation of states due to the independence assumption (Taylor, 2009), the use of delta and acceleration coefficients to model transition dynamics between frames (Tokuda, Kobayashi & Imai, 1995), and introduction of articulatory features into the training process (Ling et al., 2009). Although the HMM-based approach has achieved high intelligibility (Tokuda et al., 2013; Zen, Tokuda & Black, 2009), the synthesis tends to be over-smooth and is still not nearly as good as the best-quality concatenative systems (Hunt and Black, 1996; Taylor, 2009).

Despite the effort to take articulatory dynamics into consideration, what is lacking in both unit selection and HMM based synthesis is direct modeling of the dynamics of undershoot. Both approaches are based on the general idea of target-transition (Taylor, 2009), i.e., speech consists of steady-state targets that are connected by transitions (Holms, Mattingly & Shearme, 1964; Klatt, 1980). Following this idea, articulatory dynamics that needs to be processed mainly has to do with continuity and smoothness of the transitions. In contrast, target approximation, as part of the synchronization hypothesis, assumes that neither the starting point nor the ending point of an articulatory movement has to be *on target*. Rather, a movement may start from an incomplete realization of one target, and end with an incomplete realization of the next. But the variable realizations of the targets are nevertheless based on a consistent dynamic mechanism, which could be captured by an articulatory-based model like

those of Prom-on et al. (2009) for F_0 and Birkholz et al. (2011) for supralaryngeal articulators. Synthesis experiments based on these target approximation models have shown good naturalness in terms of both F_0 and spectral properties (Liu et al., 2016; Prom-on et al., 2009; Prom-on et al., 2013; Xu & Prom-on, 2014). Just as importantly, these studies have shown that the use of deterministic models like target approximation leads to robust reduction of training data, parameters to be trained, input linguistic features, and training time.

Ultimately, however, the issue may boil down to: Can our explicit knowledge about speech production be improved to the point that knowledge-based rules can generate acoustic signals that surpass those generated by the best technology of resequenced or reassembled natural speech? If the answer is yes, the future of speech synthesis may see a return to deterministic approaches. But the return is likely to be only in terms of articulatory dynamics, which is due to the physical nature of the speech apparatus. The learning of the target parameters, syllable duration, syllable-internal phone timing, and how they are linked to linguistic functions will nevertheless need to be data-driven, as they are likely probabilistic by nature. So a seamless integration of deterministic models with data-driven training could lead to the next generation of synthesis, with high naturalness, rich representation of linguistic and paralinguistic functions, and speaking styles.

5. Concluding remarks

What has been proposed here is not entirely new (with perhaps the exception of tactile anchoring). Many aspects of the theory have antecedents with similarities of various degrees, such as the notion of the *articulatory syllable* (Kozhevnikov & Chistovich, 1965), the proposal that the true beginning of a syllable is in the middle of the preceding syllable and that the purpose of the syllable is for the sake of synchronizing segmental and suprasegmental articulations (Ohala & Kawasaki, 1984), and the notion of C-V synchronization at the onset of the syllable (Browman & Goldstein, 1995; Menzerath and de Lacerda, 1933; Öhman, 1966). What has been lacking, however, is an account that can not only connect all the facts, but also explain why the syllable is needed in the first place. The solution, as explored in this paper, is a proposal based on the recognition that speech is a motor activity. From this recognition, the exploration here started with one of the most fundamental problems in motor control: how to reduce degrees of freedom to the extent that makes motor learning and execution possible (Bernstein, 1967). Kugler et al. (1980) have summarized the Bernstein problem as a list of principles to be followed for resolving the conundrum:

(a) keep the number of free variables to be individually regulated at a minimum; (b) keep the number of executive instructions per unit time at a minimum; (c) keep the number of executive decisions about what kind of instruction or command to issue at a minimum; (d) keep the number of executive decisions about when to issue an instruction or command at a minimum.

The *synchronization hypothesis* proposed here has offered a solution that addresses all these principles. For (a), eliminating flexibility in temporal relations between component movements of a coherent unit like the syllable, the number of free variables to be individually regulated is significantly reduced. For (b), phone-sized targets, as opposed to individual articulators, can be likened to actions in other motor movements. For (c), by assuming target

approximation as predominantly under feed-forward control, leaving only synchronization to feedback control, the number of executive decisions about what kind of instruction to issue is significantly limited. Finally, for (d), if only action-sized commands and only synchronization pulses need to be centrally issued, the number of executive decisions about when to issue an instruction is also reduced to a minimum.

Acknowledgment

This research was supported in part by National Science Foundation (NSF BCS-1355479). The funding source had no other role other than financial support. The author is heavily indebted to his collaborators of previous studies whose contributions have helped to shape the work presented here: Emily Q. Wang, Ching X. Xu, Xuejing Sun, Yiya Chen, Andrew Wallace, Fang Liu, Ying Wai Wong, Bei Wang, Hong Gao, Santitham Prom-on, Peter Birkholz, Chierh Cheng, Elbert Lee and Hao Liu. Thanks also go to Jyrki Toumenen, Adrian Fourcin, Fred Cummins, William S. Wang, Harvey Sussman, Louis Goldstein and Anotonis Botinis for comments on an early version of the manuscript. All errors are with the author.

References

Abramson, A. S. (1978). Static and dynamic acoustic cues in distinctive tones. *Language and Speech **21***(4), 319-325.

Adler, R. (1946). A study of locking phenomena in oscillators. *Proceedings of the IRE, 34*(6), 351-357.

Arvaniti, A. and Ladd, D. R. (2015). Underspecification in intonation revisited: a reply to Xu, Lee, Prom-on and Liu. *Phonology 32*, 537-541.

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

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

Baldissera, F., Cavallari, P., Marini, G. and Tassone, G. (1991). Differential control of in-phase and anti-phase coupling of rhythmic movements of ipsilateral hand and foot. *Experimental Brain Research 83*, 375-380.

Barbosa, P. and Bailly, G. (1994). Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication 15*(1-2), 127-137.

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

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

Bennett, M., Schatz, M. F., Rockwood, H., & Wiesenfeld, K. (2002). Huygens's clocks. *Proceedings: Mathematics, Physical and Engineering Sciences*, 563-579.

Benoit, C. (1986). Note on the use of correlation in speech timing. *Journal of the Acoustical Society of America 80*, 1846-1849.

Bernstein, N. A. (1967). *The co-ordination and regulation of movements*. Oxford: Pergamon Press.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development, 4*(0), 247-260.

Bingham, G. P. (2004). A perceptually driven dynamical model of bimanual rhythmic movement (and phase perception). *Ecological Psychology 16*(1), 45-53.

Birkholz, P., Kröger, B. J. and Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonantal-vowel sequences. *IEEE Audio, Speech and Lang. Process 19*.

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

Blevins, J. (2001). The syllable in phonological theory. In J. Goldsmith (Ed.), *Handbook of Phonological Theory* (pp. 206-244). Cambridge, MA: Blackwell.

Blevins, J. (2003). *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language 37*, 83-96.

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

Bradley, D. C., Sánchez-Casas, R. M., & García-Albea, J. E. (1993). The status of the syllable in the perception of Spanish and English. *Language and Cognitive Processes, 8*(2), 197-233.

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

Browman, C. P., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. *Papers in laboratory phonology I: Between the grammar and physics of speech*, 341-376.

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

Browman, C. P., & Goldstein, L. (1992b). Targetless schwa: an articulatory analysis. In R. Ladd (Ed.), *Papers in Laboratory Phonology II: Gesture, segment, prosody* (pp. 26-36): Cambridge University Press.

Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in American English. In F. Bell-Berti & L. Raphael (Eds.), *Producing speech: Contemporary issues* (pp. 19-33). New York: American Institute of Physics.

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

Buchanan, J. J. and Ryu, Y. U. (2005). The interaction of tactile information and movement amplitude in a multijoint bimanual circle-tracing task: Phase transitions and loss of stability. *The Quarterly Journal of Experimental Psychology Section A 58*(5), 769-787.

Byrd, D. and Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics 31*, 149–180.

Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics, 19*, 37-47.

Carney, P. J. and Moll, K. L. (1971). A cinefluorographic investigation of fricative consonant-vowel coarticulation. *PHON 23*, 193-202.

Caspers, J. and van Heuven, V. J. (1993). Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica 50*, 161-171.

Chakraborty, M. and Jarvis, E. D. (2015). Brain evolution by brain pathway duplication. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 370*(1684).

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

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

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

Cheng, C. and Xu, Y. (2013). Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America 134*(6), 4481-4495.

Chiosáin, M. N., Welby, P. and Espesser, R. (2012). Is the syllabification of Irish a typological exception? An experimental study. *Speech Communication 54*(1), 68-91.

Chiu, F., Fromont, L., Lee, A., & Xu, Y. (2015). Long-distance anticipatory vowel-to-vowel assimilatory effects in French and Japanese. *Proceedings of the 18th International Congress of Phonetic Sciences* (pp. 1008-1012), Glasgow, UK.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Clements, G. N. (1976). Vowel harmony in nonlinear generative phonology. *Bloomington: Indiana University Linguistics Club*.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the grammar and the physics of speech* (pp. 283-333). Cambridge: Cambridge University Press.

Clumeck, H. (1976). Patterns of soft palate movements. *Journal of Phonetics 4*, 337-351.

Content, A., Kearns, R. K. and Frauenfelder, U. H. (2001). Boundaries versus Onsets in Syllabic Segmentation. *Journal of Memory and Language 45*(2), 177-199.

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences, 37*, 318-325.

Cowie, R. and Douglas-Cowie, E. (1992). Postlingually acquired deafness. In *Trends in linguistics, studies and monographs*. New York, Mouton de Gruyter. *62,* pp.

Cummins, F. (2011). Periodic and Aperiodic Synchronization in Skilled Action. *Frontiers in Human Neuroscience 5*(170).

Cummins, F. and Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics 26*, 145-171.

Cummins, F., Li, C. and Wang, B. (2013). Coupling among speakers during synchronous speaking in English and Mandarin. *Journal of Phonetics 41*(6), 432-441.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*(4), 385-400.

de Jong, K. J. (1994). The correlation of P-center adjustments with articulatory and acoustic events. *Perception & Psychophysics 56*(4), 447-460.

de Jong, K. (2001). Rate-induced resyllabification revisited. *Language and Speech, 44*, 197-216.

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

de Saussure, F. (1916). Nature of the Linguistics Sign. In C. Bally & A. Sechehaye (Eds.), *Cours de linguistique générale*: McGraw Hill Education.

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

Delattre, P. C., Liberman, A. M. and Cooper, F. S. (1955). Acoustic Loci and Transitional Cues for Consonants. *Journal of the Acoustical Society of America 27*(4), 769-773.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language 27*, 124-142.

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

Duanmu, S. (1993). Rime length, stress, and association domains. *Journal of East Asian Linguistics 2*, 1-44.

Duanmu, S. (2008). *Syllable structure: The limit of variation*: Oxford University Press.

Esposito, C. M. (2010). Variation in contrastive phonation in Santa Ana del Valle Zapotec. *Journal of the International Phonetic Association 40*(02), 181-198.

Fitch, W. T. (2010). *The evolution of language*: Cambridge University Press.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research 46*, 127-139.

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

Fowler, C. A. and Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and speech 36*(2-3), 171-195.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In Butterworth (Ed.), *Language Production* (pp. 373-420). New York: Academic Press.

Fowler, C. A., Whalen, D. H. and Cooper, A. M. (1988). Perceived timing is produced timing: A reply to Howell. *Attention, Perception, & Psychophysics 43*, 94-98.

Fox, B. and Routh, D. (1975). Analyzing spoken language into words, syllables, and phonomes: A developmental study. *Journal of Psycholinguistic Research 4*(4), 331-342.

Frota, S. (2002). Tonal association and target alignment in European Portuguese nuclear falls. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 387-418). Berlin: Mouton de Gruyter.

Fudge, E. C. (1969). Syllables. *Journal of Linguistics 5*, 253-86.

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

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

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

Gafos, A. I., & Benus, S. (2006). Dynamics of phonological cognition. *Cognitive science, 30*(5), 905-943.

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

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

Gao, H. and Xu, Y. (2013). Coarticulation as an epiphenomenon of syllable-synchronized target approximation—Evidence from F0-aligned formant trajectories in Mandarin. *Journal of the Acoustical Society of America 135, Pt. 2*.

Gao, M. (2008). *Mandarin Tones: An Articulatory Phonology Account.* PhD. Dissertation. Yale University.

Gay, T. J. (1968). Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America 44*, 1570-1573.

Gelfer, C. E., Bell-Berti, F. and Harris, K. S. (1989). Determining the extent of coarticulation: effects of experimental design. *Journal of the Acoustical Society of America 86*(6), 2443-2445.

Gick, B. (2003). Articulatory correlates of ambisyllabicity in English glides and liquids. In J. Local, R. Ogden & R. Temple (Eds.), *Papers in Laboratory Phonology VI: Constraints on Phonetic Interpretation* (pp. 222-236). Cambridge: Cambridge University Press.

Gimson, A. C. (1970). *An introduction to the pronunciation of English*. London: Arnold.

Gnanadesikan, A. E. (2010). Syllables and syllabaries: What writing systems tell us about syllable structure. In *Handbook of the Syllable* (pp. 395-414): Brill.

Goldstein, L. M., & Fowler, C. (2003). Articulatory phonology: a phonology for public language use. In A. S. Meyer & N. O. Schiller (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 159-207). Berlin: Mouton de Gruyter.

Goslin, J. and Frauenfelder, U. H. (2001). A Comparison of Theoretical and Human Syllabification. *Language and Speech 44*(4), 409-436.

Gu, W. and Lee, T. (2007). Effects of tonal context and focus on Cantonese F0. In *Proceedings of The 16th International Congress of Phonetic Sciences* (pp. 1033-1036), Saarbrucken, Germany.

Gu, W., Hirose, K. and Fujisaki, H. (2007). Analysis of Tones in Cantonese Speech Based on the Command-Response Model. *Phonetica 64*, 29-62.

Haken, H., Kelso, J. A. S. and Bunz, H. (1985). A Theoretical Model of Phase Transitions in Human Hand Movements. *Biological Cybernetics 51*, 347-356.

Hanson, H. M. and Stevens, K. N. (2002). A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *Journal of the Acoustical Society of America 112*, 1158-1182.

Hieke, A. E. (1984). Linking as a marker of fluent speech. *Language and Speech, 27*(4), 343-354.

Hoard, J. E. (1971). Aspiration, tenseness, and syllabification in English. *Language 47*, 133-40.

Holmes, J. N., Mattingly, I. G. and Shearme, J. N. (1964). Speech Synthesis by Rule. *Language and Speech 7*(3), 127-143.

Hooper, J. B. (1972). The syllable in phonological theory. *Language*, 525-540.

Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Attention, Perception, & Psychophysics 43*, 90-93.

Howie, J. M. (1974). On the domain of tone in Mandarin. *Phonetica 30*, 129-148.

hubpages.com (2014) How to Hit a Great Smash in Badminton. Accessed 9 September 8, 2016 from http://hubpages.com/games-hobbies/Badminton-Smash-How-to-Play-the-Shot# (courtesy of Michael Hayes at HowTheyPlay.com)

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP-96* (pp. 373-376), Atlanta, Georgia.

Huygens, in Ouevres Completes de Christian Huyghens, edited by M. Nijhoff (Societe Hollandaise des Sciences, The Hague, The Netherlands, 1893), Vol. 5, p. 243 (a letter to his father, dated 26 Feb. 1665).

Iskarous, K., Fowler, C. A. and Whalen, D. H. (2010). Locus equations are an acoustic expression of articulator synergy. *The Journal of the Acoustical Society of America 128*(4), 2021-2032.

Ivry, R., Diedrichsen, J., Spencer, R., Hazeltine, E., & Semjen, A. (2004). A Cognitive Neuroscience Perspective on Bimanual Coordination and Interference. In S. P. Swinnen & J. Duysens (Eds.), *Neuro-Behavioral Determinants of Interlimb Coordination: A multidisciplinary approach* (pp. 259-295). Boston, MA: Springer US.

Jarvis, E. D. (2004). Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences 1016*(1), 749-777.

Johansson, R. S. and Flanagan, J. R. (2009). Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nat Rev Neurosci 10*(5), 345-359.

Jakobson, R., Fant, C. G. and Halle, M. (1951). *Preliminaries to Speech Analysis. The distinctive features and their correlates*. Cambridge, MA: MIT Press.

Jones, D. (1932). *Outline of English Phonetics*. Cambridge: Cambridge University Press.

Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A. d. (1999). Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication 27*, 187-207.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology 9*, 718-727.

Keating, P. A. (1988). Underspecification in phonetics. *Phonology 5*, 275-292.

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

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

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

Kelso, J. A. S., Tuller, B., & Harris, K. S. (1983). A "dynamic pattern" perspective on the control and coordination of movement. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 137-173). New York: Springer-Verlag.

Kelso, S. J. A., Fink, P. W., DeLaplain, C. R. and Carson, R. G. (2001). Haptic information stabilizes and destabilizes coordination dynamics. *Proceedings of the Royal Society of London B: Biological Sciences 268*(1472), 1207-1213.

Kent, R. and Minifie, F. (1977). Coarticulation in recent speech production models. *Journal of Phonetics 5*, 115-133.

Kent, R. D. and Moll, K. L. (1972). Tongue body articulation during vowel and diphthong gestures. *Folia phoniat 24*, 278-300.

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.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America 67*, 971-995.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America 82*, 737-793.

Koh, K., Kwon, H. J., Yoon, B. C., Cho, Y., Shin, J.-H., Hahn, J.-O., Miller, R. H., Kim, Y. H. and Shim, J. K. (2015). The role of tactile sensation in online and offline hierarchical control of multi-finger force synergy. *Experimental Brain Research 233*(9), 2539-2548.

Kohler, K. J. (1966). Is the syllable a phonological universal? *Journal of Linguistics 2*(02), 207-208.

Kovacs, A. J., Buchanan, J. J. and Shea, C. H. (2010). Impossible is nothing: 5:3 and 4:3 multi-frequency bimanual coordination. *Experimental Brain Research 201*(2), 249-259.

Kozhevnikov, V. A. and Chistovich, L. A. (1965). *Speech: Articulation and Perception*. Washington, DC: Translation by Joint Publications Research Service. JPRS 30543.

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

Kreiman, J. and Gerratt, B. R. (2010). Effects of native language on perception of voice quality. *Journal of phonetics 38*(4), 588-593.

Kubikova, L., Bosikova, E., Cvikova, M., Lukacova, K., Scharff, C. and Jarvis, E. D. (2014). Basal ganglia function, stuttering, sequencing, and repair in adult songbirds. *Scientific Reports 4*, 6590.

Kugler, P. N., Scott Kelso, J. A., & Turvey, M. T. (1980). On the Concept of Coordinative Structures as Dissipative Structures: I. Theoretical Lines of Convergence. In E. S. George & R. Jean (Eds.), *Advances in Psychology* (Vol. Volume 1, pp. 3-47): North-Holland.

Kuhl, P. K. and Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America 100*(4), 2425-2438.

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

Laboissiere, R., Ostry, D. J., & Feldman, A. G. (1996). The control of multi-muscle systems: human jaw and hyoid movements. *Biological Cybernetics, 74*, 373-384.

Labrune, L. (2012). Questioning the universality of the syllable: evidence from Japanese. *Phonology, 29*(01), 113-152.

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

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

Ladefoged, P. (l982). *A course in phonetics*: University of California, Los Angeles.

Lane, H. and Webster, J. (1991). Speech deterioration in postlingually deafened adults. *Journal of the Acoustical Society of America 89*, 859–866.

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

Lee, A., Prom-on, S. and Xu, Y. (in press). Pre-low raising in Japanese pitch accent. To appear in *Phonetica*.

Lehiste, I. (1960). An acoustic–phonetic study of internal open juncture. *Phonetica, 5* (Suppl. 1), 5-54.

Levelt, W. J. M., Roelofs, A. and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences 22*(1), 1-38.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. G. (1967). Perception of the speech code. *Psychological Review 74*, 431-461.

Liberman, I. Y., Shankweiler, D., Fischer, F. W. and Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of experimental child psychology 18*(2), 201-212.

Lin, M. (1995). A perceptual study on the domain of tones in Standard Chinese. *Chinese Journal of Acoustics 14*, 350-357.

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

Lindblom, B. and Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics 40*(1), 1-19.

Ling, B. and Liang, J. (2015). Tonal alignment in shanghai chinese. In *Proceedings of COCOSDA2015,* 128-132.

Ling, Z.-H., Richmond, K., Yamagishi, J., & Wang, R.-H. (2009). Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(6), 1171-1185.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384-422.

Liu, H., Lu, H., Shao, X. and Xu, Y. (2016). Model-based Parametric Prosody Synthesis with Deep Neural Network. In *Proceedings of Interspeech 2016,* 2313-2317.

Locke, J. L. (1983). *Phonological Acquisition and Change*. London: Academic Press.

Löfqvist, A. (1991). Proportional timing in speech motor control. *Journal of Phonetics 19*, 343-350.

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

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

Mechsner, F., Kerzel, D., Knoblich, G. and Prinz, W. (2001). Perceptual basis of bimanual coordination. *Nature 414*, 69-73.

Mechsner, F., Stenneken, P., Cole, J., Aschersleben, G. and Prinz, W. (2007). Bimanual circling in deafferented patients: evidence for a role of visual forward models. *journal of Neuralphsiology 1*, 259-282.

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

Miller, J. L., Grojean, F. and Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica 41*, 215-225.

Moll, K. and Daniloff, R. (1971). Investigation of the timing of velar movement during speech. *Journal of the Acoustical Society of America 50*, 678-684.

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

Munhall, K. G. (1985). An examination of intra-articulator relative timing. *Journal of the Acoustical Society of America 78*, 1548-1553.

Myers, S. (1998). Surface underspecification of tone in Chichewa. *Phonology 15*, 367-392.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E. and Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *The Journal of the Acoustical Society of America 132*(6), 3980-3989.

Nolan, F. and Asu, E. L. (2009). The Pairwise Variability Index and Coexisting Rhythms in Language. *Phonetica 66*(1-2), 64-77.

Ohala, J. J. (1992). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In *Papers from the Parasession on the Syllable* (pp. 319-338). Chicago: Chicago Linguistic Society.

Ohala, J. J. (1994). *Towards a universal, phonetically-based, theory of vowel harmony.* Paper presented at the ICSLP, Yokohama.

Ohala, J. J. and Kawasaki, H. (1984). Prosodic phonology and phonetics. *Phonology 1*, 113-127.

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

Ostry, D. J., Gribble, P. L. and Gracco, V. L. (1996). Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? *Journal of Neurosicence 16*, 1570-1979.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics 25*(5), 382-407.

Perkell, J. S. and Chiang, C.-M. (1986). Preliminary supposrt for a "hybrid model" of anticipatory coarticulation. In *Proceedings of The 12th International Congress of Acoustics*, Toronto. Canadian Acoustical Association*,* A3-6.

Perkell, J., M., M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J. and Guiod, P. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication 22*, 227-249.

Peterson, G. E., Wang, W. S. Y. and Sivertsen, E. (1958). Segmentation Techniques in Speech Synthesis. *The Journal of the Acoustical Society of America 30*(8), 739-742.

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

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

Pierrehumbert, J. and Beckman, M. (1988). *Japanese Tone Structure*. Cambridge, MA: The MIT Press.

Pinker, S. (1995). *The language instinct: The new science of language and mind* (Vol. 7529): Penguin UK.

Prom-on, S., Birkholz, P. and Xu, Y. (2013). Training an articulatory synthesizer with continuous acoustic data. In *Proceedings of Interspeech 2013,* 349-353.

Prom-on, S., Birkholz, P. and Xu, Y. (2014). Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *EURASIP Journal on Audio, Speech, and Music Processing 2014*(1), 1-11.

Prom-on, S., Xu, Y. and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America 125*(1), 405-424.

Pulgram, E. (1970). *Syllable, word, nexus, cursus*. The Hague: Mouton.

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

Recasens, D. (1984b). V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study. *Journal of Phonetics 12*, 61-73.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic bulletin & review 12*(6), 969-992.

Ridderikhoff, A., Peper, C. E. and Beek, P. J. (2007). Error correction in bimanual coordination benefits from bilateral muscle activity: evidence from kinesthetic tracking. *Experimental Brain Research 181*, 31-48.

Ringel, R. L. and Ewanowski, S. J. (1965). Oral Perception: 1. Two-Point Discrimination. *Journal of Speech, Language, and Hearing Research 8*(4), 389-398.

Saitou, T., Goto, M., Unoki, M. and Akagi, M. (2009). Speech-to-Singing Synthesis System: Vocal Conversion from Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices. NCMMSC2009.

Saitou, T., Unoki, M. and Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Communication 46*(3-4), 405-417.

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

Schiller, N. O., Meyer, A. S. and Levelt, W. J. (1997). The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants. *Language and Speech 40*(2), 103-140.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological review 82*(4), 225-260.

Schmidt, R. C., Carello, C. and 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.

Scripture, E. (1902). *The Elements of Experimental Phonetics*. New York: Charles Scribners Sons.

Selkirk, E. O. (1982). The syllable. In H. v. d. Hulst & N. Smith (Eds.), *The structure of phonological representations, part II* (pp. 337-383). Dordrecht: Foris.

Selkirk, E. O. (1984). On the major class features and syllable theory. In M. Aronoff & R. T. Oehrle (Eds.), *Language Sound Structure: Studies in Phonology* (pp. 107-136). Cambridge: MIT Press.

Shattuck-Hufnagel, S. (2010). The role of the syllable in speech production in american english: A fresh consideration of the evidence. In *Handbook of the Syllable* (pp. 195-224): Brill.

Spencer, R. M. C., Ivry, R. B., Cattaert, D. and Semjen, A. (2005). Bimanual coordination during rhythmic movements in the absence of somatosensory feedback. *journal of Neuralphsiology 94*, 2901-2910.

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

Stagray, J. R., Downs, D. and Sommers, R. K. (1992). Contributions of the fundamental, resolved harmonics, and unresolved harmonics in tone-phoneme identification. *Journal of Speech, Language, and Hearing Research 35*(6), 1406-1409.

Steriade, D. (1982). *Greek prosodies and the nature of syllabification*, PhD. Dissertation, Massachusetts Institute of Technology.

Steriade, D. (1995). Underspecification and markedness. In J. A. Goldsmith (Ed.), *Handbook of Phonological Theory* (pp. 114-174). Oxford: Basil Blackweell.

Steriade, D. (1999). Alternatives to syllable-based accounts of consonantal phonotactics. In O. Fujimura, B. D. Joseph & B. Palek (Eds.), *Proceedings of linguistics and phonetics 1998: Item order in language and speech* (pp. 205-245). Prague: Karolinum Press.

Stetson, R. H. (1951). *Motor Phonetics:A study of Speech Movements in Action*. Amsterdam: North Holland.

Stevens, K. N. and Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics* **38**: 10-19.

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

Sussman, H. M., McCaffrey, H. A. and Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America 90*(3), 1309-1325.

Swinnen, S. P. and Wenderoth, N. (2004). Two hands, one brain: cognitive neuroscience of bimanual skill. *Trends in Cognitive Sciences 8*(1), 18-25.

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

Tanakaa, M., Alvarado, J. S., Murugan, M. and Mooney, R. (2016). Focal expression of mutant huntingtin in the songbird basal ganglia disrupts cortico-basal ganglia networks and vocal sequences. *Proceedings of the National Academy of Sciences 113*(12), E1720–E1727.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America 107*, 1697-1714.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

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

Tokuda, K., Kobayashi, T. and Imai, S. (1995). Speech parameter generation from HMM using dynamic features. In *Proceedings of International Conference on Acoustics Speech and Signal Processing 1995,* 660-663.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K. (2013). Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE 101*(5), 1234-1252.

Tokuda, K., Zen, H. and Black, A. W. (2002). An HMM-based speech synthesis system applied to English. In *Proceedings of Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002,* 227- 230.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language, 27*, 87-104.

Tsao, Y.-C. and Weismer, G. (1997). Interspeaker Variation in Habitual Speaking Rate, Evidence for a Neuromuscular Component. *Journal of Speech, Language, and Hearing Research 40*(4), 858-866.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic Segment Durations in Prosodic Research: A Practical Guide. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter & J. SchlieBer (Eds.), *Methods in Empirical Prosody Research* (pp. 1-28). Berlin, New York: De Gruyter.

van Santen, J. P. H. and Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *Journal of the Acoustical Society of America 107*, 1012-1026.

Vennemann, T. (1988). *Preference laws for syllable structure and the explanation of sound change*. Berlin: Mouton de Gruyter.

Wang, W. S. Y. (1967). Phonological features of tone. *International Journal of American Linguistics 33*, 93-105.

Wells, J. C. (1990). Syllabification and allophony. In S. Ramsaran (Ed.), *Studies in the pronunciation of English: A commemorative volume in honour of A. C. Gimson* (pp. 76-86). London: Routledge.

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

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics 18*, 3-35.

Wilson, A. D., Bingham, G. P. and Craig, J. C. (2003). Proprioceptive perception of phase variability. *Journal of Experimental Psychology: Human Perception & Performance 29*, 1179-1190.

Wilson, A. D., Collins, D. R. and Bingham, G. P. (2005). Perceptual coupling in rhythmic movement coordination: stable perception leads to stable action. *Experimental Brain Research 164*(4), 517-528.

Wong, Y. W. (2006). Contextual Tonal Variations and Pitch Targets in Cantonese. In *Proceedings of Speech Prosody 2006*, Dresden, Germany*,* PS3-13-199.

Wood, S. A. J. (1996). The temporal coordination of articulator gestures. In *Proceedings of the First ETRW on Speech Production Modeling*, Autrans*,* 61-64.

Xu, C. X. and Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association 33*, 165-181.

Xu, Y. (1986). Acoustic-phonetic characteristics of junctures in Mandarin Chinese [in Chinese]. *Zhongguo Yuwen [Journal of Chinese Linguistics]*(4), 353-360.

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

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

Xu, Y. and Prom-on, S. (2010-2016). PENTAtrainer1.praat. Available from: http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/.

Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication 57*, 181-208.

Xu, Y. and 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. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication 33*(4), 319-337.

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

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

Yang, S. and Xu, Y. (1988). Acoustic-phonetic Oriented Chinese Speech Synthesis Technology. *Speech Communication 7*(3), 317-325.

Zen, H., Tokuda, K. and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication 51*(11), 1039-1064.

Zhang, Y. (2011). An Investigation of the Acquisition of Linking by Chinese EFL Learners. *HKBU Papers in Applied Language Studies 15*, 42-64.