Title
Multi-timescale Dynamical Interactions Between Speech Rhythm and Gesture

Journal
UC Berkeley PhonLab Annual Report, 4(4)

ISSN
2768-5047

Author
Tilsen, Sam

Publication Date
2008

DOI
10.5070/P759m4g8jj

Multi-timescale dynamical interactions between speech rhythm and gesture

Sam Tilsen

Abstract

Temporal patterns in speech occur on multiple timescales. Regularities in such patterns have been observed between gestures, which are relatively quick, and between rhythmic units, which are slower. Previous work has shown that patterns in both domains can be modeled with oscillatory dynamical systems. This paper builds upon prior work by investigating how gestural and rhythmic patterns interact. An experiment was conducted in which speakers repeated the phrase “take on a spa” to a metronome with several different rhythms, some of which were more difficult to produce than others. Gestural kinematics were recorded using electromagnetic articulometry, allowing for both rhythmic and gestural variability to be measured. Relative timing of jaw, lip, and tongue movements associated with the [sp] cluster in the word “spa” were analyzed. Significant relations were found between rhythmic variance and gestural variance. In addition, within-gesture effector synergies differed systematically across rhythmic conditions. These results demonstrate an interaction between rhythmic and gestural systems. A dynamical model of phase-coupled oscillators is presented, which can simulate the observed variability patterns.

1. Introduction

Speech is an interesting motor behavior because it involves systems whose dynamics are associated with different timescales. These systems are normally referred to—from shortest to longest timescale—as gestures, moras, syllables, feet, and phrases. The gestural systems govern articulator movements, such as the raising of the tongue for an [s] or raising of the lower lip for a [p]. The rhythmic systems coordinate groups of gestures or other rhythmic systems, e.g. a metric foot contains multiple syllables. Rhythmic systems generally exert their influence over longer periods of time than gestural systems. The distinction between gestural and rhythmic domains constitutes one source of complexity in speech. Another important source of complexity is the use of multiple effectors to accomplish a single gestural goal, e.g. the simultaneous, synergistic use of the jaw and lower lip to achieve the bilabial closure for [p]. This paper addresses experimentally the question of how gestural systems interact with rhythmic systems, and whether synergistic relations are influenced by speech rhythm.

Both gestural and rhythmic systems have been described independently with dynamical models by various researchers, and in subsequent sections we will review some of this work. Based on this prior research, a natural question to ask is whether these two domains interact dynamically. If such interactions occur, then experimental perturbations of speech rhythm may have two primary effects upon gestures: 1) to influence the relative timing of the movements associated with different gestures, and 2) to influence the synergistic relations between the effectors used to accomplish a single gesture.

In many domains of motor control, we observe rhythmic movements involving multiple goals and effectors, and a useful tool in the analysis of such movements is the concept of relative phase. To better understand walking, for example, we can analyze the difference in phases of the limbs. This approach reveals that the relative phase (or phase difference) in normal walking is nearly constant. Why go to the trouble of defining a relative phase, when we can use separate phases for each limb? The advantage of the relative phase is that it reduces the description of the system to a single variable. Reducing the number of dimensions needed to describe a self-organizing dynamical system is a common technique in the physical sciences (Haken 1983, 1993), and has been employed with success in models of biological systems (Winfree 1980), including human behavior (Haken, Kelso, & Bunz 1985).

Using lower-dimensional descriptions allows us to more readily conceptualize the behavior of a complex system, and can provide deeper insight into principles underlying its organization. For example, in a classic study, Haken, Kelso, & Bunz (1985) found that as movement frequency increases, anti-phase finger-wagging movements become less stable than in-phase finger wagging. Both in-phase and anti-phase finger wagging are much more stable than any other relative phasing of finger movements. The stability of these two regimes is a profound finding, and has been applied insightfully to modeling the coordination of speech systems.

Walking is a relatively simple example because we do not have to consider the possibility that one leg can move through multiple cycles in the time the other moves through just one cycle: i.e. the left leg does not take two steps in the time that the right leg takes one. In other words, the dynamics of the component systems adhere to the same timescale. In contrast, finger-wagging and other forms of manual coordination can involve multiple frequencies (c.f. Haken et. al. 1996), and this requires a more general notion of relative phase. Such studies have revealed that that low-order frequency ratios in multifrequency (or multi-timescale) systems are more stable than higher-order ratios.

Speech coordination is also more complex than locomotion because any adequate dynamical description of the system must involve a hierarchy of timescales. What exactly are these timescales? The shortest one encompasses speech gestures, and important work has been done exploring the idea that the lexical specification of speech involves target relative phases of gestures derived from coupled systems (Goldstein & Browman 1990; Saltzman 1986; Saltzman & Munhall 1989; Saltzman & Nam 2003). Dynamical treatments of rhythmic (or prosodic/metrical) units, such as moras, syllables, feet, and phrases, have also been proposed (O'Dell & Nieminen 1999; Cummins & Port 1998; Barbosa 2002). These descriptions of rhythmic systems require a more complicated notion of generalized relative phase, in which individual oscillator phases are converted to a common timescale before their difference is calculated (c.f. Byrd & Saltzman 2000, Kopell 1988; Pikovsky et. al. 2001). Generalized relative phase allows for a straightforward extension of the concept of phase coupling to oscillatory systems with frequencies related by approximately integer ratios. It provides for a more intuitive description of multifrequency rhythmic systems, and for a potential integration of rhythmic and gestural dynamical models.

If short-timescale gestural systems and longer-timescale rhythmic systems do not interact with each other, then their dynamics can be understood independently and there should be nothing particularly interesting about the behavior of the system as whole that cannot be learned from studying its parts. It is worth pointing out that there is no a priori reason for such interactions to occur, although some sort of trivial law relating the two seems necessary. Absence of an interaction would entail, for example, that the relative timing of tongue and lip gestures in an [sp] cluster in the word "spa" is unaffected by the rhythmic context in which these gestures are articulated.

If to the contrary, rhythmic and gestural systems do interact, then intergestural timing should be influenced by the rhythmic context in which the gestures are produced. In that case the description of the system becomes more complex—particularly because the gestures occur on a faster timescale than feet and phrases. There are several indirect reasons to presuppose that rhythmic and gestural systems interact. For one, there are correlations between patterns of deletion/reduction of speech gestures and the typological classes of stress- and syllable-timed languages (Ramus et. al. 1999, Dauer 1983), and there are also indications that segmental deletion is associated with more rhythmic speech. However, to date there has been no conclusive demonstration that rhythmic and gestural systems interact within a given utterance. The present study provides such evidence.

1.1 Multifrequency coupling between feet and phrases

Multiscale dynamics involving rhythmic systems were discovered by Cummins & Port (1996) and (1998), who used a *speech cycling task* to probe the relative timing between phrases and feet. In this task, subjects hear a high-low two tone metronome pattern and repeat a phrase (e.g. *big for a duck*) so that the first and last stressed syllables of the phrase (i.e. *big* and *duck*) align with the metronome beats. After a number of metronome pattern repetitions, the metronome fades out while subjects continue repeating the phrase, trying to maintain the metronome rhythm. The controlled variable in this design is the target phase (Φ) of the second (L) metronome tone relative to a phrasal period defined by successive first (H) tones. Target phases were 0.3, 0.4, 0.5, 0.6, and 0.7. Fig. 1 schematizes the design of the Cummins & Port (1998) experiment, in which the H to L tone interval is fixed at 700 ms and the phrasal period is varied.

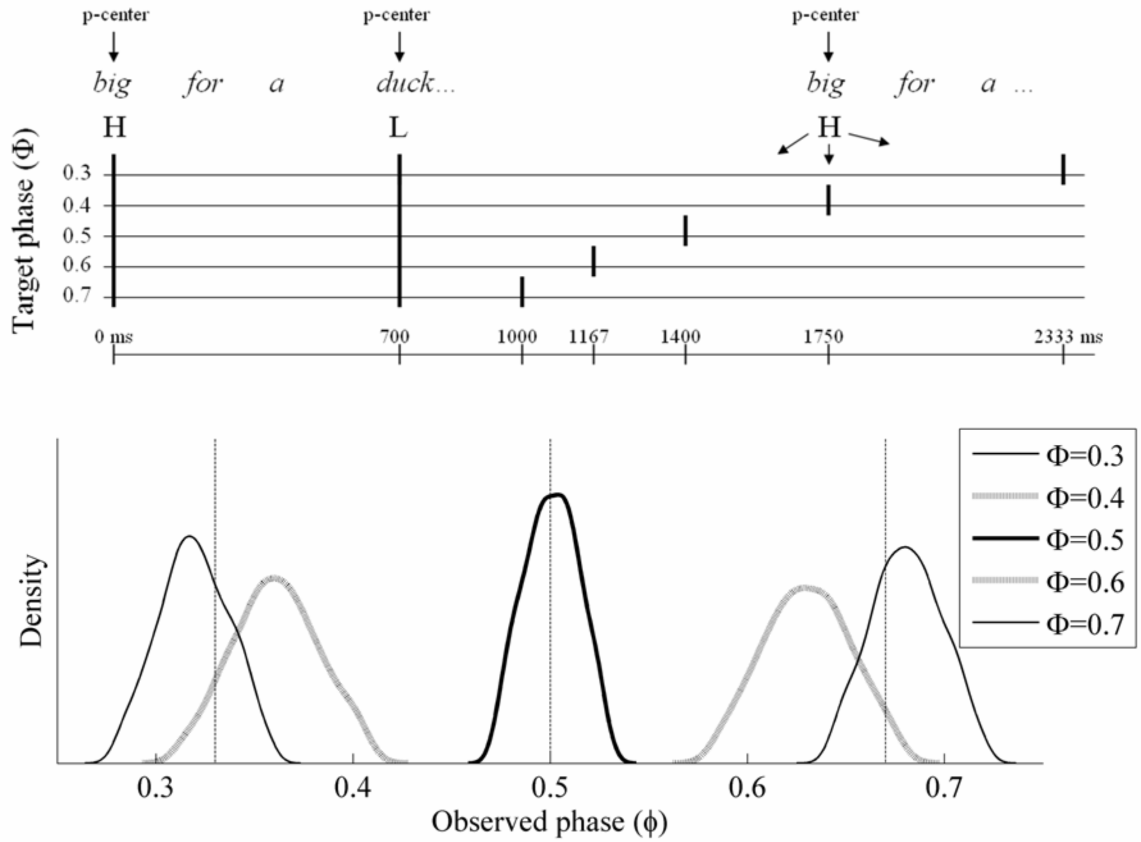


Fig. 1. Schematic illustration of speech cycling task design (top) and simulated relative phase distributions for each target phase condition (Φ) (bottom), exhibiting a harmonic timing effect. Biases in observed phases toward harmonic ratios (dashed vertical lines) are evident in the simulated distributions, as well as increased variability in the higher-order target phase conditions. See text for further details.

With this design, Cummins and Port discovered a *harmonic timing effect*: productions of the second stressed syllable were biased toward phases of the phrasal period that were close to low-order integer ratios: 1/3, 1/2, and 2/3. Observed phases were measured by approximating the locations of p-centers, which are perceptually salient instants associated with syllables (Allen 1972, 1975; Morton, Marcus, & Frankish 1976; Howell 1988; Pompino-Marschall 1989). In Fig. 1, the observed relative phase φ is the ratio of the interval between p-centers of *big* and *duck* to the interval between successive p-centers of *big*. Cummins and Port found that in each target phase condition, the distributions of phases were biased toward the nearest low-order harmonic ratio. This is shown in Fig. 1 by a shift of each distribution toward the nearest low-order ratio (dashed vertical lines).

Crucially, variance in produced phase was lower for target phases nearer to the harmonic ratios; hence variability in produced phase was highest in the 0.4 and 0.6 conditions, intermediate in the 0.3 and 0.7 conditions, and lowest in the 0.5 condition. This suggests that the rhythm in the 0.5 target phase condition is easiest to produce, while the other target phases require more difficult rhythms.

Cummins & Port (1998) and Port (2003) suggested that the results of the speech cycling task can be modeled with phase-locked pulses from a multifrequency system of coupled oscillators, where harmonic phrase and foot oscillators are either 1:2 or 1:3 frequency-locked and the foot oscillator produces pulses at phase 0. The oscillators in this model are phase-locked with a generalized relative phase of 0, and stressed syllable p-centers are attracted to the pulses. The basic idea of this approach is that low-order harmonic frequency ratios are more stable than higher-order ones (Haken, Kelso, & Bunz 1985), which explains the harmonic timing biases. Correlations between target phase and variability may arise due to increased competition between less stable 1/3 and 2/3 harmonic attractors and higher-order target rhythms.

The results of the speech-cycling task are interesting because they support the notion that there exist multi-timescale dynamical interactions between linguistic systems, in this case feet and phrases. The Cummins & Port analysis reconceptualizes hierarchical metrical structures as dynamical systems that are responsible for rhythmic coordination. This constitutes a fairly radical departure from the mainstream linguistic understandings of metrical feet and their combinations that were developed in Liberman (1975) and Liberman & Prince (1977).

1.2 Intergestural coupling

Now we focus on the smaller timescale appropriate for intergestural coordination. An important phenomenon here is the *c-center effect*, discovered by Browman & Goldstein (1988). This effect refers to timing patterns observed between multiple onset consonant gestures and vowel gestures within a syllable, e.g. in CCV syllables like *spa*. In syllables with a single onset C, the beginnings of gestures associated with consonant and vowel are approximately in-phase synchronized (note that consonantal gestures are produced more quickly). However, in syllables with complex CC onsets, the beginnings of the consonant gestures are equally displaced in opposite directions from the onset of the vocalic gesture, as schematized in Fig. 2 (c).

Browman & Goldstein (2000) argued that the c-center effect arises from competing lexical specifications of relative phases between pairs of gestures. All consonant gestures in a syllable are specified to be in-phase coordinated with the initial vowel gestures of a syllable, as represented in Fig. 2 (a). In other words, the gestures belonging to each consonant in C_1C_2V have the same target relative phase specification as the gestures associated with the single consonant in CV. However, in C_1C_2V there is an additional anti-phase target between the consonants, as represented in Fig. 2 (b). Both of these targets cannot be simultaneously achieved. Instead they compete, the result being similar to the temporal configuration in Fig. 2 (c). In this case, the moment that ends up being in-phase synchronized with the beginning of the vowel gesture is the *c-center*, which refers to the point halfway between the beginnings of the C gestures.

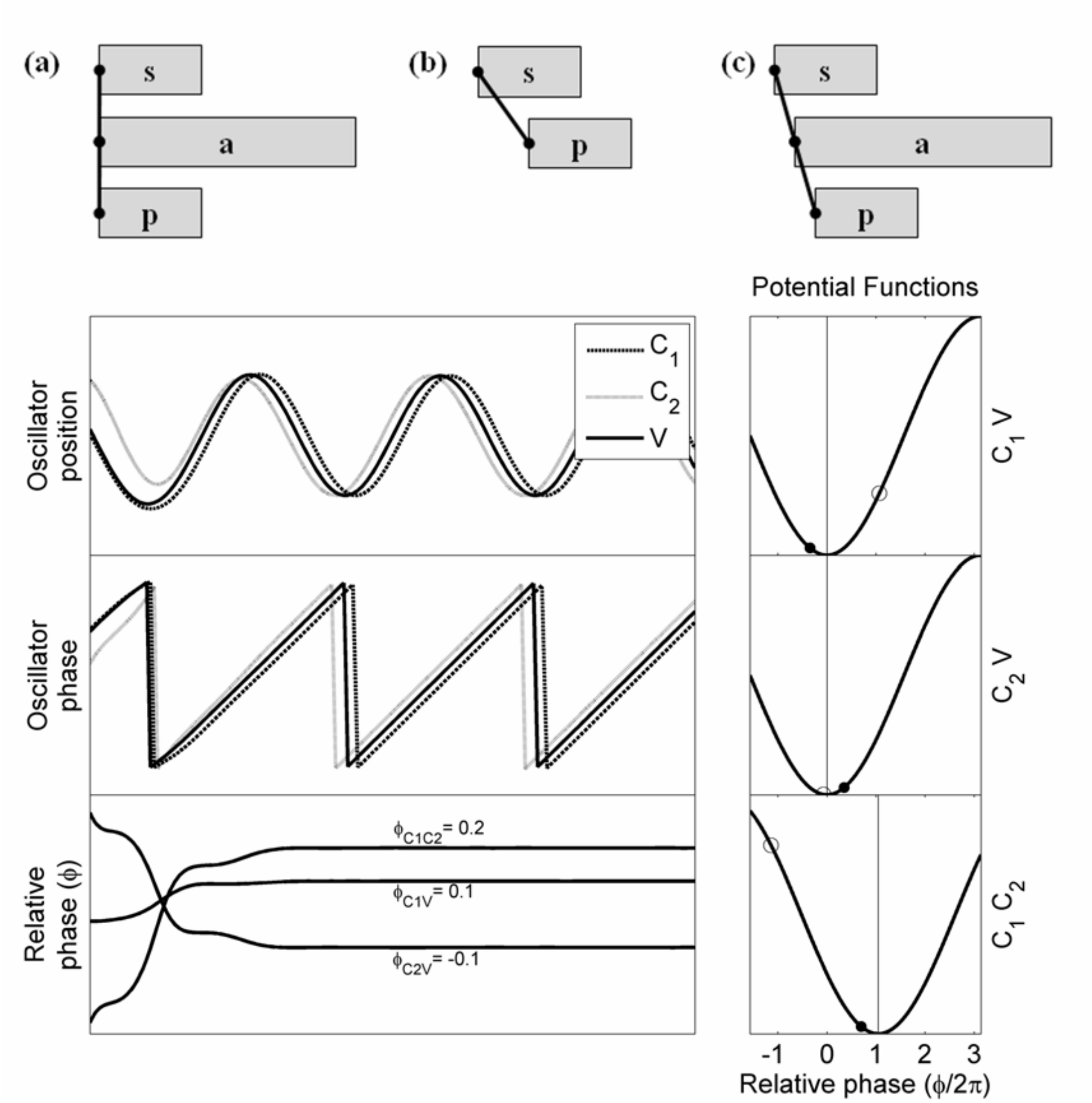


Fig. 2. Coupling graphs and simulation of c-center effect through competitive coupling. Potential functions depict relative phase targets with vertical lines, initial relative phases (randomly chosen) with open circles, and final relative phases with filled dots. Evolutions of oscillator positions, phases, and relative phases are shown on the left. Note that transients die off rather quickly, and that once the system has stabilized, C_1 and C_2 are equally displaced from V in opposite directions.

Saltzman & Nam (2003) simulated the competitive coupling proposed by Browman & Goldstein (2000) using a system of coupled oscillators. Rather than treating the relative phasing of gestures as arbitrarily specified lexical information (as was done in Saltzman & Munhall 1989), the newer model incorporates the idea that target relative phases are stable fixed points of a system of coupled oscillators. These oscillators can be conceptualized as *gestural planning systems*, whose properties are developed partly in Saltzman & Byrd (2000). There is an important distinction here between gestural systems and gestural planning systems. In the task dynamic framework, gestures are modeled as critically-damped mass-spring systems having fixed-point attractors. With simple parameter modifications, gestures can vary in speed, amplitude, and duration. The gestural planning oscillators are very different. They exhibit limit cycle dynamics and their relative phases are governed by potential functions, which are force fields directing change of relative phase. These forces advance or delay the phases of individual oscillators with the effect of decreasing the distance between relative phases and their associated targets. The planning oscillators bear an indirect relation to observable articulator movements. It is the stabilized relative phases of the planning oscillators which in the Saltzman & Nam (2003) model are hypothesized to determine the relative phases of the mass-spring gestural systems.

Fig. 2 shows a simulation of the c-center effect in planning oscillators based upon the Saltzman and Nam model. For each pair of coupled oscillators (and there are three such pairs in a C_1C_2V system), there exists a potential function governing the evolution of the relative phase between the oscillators. Regardless of initial relative phase (here chosen randomly), the system quickly settles into a stable configuration. The stable configuration corresponds to a balance between competing forces associated with the three potential functions. The minimization of all potential energy is accomplished by equal displacement of C gestural phases from phase of the V gesture—this replicates precisely the c-center effect observed in behavioral data.

1.3 Coupling between gestural and rhythmic systems

Given that dynamical coupling can model empirically observed temporal patterns involving feet and phrases, and also gestures, a logical question to ask is whether evidence can be found for a more general model which accounts for patterns on both timescales. A synthesis of the two models requires one to bridge the gap between gestural timing and rhythmic timing. One sensible way to connect these disparate timescales is through a generalized phase coupling. This coupling could be direct—from foot oscillators to gestural oscillators, or could operate through intermediate systems, such as syllables and/or moras.

A coupled-oscillators model of the relative timing of feet and syllables is described in O'Dell and Nieminen (1999). The timespan of the metrical foot is generally reinterpreted in dynamical treatments as the *stress-foot*, which is an interval between stressed syllables. O'Dell and Nieminen model a foot with *n* syllables as a system composed of a stress and syllable oscillator exhibiting 1:*n* frequency-locking. A coupling strength parameter describes the relative strength of the stress and syllable oscillators. Their model can predict durational attributes that correlate with the much-studied distinction between stress and syllable-timed languages (Pike 1945; Abercrombie 1965). In languages in which the stress oscillator dominates, the duration of an interstress interval is longer, due to additional duration in stressed syllables; in languages in which the syllable oscillator dominates, stressed syllables contribute as much as unstressed syllables to interstress interval duration. This accords with the analysis of Erickson (1991), who argued that the distinction between syllable-timing and stress-timing results from additional duration endowed to stressed syllables in stress-timed languages.

The generality of coupling in dynamical models of speech gesture and rhythm begs for the development of a model integrating both domains. One would expect this model to cover the gamut of relevant timescales, from phrases (groups of feet), to feet (intervals between stressed syllables), to syllables, moras, and on the lowest level, gestures and effectors. Fig. 3 illustrates how prosodic units in the canonical hierarchical model of speech are related to oscillatory systems associated with a range of timescales. Of course, not all languages exhibit evidence for systematic patterning on all of these scales; for example, some languages exhibit no evidence of moras, others have no stress. Any satisfactory model should have parameters available to accommodate cross-linguistic variation, although these will not be our concern here. Further, the values of some model parameters may vary from speaker to speaker and from utterance to utterance.

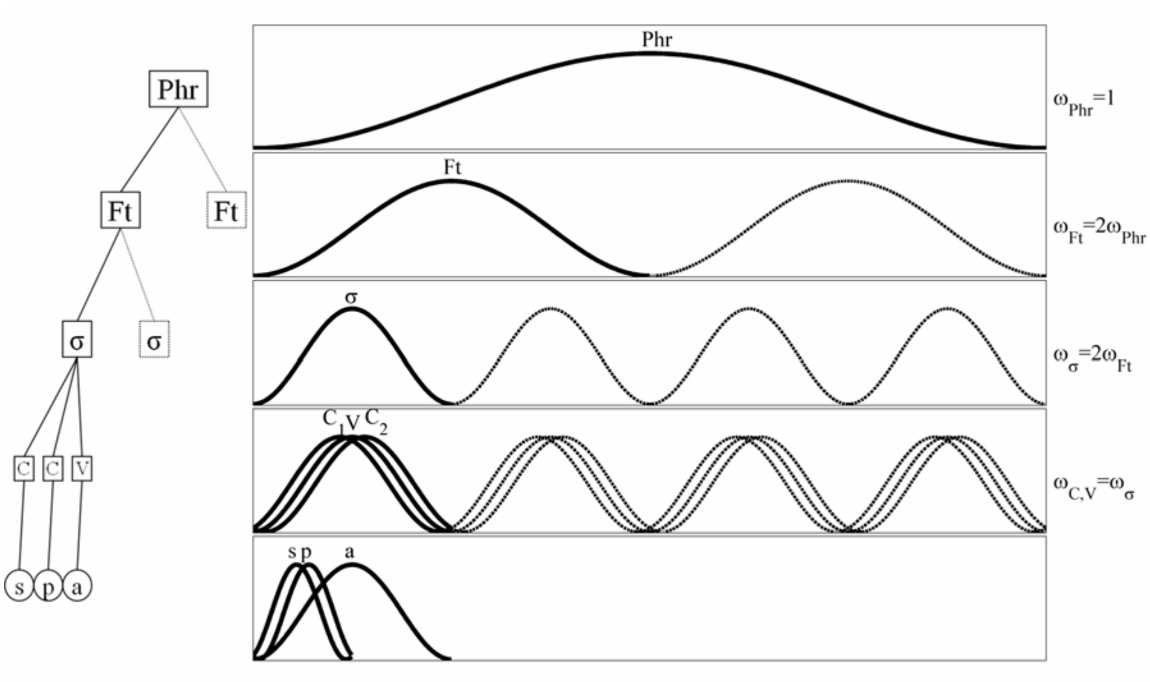


Fig. 3. Relations between a hierarchical model of rhythmic/gestural units and a multiscale dynamical model.

The multi-timescale dynamical model reconceptualizes the units of the traditional model as oscillators, which can be described with differential equations. The “levels” (or types of units) in the traditional model become timescales, which correspond to the inherent frequencies of their associated oscillators. Ideally, these frequencies are related by low-order ratios of integers. Connections between units are represented by coupling terms. Hence the two most fundamental metaphors structuring the canonical model are preserved: CONTAINMENT (of feet within phrases, syllables within feet, etc.) is implicit in the frequency relations between timescales, and CONNECTION (between systems) is a coupling interaction between oscillators, parameterized by a potential function. Note that gestural systems (as opposed to gestural planning systems) are not oscillatory; rather, these systems exhibit mass-spring dynamics, and the relative phases of their onsets are determined by the planning systems.

There are innumerable possibilities for exactly what form the equations describing a multiscale model can take, and in the end this is an empirical question. The issue of how the coupling relations between oscillators should be structured will be considered in section four. Crucially, regardless of how the multiscale dynamical model is formulated, the model predicts the possibility of interactions between disparate timescales. For example, consider repeated productions of the phrase *take on a spa*, in which the words *take* and *spa* are stressed syllables. In addition to rhythmic temporal patterns between the feet and phrase, there are intergestural patterns involving the tongue movement associated with [s] and the lower lip movement associated with [p]. The multiscale dynamical model allows for the possibility of interactions between gestural timing and lower frequency foot and phrase timing. The reasoning behind this is as follows:

Based on the analyses of Cummins & Port (1998) and Port (2003), assume that in more difficult rhythmic conditions, foot-phrase coupling is less stable (i.e. weaker). This assumption is motivated by the observed increases of variability in foot-phrase timing for higher-order target rhythms. It follows that if gestural systems are coupled to rhythmic systems, then intergestural coupling should be more variable when rhythmic coordination is more variable. In the case of *take on a spa*, the relative timing of vertical height maxima associated respectively with the tongue and lip gestures in [s] and [p], will be least variable when the phrase is spoken to a metronome pattern requiring $\Phi_{0.5}$, and will be more variable when the pattern requires $\Phi_{0.3}$, $\Phi_{0.6}$, etc. Exactly how this arises from coupling between oscillators will be discussed in detail later, but the basic principle involves how quickly the forces governing relative phase correct for random perturbations due to noise.

Conventional, non-dynamical models of speech production do not predict differences in temporal patterns due to interaction between prosodic units and gestures. Nor can such models be readily amended to account for such: without dynamic interaction, an [sp] onset cluster should be coordinated the same way regardless of the rhythmic context in which it is produced. The general problem with non-dynamical models is that they provide no mechanism for predicting the temporal patterns of articulations. These models treat gestures as sequential sets of actions that are *linked* or *connected* to higher-level and lower-level units (as in Fig. 3). These linkages themselves offer no provisions for a quantitative characterization of the temporal influences between levels.

Another effect that might arise due to rhythmic-gestural interaction is a change in synergistic relations between the effectors involved in a single gesture. Gestures such as the bilabial closure of [p] involve jaw movement and lower lip movement. However, the extent to which the jaw and lower lip contribute to the closure can vary as a function of the gestural context. The task dynamic model of multi-effector movements in Saltzman & Kelso (1983) and Saltzman (1986) uses an arbitrary weighting matrix to determine the relative contributions of effectors with a common goal. It is conceivable that distinct weighting parameters are associated with differences in coupling between rhythmic and gestural systems—in which case, differences in rhythmic organization may correspond to differences in synergistic relations between effectors.

To test for evidence of rhythmic-gestural interactions, a speech cycling task (as in Cummins & Port 1998) was conducted using the phrase *take on a spa*, while the movements of the jaw, lower lip, and tongue blade were recorded using electromagnetic articulometry. The following specific hypotheses were tested:

H1. *Rhythmic-gestural co-variability*: when rhythmic timing is more variable, intergestural timing between [s] and [p] will also be more variable. This follows from coupling between rhythmic and gestural planning oscillators.

H2. *Rhythmic effects on synergies:* different rhythmic targets will be associated with different within-gesture synergistic patterns between articulators, specifically between the jaw and tongue movements for [s], and between the jaw and lower lip movements for [p].

2. Method

2.1 Task and measurement

8 subjects participated in the experiment, which consisted of a speech cycling task. The same five target rhythms used in Cummins & Port (1998) were used for the current experiment, with the same target relative phases, Φ = [0.3, 0.4, 0.5, 0.6, 0.7]. At the beginning of their first session, subjects were given instructions and practiced a $\Phi_{0.5}$ rhythmic condition. Sessions were organized in blocks of five trials, so that subjects performed the task with a particular target rhythm for five consecutive trials, after which they switched to different target rhythm. The order of blocks (i.e. target rhythms) was restricted so that the targets for consecutive blocks always differed by more than 0.1, in order to reduce the possibility of carryover effects from one block to another. The orders of blocks were randomly assigned to subjects. After performing five blocks, subjects rested for several minutes, and then performed another five blocks in the same order. Thus each subject produced 10 trials per target condition in each session.

On each trial, subjects waited until the third repetition of the high-low two tone metronome pattern, and then began producing the phrase “take on a spa” so that the word *take* coincided with the first (H) metronome tone, and the word *spa* coincided with the second (L) metronome tone. The metronome tones were 50 ms long and separated by 700 ms in all target phase conditions. The high tone was 1200 Hz, the low tone 600 Hz; both were windowed with a Tukey window (r = 0.5). The metronome pattern repeated 12 times from the start of each trial, and the last three pairs of tones were faded out by successive 50% decreases in amplitude. Subjects were instructed to continue repeating the phrase with the same rhythm after the metronome pattern faded out. After a duration of time corresponding to 14 cycles of the metronome pattern, they were signaled to stop. Subjects were told to speak with a normal volume and pitch, and not to tap their feet or hands, or to imagine doing so. They were instructed not to take shallow breaths between each repetition of the phrase, but rather, to take a deep breath when necessary and skip one cycle of the phrase. When taking a breath without the metronome, they were told to estimate the duration of one cycle.

The Carstens Articulograph AG200 (EMA) was used to track articulator movement data at 200 Hz (Hoole 1996). Audio was recorded at 22050 Hz using a microphone clipped onto the shirt of the subject. A secondary audio signal was collected for synchronization with the EMA using a table microphone. So that the metronome tone would not be recorded along with the speech signal, subjects wore earbud headphones. Inspection of kinematic data with and without the earbuds revealed that the presence of the earbuds in the magnetic field generated by the EMA helmet produced no noticeable increase in noise or signal distortion.

Four transducer coils were used, all in the midsagittal plane, in the following locations: 1) on the forehead for reference, 2) on the lowermost projection of the jaw when the subject looked straight ahead, 3) on the outmost projection of the lower lip, and 4) on the blade of the tongue, 2-3 cm from the tongue tip. In electromagnetic articulometry studies, a bite plate is commonly used to discern the angle of the occlusal plane relative to the transmitter coils. Unlike other EMA experiments, subjects were required to wear the (somewhat uncomfortable) transmitter helmet for a rather extended period of time, which necessitated occasional readjustment. The weight of the helmet is balanced by a line above the head of the subject; this presents the dilemma that the more weight is taken off the helmet, the more subject to movement the helmet becomes, especially when subjects do not sit perfectly still. This means that to estimate the occlusal plane throughout the experiment, one would have to recalibrate with a bite plate every time an adjustment is made, which is impractical. Fortunately, the experimental hypotheses are concerned with the relative timing and displacement of articulatory movements, rather than absolute positions of the articulators. Variation in the location of the occlusal plane is not problematic for measurements of relative timing and relative gestural magnitude within trials, and thus a bite plate was not used.

To reduce the influence of preceding and subsequent context on the gestural trajectories of [s] and [p], a number of measures were taken in the design of the carrier phrase, “take on a spa”. A guiding principle behind this design was that the tongue and lip gestures entering and exiting [sp] should be of large magnitude, in order to make landmarks associated with their articulatory trajectories easily detectable. First, the [ə] preceding [sp] encouraged subjects to pass through a relatively neutral vocal tract configuration immediately prior to the articulations of interest. It was deemed important to employ a four-syllable phrase, because a three-syllable phrase is produced abnormally slowly in a 700 ms interval, which was the constant intertone duration employed in Cummins & Port (1998). A phrase with the fewest potential coarticulatory confounds would require no tongue raising or fronting, and no large magnitude lip closure, protrusion, or retraction gestures in the two stressless syllables preceding [sp]. However, this ideal design is not achievable with the set of English function words. The function word *on* was chosen because pilot investigations revealed that both the vowel and alveolar nasal in *on* [an] tended to be coarticulated with the preceding velar stop in *take* [teik]. Moreover, rather than a distinct tongue gesture for [n] occurring, a nasalized back vowel [ã] was normally produced. Due to the deletion of the nasal and resultant vowel nasalization, the phonetic value of the sequence was generally [te^{i}kã:əspa].

In some cases, the horizontal velocity maximum of tongue movement from [ã:ə] to [s] is a more distinct marker of the [s] gesture than the vertical (raising) velocity maximum. The sensor coil on the tongue blade undoubtedly contributed to the observed articulatory patterns. It is possible that subjects may have made abnormal alveolar closures to adjust for the presence of the sensor. It is unlikely, however, that this had a profound effect upon temporal relations between [s] and [p], or that this adjustment factor differed between rhythmic conditions. The vowel following [sp] was chosen to be a low, central/back vowel [a] in order to maximize the speed of the releases associated with [s] and [p]. No coda consonant followed the vowel in order to further maximize the movement amplitude and avoid potential anticipatory coarticulation.

2.2 Analysis

Data from one session were not analyzed because the subject suffered from fatigue and exhibited a lack of attention to the task. Another two sessions from different subjects were lost due to equipment malfunction; hence for three subjects data from only one session are analyzed. One of the female subjects performed three sessions, and the four remaining subjects each performed two sessions. The first trial that each subject performed in each target rhythm condition in each session was disincluded from the analysis. Thus, for each subject who performed 2 sessions, 18 trials in each target rhythmic condition were analyzed. Several trials in which subjects were too inaccurate were also discarded, as were several trials in which sensor coils stopped functioning properly.

The kinematic data from each trial were smoothed with a moving-average filter to reduce noise. For each subject, data were rotated in the X-Y plane to attain a common angle from mean tongue position to mean lip position. For point-based analyses, all horizontal and vertical minima (x, y), maxima (X, Y), velocity minima (dx, dy) and velocity maxima (dX, dY) were identified. Because speech targets are generally accomplished through synergies of articulators (e.g. [s] involves tongue and jaw movements), the tongue (Tj) and lower lip (Lj) trajectories analyzed here will in most cases include components associated with jaw movement. Where otherwise noted, tongue (T) and lip (L) positions from which jaw (J) position has been subtracted are analyzed, offering a view of tongue and lip movement independent from jaw movement.

P-centers of the syllables *take*, *on*, and *spa* served as the basis for rhythmic analyses. The p-center is an approximate location of the "beat," or perceptual-center of a syllable (cf. Allen 1972, 1975; Morton, Marcus, & Frankish 1976; Howell 1988; Pompino-Marschall 1989; DeJong 1994; Scott 1993, 1998 for more information on the concept of the p-center). Following Cummins & Port (1998), the acoustic signal was filtered with a passband of [700-1300] Hz, using a 1st-order Butterworth filter, which has gradual roll-offs. This effect of this is to greatly diminish spectral energy corresponding to F0 and higher-frequency fricative energy in the signal. The magnitude (absolute value) of the bandpass-filtered signal was then low-pass filtered using a 4th-order Butterworth filter with a 10 Hz cutoff. The result is a smoothly-varying representation of mostly vocalic energy in the signal. P-centers were estimated to be the midpoints in time between the points when the signal amplitude is 10% above its local minimum and 10% below its local maximum. These estimates consistently yield points that are near vocalic energy velocity maxima.

Also automatically detected in each phrase was the *s-center*, i.e. the sibilance center, which is the point in time corresponding to maximum sibilance energy associated with the [s] in *spa*. In this case, a passband of [5000-7500] Hz was used. To identify locations where subjects took a breath, the acoustic data from every trial were visually and auditorily inspected; in the course of this process, spurious and missing p-centers and s-centers were corrected. Where relevant in subsequent analyses, the data from each trial were separated into inter-breath groups, so that analyses are not confounded by the pauses associated with breaths.

3. Results

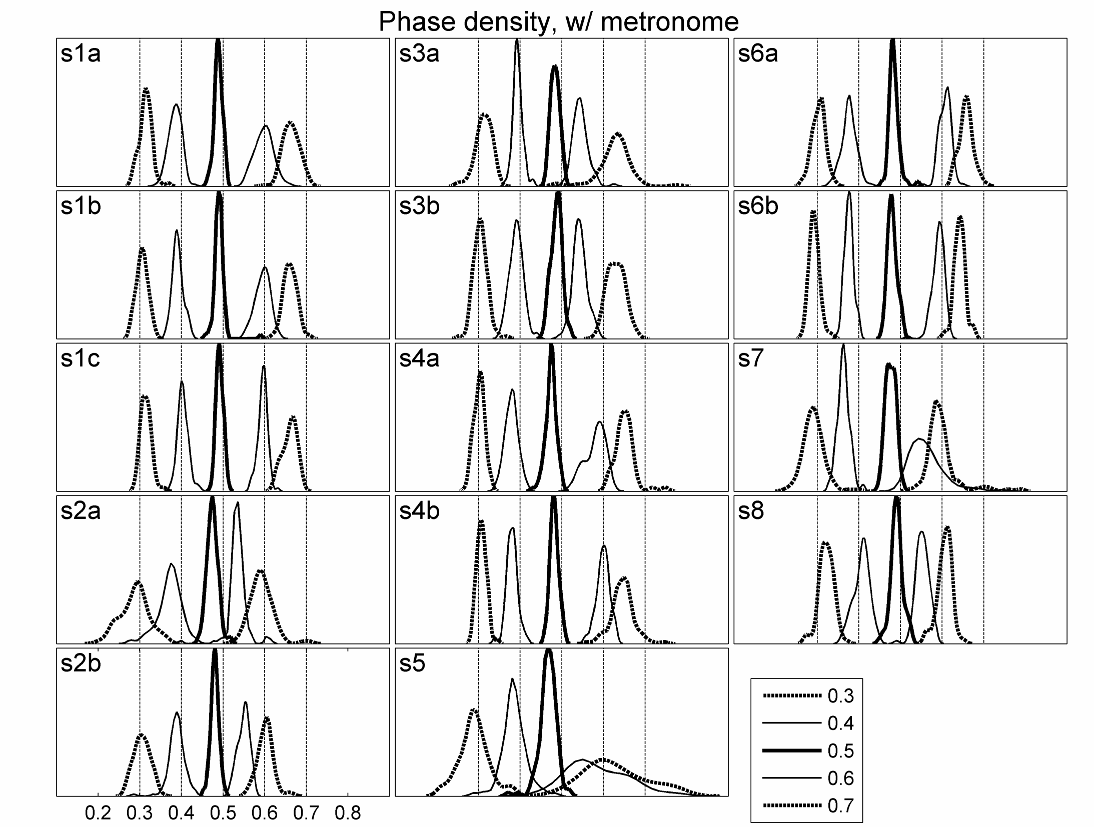
In the following, the symbols $\Phi_{0.3}$, $\Phi_{0.4}$, etc. are used to refer to the target relative phase conditions, and φ is used to refer to produced relative phases. Subjects and sessions are referred to with numbers and letters, e.g. *s1c* represents the third session of the first subject. PC1, PC2, and PC3 refer to the p-centers of *take*, *on*, and *spa*; SC refers to the s-center of *spa*. M1 and M2 refer to the first and second metronome tones. Unless otherwise stated, the following analyses consider jaw-tongue (Tj) and jaw-lip (Lj) synergies rather than independent lip (L) and tongue (T) movements, which are obtained by subtraction of jaw movement (J).

3.1 Metrical variability

Previous observations of increased rhythmic variability in the higher-order target rhythm conditions of the speech cycling task were replicated (Cummins & Port 1998). A partial harmonic timing effect was observed, but the interpretation of its cause is not unambiguous. Unlike previous work, a linear bias in produced phases was seen, and certain kinematic measures were more closely timed than p-centers to metronome tones.

Fig. 4 (top) shows for each session kernel density estimates of the distributions of the phase of PC3 (the p-center of *spa*) relative to the metronome period, i.e. [PC3(n) - M1(n)]/[M1(n) – M1(n+1)], where n is the phrase number within a trial. First examine the $\Phi_{0.3}$, $\Phi_{0.4}$, and $\Phi_{0.5}$ target phase conditions with the metronome present. For some subject-sessions, φ in $\Phi_{0.3}$ tended slightly toward the 1/3 harmonic, but for other subject-sessions, φ deviated away from this harmonic phase (an "anti-harmonic" bias). Most subjects (except for the anomalous *s8*) exhibited a small but significant deviation toward the nearest harmonic in the $\Phi_{0.4}$ condition. These biases would constitute firm evidence for harmonic timing, if not for biases in the same direction—i.e. for lower φ—in the most harmonic phase condition, $\Phi_{0.5}$. The likely cause of this early phase bias will be discussed below.

Considering $\Phi_{0.6}$ and $\Phi_{0.7}$, more variation between subjects occurred. Some subjects (*s1*, *s4*,*s6*) produced φ tending toward the 2/3 harmonic, particularly in $\Phi_{0.7}$. This could be due either to an early phase bias or to a harmonic timing effect—yet both the early-phase bias and a harmonic timing effect appear to be absent from the $\Phi_{0.6}$ condition for these subjects. An alternative explanation is that $\Phi_{0.6}$ and $\Phi_{0.7}$ are more difficult to attain (less stable) and so φ in these conditions were attracted to 1/2. This appears to have been the case for the other subjects (*s2*, *s3*, *s5*, *s7*, *s8*) who generally did not achieve the target phase in the $\Phi_{0.6}$ and $\Phi_{0.7}$ conditions. This is particularly obvious when the metronome was not present, as shown in Fig. 4 (bottom), where these same subjects produced φ around 1/2 in $\Phi_{0.6}$ and $\Phi_{0.7}$. Compared to the Cummins & Port (1998) experiment, some subjects were much less adept at maintaining the target rhythms without the metronome.


Phase density, w/ metronome
s1a
s3a
s6a
s1b
s3b
s6b
s1c
s4a
s7
s2a
s4b
s8
s2b
s5
0.2 0.3 0.4 0.5 0.6 0.7 0.8
0.3
0.4
0.5
0.6
0.7

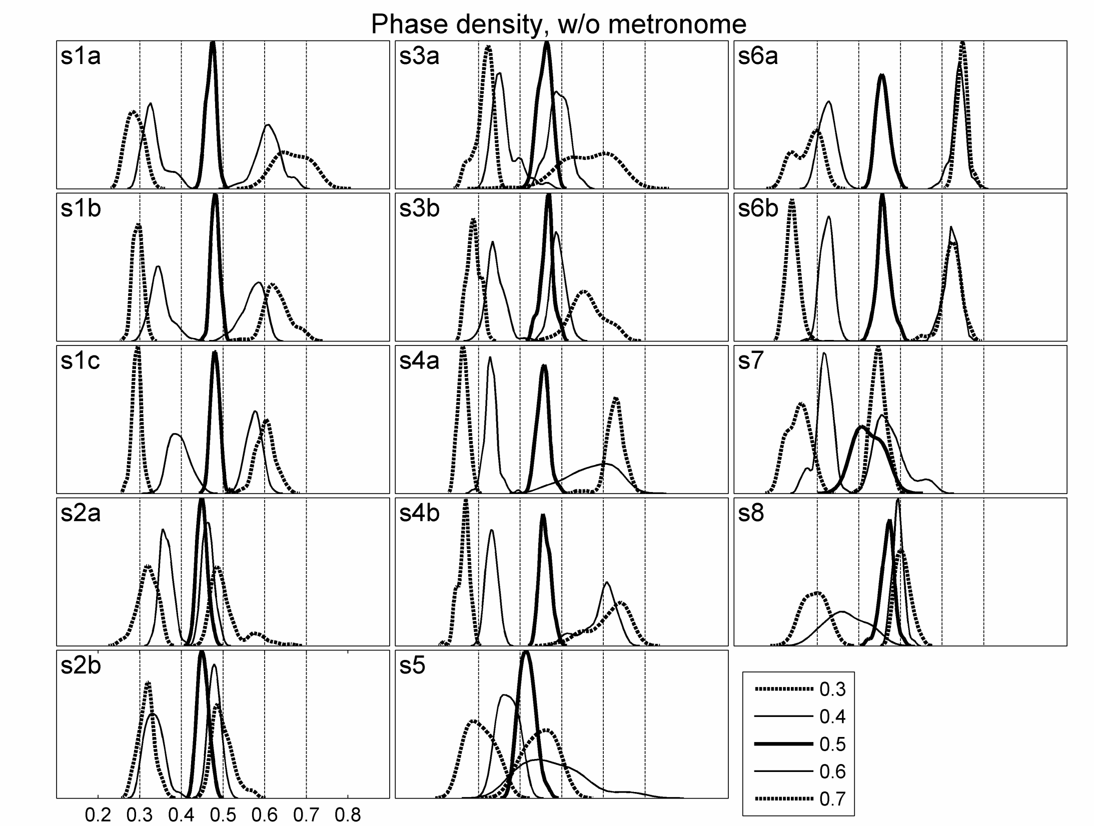

Phase density, w/o metronome
s1a
s3a
s6a
s1b
s3b
s6b
s1c
s4a
s7
s2a
s4b
s8
s2b
s5
0.2 0.3 0.4 0.5 0.6 0.7 0.8
0.3
0.4
0.5
0.6
0.7

Fig. 4 Relative phase densities in each subject-session for the p-center of *spa*, with the metronome (top) and without the metronome (bottom).

The several anti-harmonic φ for $\Phi_{0.3}$ and the early phase bias for $\Phi_{0.5}$, suggest that some factor influencing rhythmic coordination in the experiment may have effected a general bias toward earlier φ. Notably, the measures in Fig. 4 (top) differ from the measure used in Cummins & Port (1998), in that the metronome period is used as the reference frame, rather than the first p-center, PC1. For the present data, using PC1 as a reference period for defining observed relative phase exacerbates the early phase bias. This can be seen in Fig 4 (bottom), which presents φ distributions from phrases without the metronome present (and hence uses PC1 as the reference period).

The probable source of the early phase bias can be seen in Fig. 5, which shows a phrase produced in $\Phi_{0.5}$ by *s1c*. First, observe in this phrase that both PC1 and PC3 occurred after M1 and M2 respectively. For almost all subject-sessions, these p-centers tended to occur after the metronome tones. This indicates that generally subjects did not time their p-centers to coincide with metronome beats. Instead, for many subjects other gestural landmarks appear to be more closely timed to the metronome tones. In particular, for most subjects the measures J dy (i.e. jaw vertical velocity minima) and Lj dy or Tj dy exhibited relatively smaller mean deviations from M1 and M2, and furthermore, those deviations were comparably variable to those of the p-centers. Table 1 shows mean interval durations between these measures and M2 for each subject-session in $\Phi_{0.5}$. Except for *s2* and *s6a*, the other subjects exhibited closer timing between M2 and one or more of the kinematic measures than they did between M2 and PC3.

Table 1. Mean interval durations between the 2nd metronome tone and p-center / articulator velocity minima of *spa*.

	PC3 mean	(s.d.)	Tj_dy		Lj_dy		J dy	
s1a	.045	(.022)	-.005	(.024)	-.018	(.021)	-.012	(.022)
s1b	.012	(.025)	.015	(.025)	.031	(.027)	.012	(.026)
s1c	.038	(.019)	-.026	(.020)	-.036	(.021)	-.030	(.019)
s2a	-.010	(.036)	.044	(.040)	.016	(.039)	.022	(.038)
s2b	.002	(.041)	.034	(.042)	.002	(.043)	.007	(.041)
s3a	.046	(.029)	-.008	(.028)	-.046	(.028)	.096	(.055)
s3b	.062	(.036)	-.062	(.036)	-.070	(.035)	.202	(.035)
s4a	-.026	(.030)	.041	(.030)	.012	(.032)	.116	(.044)
s4b	.046	(.024)	-.017	(.023)	-.041	(.022)	-.023	(.117)
s5	.048	(.029)	-.015	(.030)	-.017	(.028)	-.018	(.029)
s6a	.001	(.022)	.030	(.024)	.016	(.022)	.021	(.022)
s6b	.015	(.023)	.002	(.022)	-.016	(.024)	-.007	(.025)
s7	.028	(.027)	-.023	(.035)	-.029	(.033)	-.026	(.061)
s8	.023	(.021)	.041	(.021)	-.017	(.019)	.122	(.020)

These data suggest that the p-center was not the moment of the stressed syllable that the majority of subjects most accurately aligned with metronome tones—instead they timed maximum velocity of jaw opening, or [s] or [p] release, to coincide with the tone.

Furthermore, jaw opening tended to occur earlier relative to the vowel in [spa] compared to [teik], as evident in Fig. 5. A potential cause of this may have been that the vowel [ei] required a smaller magnitude release gesture from [t] than did the vowel [a] from [sp], resulting in a slower release in the former. Because of this asymmetry, using PC1 as a reference frame introduces even more early phase bias, because the discrepancy in timing between jaw opening and vowel onset between *take* and *spa* makes the numerator of the relative phase measure smaller.

With and without the metronome, φ variability was generally lowest in $\Phi_{0.5}$. This accords with previous findings that the lowest-order harmonic mode is the most stable, and is crucial for subsequent analyses in this study. Regarding the other Φ, there was a fair amount of between-subject variation in patterns of variability. With the metronome present, some subjects exhibited more variability in $\Phi_{0.6}$ and $\Phi_{0.7}$ than in $\Phi_{0.3}$ and $\Phi_{0.4}$, although others showed comparable amounts. Without the metronome, most subjects produced more variable φ in $\Phi_{0.6}$ and $\Phi_{0.7}$.

No general pattern for greater variability in the less-harmonic $\Phi_{0.4}$ and $\Phi_{0.6}$ compared to $\Phi_{0.3}$ and $\Phi_{0.7}$ was observed. Why did these higher-order Φ patterns differ from those reported by Cummins & Port? One possibility is that subjects in the present experiment were not trained sufficiently; perhaps with more practice and/or coaching the subjects would have exhibited the expected patterns. Due to limitations of the experimental timeframe this was not possible. However, the subject who participated in three sessions (*s1c*) did not show this expected pattern either, and instead exhibited less variability in $\Phi_{0.4}$ and $\Phi_{0.6}$ compared to $\Phi_{0.3}$ and $\Phi_{0.7}$. A different explanation for this discrepancy is that the 1/3 and 2/3 phase attractors were for some reason weaker in this experiment, perhaps because the subjects were less rhythmically adept. The greater difficulty of the $\Phi_{0.6}$ and $\Phi_{0.7}$ rhythms presumably arises from the awkwardness of having to say *take* immediately after *spa*. This controverts the normal prosody of the phrase, which would have a break after *spa*—although this same factor would have been present in the phrases used by Cummins & Port.

Overall the most important observation for present purposes is that the $\Phi_{0.5}$ condition incurs the lowest rhythmic variability. Hence in subsequent analyses of gestural timing, we will focus on comparisons of gestural coordination between the more rhythmically variable Φ with the most stable condition, $\Phi_{0.5}$.

3.2 Intergestural phasing variability

The gestural trajectories on each trial are most basically conceptualized as continuous objects (sampled discretely), but most analyses of gestural timing have extracted relevant landmarks (points in time) from these trajectories and analyzed the distributions of temporal intervals between the landmarks. A sensible choice of landmarks includes horizontal and vertical position extrema—points in time when an articulator reaches a maximum or minimum in the horizontal or vertical dimension. Additional useful landmarks are velocity maxima and minima, i.e. when an articulator moves most quickly between position extrema. Here we will only consider landmarks most relevant to the timespan immediately preceding and following M2 and PC3, i.e. those associated with articulation of [spa]. Furthermore, because of the substantial between-subject variation in rhythmic behavior without the metronome, only those phrases produced with the metronome will be analyzed.

Fig. 5 shows a representative example of a phrase from *s1c* in $\Phi_{0.5}$. The top panel shows the magnitude of the bandpass-filtered waveform and its amplitude envelope obtained with lowpass filtering, which is used to estimate p-centers. Also shown is the signal used to estimate the s-center of [spa]. Metronome beats and p-centers are labeled as well. The middle panels show 2D plots of jaw, lip, and tongue positions. Also labeled along these trajectories are the locations in time when metronome beats, p-centers, and s-centers occurred. The bottom panels show the horizontal and vertical components of motion. Coordinate axes are in cm^{-3} and relative to the center of the mid-saggital plane within the EMA helmet. Most movement ranges are on the order of 1 cm. P-center and metronome locations are marked with vertical lines. Position maxima are labeled with open circles, minima with closed circles, velocity maxima with open squares, minima with stars.

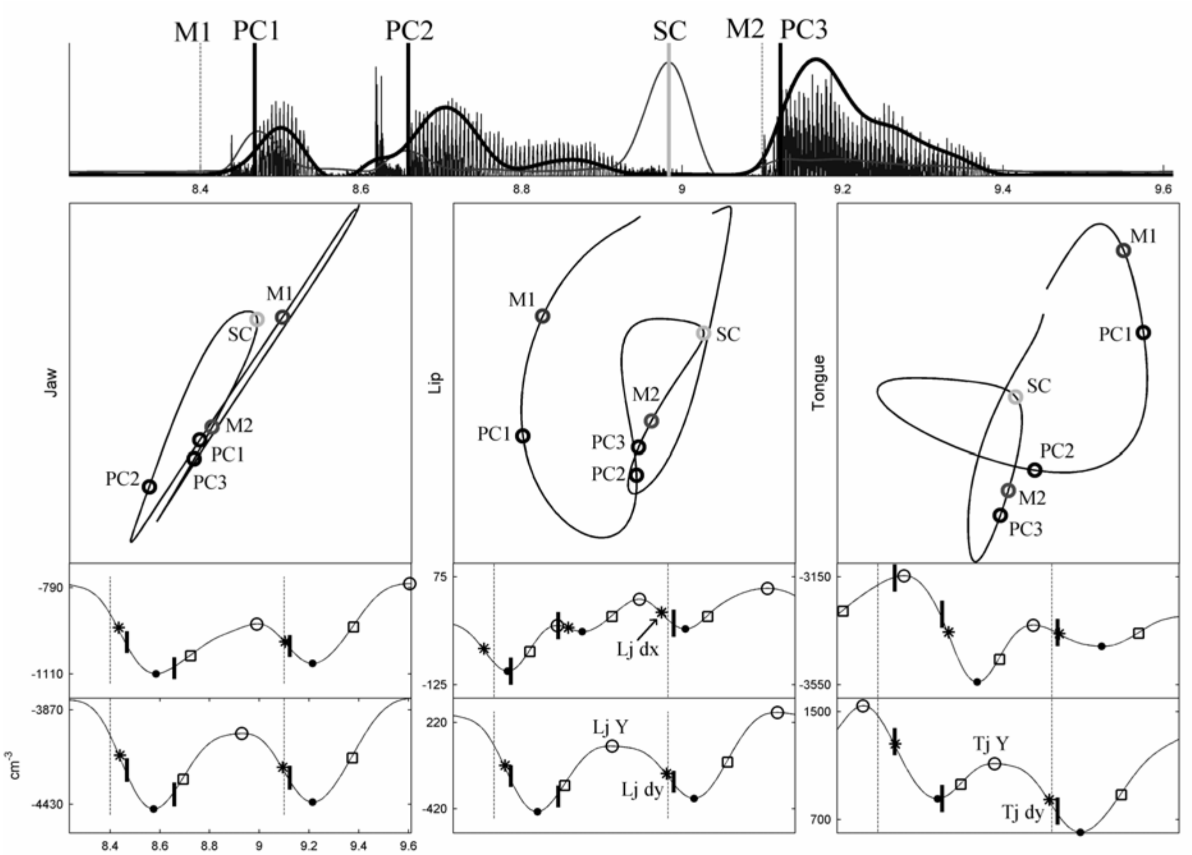


Fig. 5. Example of a phrase in $\Phi_{0.5}$ from *s1c*. Shown are: amplitude envelope used for p-center detection (top), 2D articulator positions (middle), and 1D articulator movement (bottom). See text for details.

Certain gestural landmarks in the above figure of are particular interest here. One is the maximum lip height (Lj Y) preceding PC3, which is associated with the labial closure of [p] in *spa*. Maximum tongue height (Tj Y), associated with the raising of the tongue for [s], is also of interest. However, this was not the most reliable landmark of the tongue gesture for some subjects. It was common for tongue raising to occur much earlier than the onset of the [s] gesture, due to raising associated with the presence of the preceding schwa and nasal. In these cases, the constriction narrowing expected for [s] was accomplished via horizontal motion of the tongue. Fig. 5 illustrates this in the 2D trajectory: the tongue reaches a height plateau relatively early (prior to SC), throughout which a rapid horizontal fronting occurs. Other relevant landmarks are tongue and lip velocity minima (Tj dy, Lj dx, Lj dy), which correspond to the release portions of the [s] and [p] gestures.

As hypothesized, patterns of variance across rhythmic conditions provided evidence for interaction between rhythmic and gestural timing. The more difficult rhythmic conditions were generally associated with increased temporal variability in interval durations between tongue and lower lip gestural landmarks. The panels in Fig. 6 show for each subject (averaged across sessions) the ratio of the variance of a specific interval in each Φ condition to the variance in the $\Phi_{0.5}$ condition. Fig. 6 (a) shows this ratio for the interval between vertical tongue blade maxima (Tj Y) and lower lip maxima (Lj Y). The ratio of variances corresponds to an F-statistic, and horizontal bars show a 95% confidence region for equality of variance (based on a sample size of 150, which is a conservative approximation of the number of phrases each subject produced).

The patterns in Fig 6 (a-d) confirm the hypothesized greater variability in the more variable rhythmic conditions for combinations of Tj Y, Tj dy, Lj Y, and Lj dy, which represent tongue and lip closure and release gestures. With only a few exceptions, most subjects exhibited significantly greater intergestural variability in $\Phi_{0.3}$ and $\Phi_{0.4}$. The pattern is less general for the higher target phases. For $\Phi_{0.6}$ the majority of the subjects tended to produce more variable intervals, but a few subjects produced variances that either did not differ significantly from variances in $\Phi_{0.5}$ or were actually lower. For $\Phi_{0.7}$, variance does not appear to pattern consistently one way or the other relative to $\Phi_{0.5}$.

There is, however, some consistency in which subjects produced the unexpected patterns of lesser variance in the higher Φ. Subjects *s2*, *s7*, and *s8* were generally responsible for these occurrences; in contrast, subjects *s1*, *s3*, and *s5* produced more of the expected patterns across the four measures. These differences between subjects suggest that some other mechanism affected intergestural timing in the high Φ. In the following section, we will consider how different strategies for altering synergistic relations between effectors may have been responsible for the variation.

Fig. 6 (e-f) exchanges landmarks of the tongue gesture trajectory for the s-center (sibilance-center), which corresponds to the point of loudest high-energy noise associated with [s]. Although the s-center is not derived directly from a kinematic measurement, it approximates fairly well the point in time when the [s] constriction is most narrow. Indeed, a reasonable argument can be made that the s-center is an acoustic target of the [s] gesture, and thus a suitable candidate for a coordinated event. Moreover, for many subjects, intervals between s-centers and the metronome tone are no less variable than the other measures considered here. Interval variance patterns between s-centers and [p] gesture landmarks in Fig. 6 (e-f) reflect the predicted variance patterns even more closely. In (f) we see that only two subjects violate the expected patterns in the high Φ conditions.

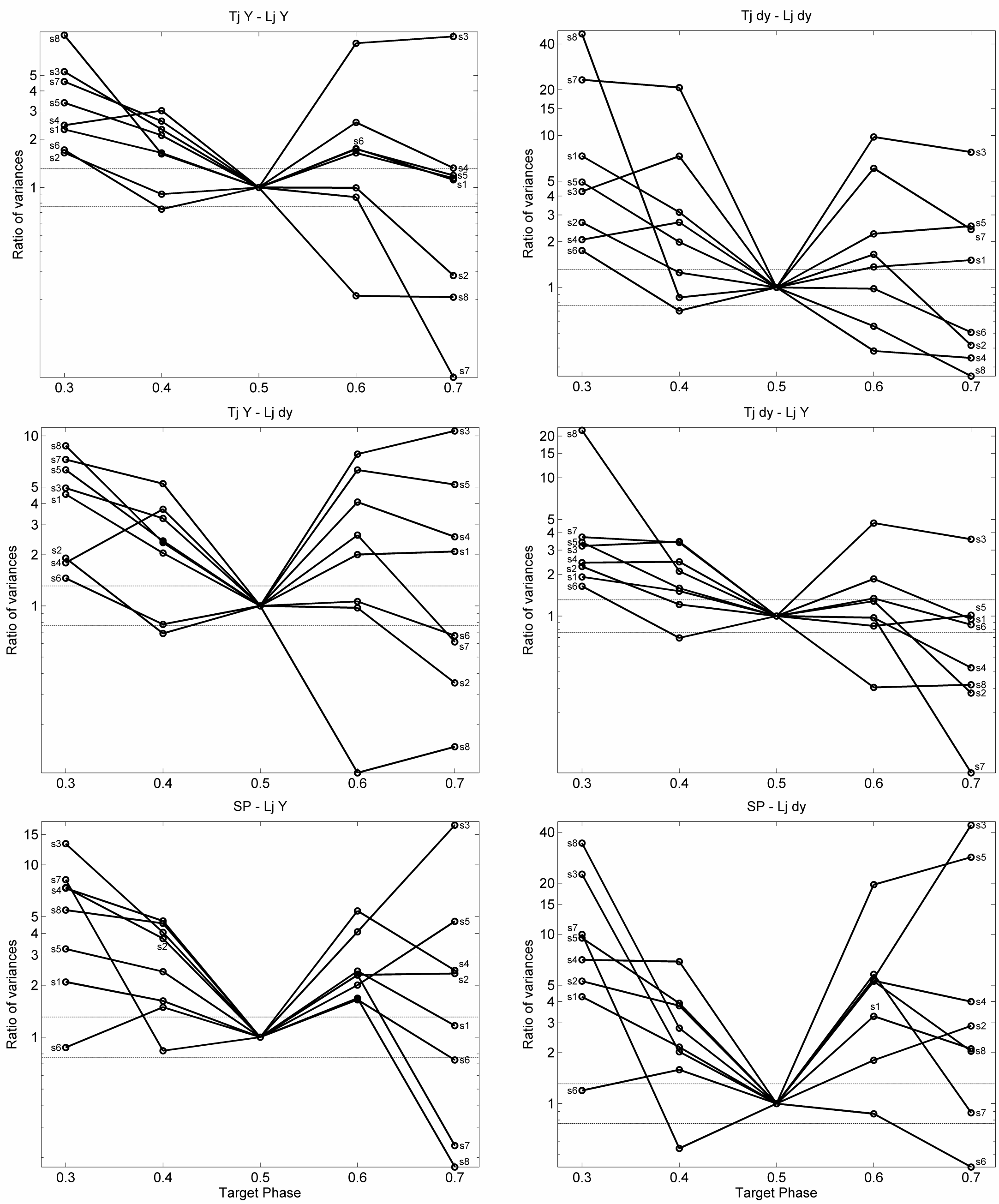


Fig. 6. Ratios of variance in intergestural timing across rhythmic conditions. Variances were averaged across sessions within subjects. Ratio denominator is variance in $\Phi_{0.5}$. The ratios are equivalent to F-statistics; approximate 95% confidence intervals are shown.

These data confirm the hypothesis of covariability between rhythmic and gestural systems. When rhythmic timing was more variable, intergestural timing between [s] and [p] was also more variable, at least for the majority of subjects. Synergistic patterns in the following section suggest that some of the exceptions, most of which occurred in high Φ, may be related to differences in how subjects altered the relative use of articulators in $\Phi_{0.6}$ and $\Phi_{0.7}$.

3.3 Synergistic patterns

As hypothesized, the relative contributions of effectors to achieving gestural targets also varied across rhythmic conditions. The variation was systematic for effectors in the [s] synergy, but no clear pattern was observed for the [p] synergy. Fig. 7 shows the average ratio of ranges of vertical jaw and tongue movement in the neighborhood of [spa]. Here tongue movement has been estimated by subtracting jaw position from tongue position, and ranges were taken as the differences between local vertical maximum and minimum. Almost all subjects exhibited a tendency for reduced jaw movement (relative to tongue movement) in the high Φ compared to the low Φ. Note that all subjects also exhibited less absolute movement magnitude in high Φ (not shown).

Why do most subjects employ relatively less jaw movement in $\Phi_{0.6}$ and $\Phi_{0.7}$? One reason may involve anticipation of the upcoming alveolar stop gesture in the following phrase. In $\Phi_{0.7}$ there are only 300 ms after M2 before the following M1, and so subjects have to make the closure gesture for the [t] of *take* very quickly after the [s] and [p] release gestures. Decreasing jaw movement amplitude makes this transition from opening to closure easier to accomplish in a relatively short period of time.

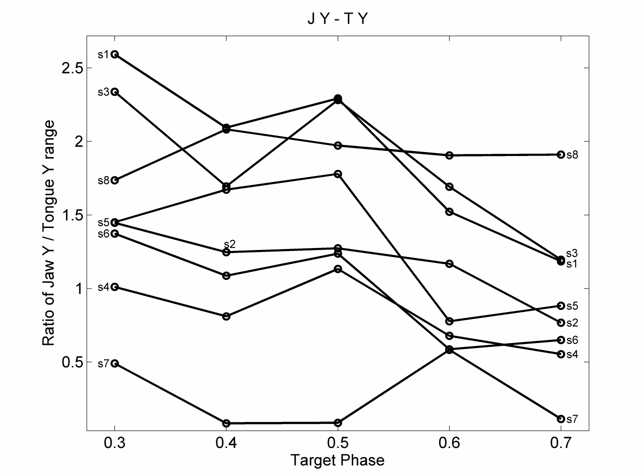


Fig 7. Average ratio of range of vertical jaw movement to vertical tongue movement in the neighborhood of *spa*, across target phases. Ratios were averaged across sessions within subject.

There are two exceptions to the trend of reduced contribution of jaw movement in high Φ. *s8* maintained a relatively high ratio of jaw-to-tongue movement across Φ conditions, and *s7* maintained a relatively low ratio of jaw-to-tongue movement in all Φ. It may be no coincidence that these two subjects were also responsible for several of the unpredicted variance patterns seen in the analysis of intergestural variability. The relatively greater use of the jaw effector in the jaw-tongue synergy involved in [s] may make intergestural timing between the tongue and lip less variable, because both [s] and [p] gestures involve the jaw. If these gestures become more predominantly jaw-based, then lower intergestural variability is expected. This raises the question of whether the increased variability seen in $\Phi_{0.6}$ and $\Phi_{0.7}$ is associated more with a reduction in jaw use relative to tongue use, rather than rhythmic difficulty. To this point, synergistic changes cannot be the sole explanation for increased variability, because intergestural variability increases in $\Phi_{0.3}$ and $\Phi_{0.4}$ as well. A more complicated but also more accurate account of the observed patterns might involve consideration of both rhythmic difficulty and intragestural synergistic relations.

4. General discussion

The three main findings of the above experiment are reiterated below:

(1) Higher rhythmic variability in phrase-foot timing is associated with more difficult (or less harmonic) target rhythms. This replicates findings of Cummins & Port (1998).

(2) Higher variance in intervals between gestural landmarks co-occurs with higher rhythmic variability for more difficult target rhythms.

(3) Synergistic relations, in the form of relative magnitudes of effector movements, change across different rhythmic contexts.

The following discussion will be primarily concerned with understanding rhythmic-gestural covariability in the context of a model of coupled oscillators. To that end, a simple instantiation of a dynamical model will be presented. Simulations of the model will be used to demonstrate how rhythmic-gestural coupling can lead to increased intergestural variability.

4.1 A dynamical model of rhythmic-gestural interaction

There are innumerable possibilities for exactly what form the equations describing a model of coupled oscillatory systems can take, and empirical data will inform the future development of similar models. The model presented here is a proof-of-concept, rather than a claim about the most realistic formulation. To this end, important questions will be put aside regarding exactly how the coupling relations between oscillators should be structured, and whether the coupling can be simple phase-coupling or should also involve amplitude. The most important thing for current purposes is that the model account for interactions between the disparate timescales of rhythmic systems and gestural systems. For more advanced instantiations of a model capable of such behavior, confer Keith & Rand (1984), Saltzman & Byrd (2000), and especially Haken et. al. (1996).

Each oscillator in a multiscale model can be described in a simple fashion by differential equations governing its phase and amplitude, shown in Eqs. (1):

(1) $$\dot{\theta}_i = \omega_i + \sum_j A_{ij} F(\phi_{ij}) + \eta_i$$
$$\dot{r}_i = 1$$

(2) $$\phi_{ij} = \omega_i \theta_j - \omega_j \theta_i \quad mod\ 2\pi$$

(3) $$F(\phi_{ij}) = -\frac{dV_{ij}}{d\phi_{ij}}$$
$$V(\phi_{ij}) = -\cos(\phi_{ij} - \Phi_{ij})$$

The equations in (1) treat the phase velocity of each oscillator as the inherent frequency of the oscillator plus the sum of phase changes due to coupling forces exerted on the oscillator (*F*). The coupling forces are weighted by corresponding coupling strengths (*A*). A noise term (η) is present as well, adding Gaussian noise to the phase velocity. The polar amplitudes (radii) of the oscillators are constant, and so in the absence of the noise and coupling terms, the equations in (1) correspond to a polar form of a harmonic oscillator. More biologically plausible models usually incorporate some form of nonlinear damping, such as Van der Pol and/or Rayleigh damping. For current purposes, these nonlinear terms are unnecessary, because the entirety of the conceptual workload can be carried by the notions of phase coupling and synchronization.

The coupling function *F* takes as its argument a generalized relative phase difference (φ) between a pair of oscillators. The generalized relative phase difference is defined in Eq. (2), and is modulo 2π like the phase variable. Oscillator frequencies (ω) are restricted to integer multiples of the lowest frequency, which belongs to the phrase oscillator. The phrase oscillator can be defined to have an inherent frequency of 2π with the introduction of an appropriately normalized time variable. Following Haken (1983), the coupling force between a pair of oscillators is the negative of the derivative of the potential function (*V*) governing their relative phase, shown in Eqs. (3). The potential is greater when the relative phase between two oscillators is farther from the target relative phase (Φ_{ij}), which is a stable equilibrium. The absolute value of its derivative, which corresponds to the speed with which relative phase changes, is greatest when the relative phase is at $\Phi_{ij} \pm \pi$, which halfway between the stable and unstable equilibria.

The coupling forces can either speed up or slow down the instantaneous frequency (phase velocity) of an oscillator. For any pair of oscillators, osc_1 and osc_2, the strengths of the coupling forces exerted upon osc_1 by osc_2 and vice versa are represented separately in the coupling strength parameter matrix (A_{ij}). This allows for either oscillator to drive the other one, or for mutual interaction, which can be symmetric or asymmetric. General hypothetical parameter matrices relevant for rhythmic-gestural interaction in a production of the phrase *take on a spa* are shown in Table 2.

Table 2. Coupling strength and target relative phase parameter matrices relevant to the syllable *spa* in the phrase *take on a spa*.

		Coupling strengths* ***A***						**Generalized relative phase targets** Φ					
		Phr	Ft	σ	$C_{[s]}$	$C_{[p]}$	$V_{[a]}$	Phr	Ft	σ	$C_{[s]}$	$C_{[p]}$	$V_{[a]}$
$\omega_{Phr} = 2\pi$	Phr		*a*						0				
$2\omega_{Phr}$, $3\omega_{Phr}$	Ft	*a*		*b*	x_1	x_2	x_3	0		0	0	0	0
$2\omega_{Ft}$, $3\omega_{Ft}$	$\sigma_{/spa/}$		*b*		y_1	y_2	y_3		0		0	0	0
ω_σ	$C_{[s]}$					*c*	*d*					0.5	0
ω_σ	$C_{[p]}$				*c*		*d*				0.5		0
ω_σ	$V_{[a]}$				*d*	*d*					0	0	

*Coupling strengths represent the influence of row *i* upon column *j*.

In the coupling structure in Table 2, the syllable [spa] belongs to its own foot, and interactions between this syllable and others are not considered. Moras have been omitted, as well as many other gestures involved. C[s] represents only the tongue blade gesture associated with [s], C[p] only the lower-lip gesture, and V[a] represents a tongue gesture associated with [a]. A target relative phase of 0 indicates in-phase synchronization, 0.5 anti-phase synchronization. Although there are two Ft systems involved in production of the phrase, they are synchronized in-phase and collapsed into a single system in the present treatment. The only anti-phase specification in the model adheres between the [s] and [p] gestures, and is responsible for a c-center effect relative to the vowel (cf. section 1.2).

In the coupling matrix, if a term appears on both sides of the diagonal, either asymmetric or symmetric coupling occurs between the corresponding pair of systems. In the above parameterization, gestural systems mutually influence each other symmetrically, and rhythmic systems do as well. If a term only appears on one side of the main diagonal, this indicates a unidirectional coupling (often called *driving*), which means that the system in the corresponding row influences the system in the corresponding column, but not vice versa. In the coupling structure above, if the *x* and *y* parameters are greater than zero, the foot and syllable oscillators drive the gestural systems.

The hypothesized interaction between rhythmic and gestural systems in the present experiment can be modeled with variation of coupling strength between rhythmic and gestural systems (*x*, *y*). Manipulation of coupling strength parameters corresponds in an intuitive way to changes in rhythmic difficulty and rhythmic-gestural coupling. More difficult rhythms are associated with greater variance in rhythmic timing, and the model can replicate this with a decrease in inter-rhythmic coupling strength. The reason that coupling interacts with variability in the first place has to do with the effect of noise on the phase velocities of oscillators. Due to its Gaussian nature, the noise normally has only a small effect on oscillator phases (which translates to a small effect on relative phase), but on occasion it has a larger effect. With relatively strong coupling forces between systems, large fluctuations in oscillator phases and relative phases are corrected quickly, making relative phases less variable. However, when coupling forces are weaker, noise-induced fluctuations in relative phase are corrected more slowly, and thus more variance is observed in the system. This effect can operate on gestural systems through driving forces from rhythmic systems: stronger coupling from rhythmic to gestural systems reduces variability in intergestural timing.

4.2 Simulation

The idea that a multiscale model can replicate the experimentally observed relations between rhythmic and intergestural variability was verified through numerical simulation. Fig. 8 presents an example simulation of the model. The top panel shows oscillator positions (i.e. $\cos \theta_i$) and selected relative phases over time. For both rhythmic and gestural systems, the oscillators begin out of phase but eventually synchronize. This synchronization can be observed in the stabilization of generalized relative phases (φ). Initial phases in this example were θ = [1.02, 5.75, 0.30, 5.40, 0.09, 2.20] radians, corresponding in degrees to [58°, 329°, 17°, 309°, 5°, 126°]. Regardless of the initial phases of the oscillators, the same relative configuration is eventually reached; initial phases only have an effect on how long relative phases take to stabilize. Following Saltzman & Nam (2003), Φ_{C1C2} was set lower than $-\pi$ (ideal anti-phasing), in this case $\Phi_{C1C2} = -\pi/2$. Further details regarding simulation parameters can be found in Appendix A.

For the Phr, Ft, and σ systems, the pattern is such that the Ft repeats twice within the Phr period, the σ repeats four times, and both of these oscillators hit a peak (i.e. $\theta = 0 = 2\pi$) whenever the Phr oscillator does. For the C_1, C_2, and V systems, C_1 phase precedes V by the same amount that C_2 follows V—this exemplifies a c-center effect.

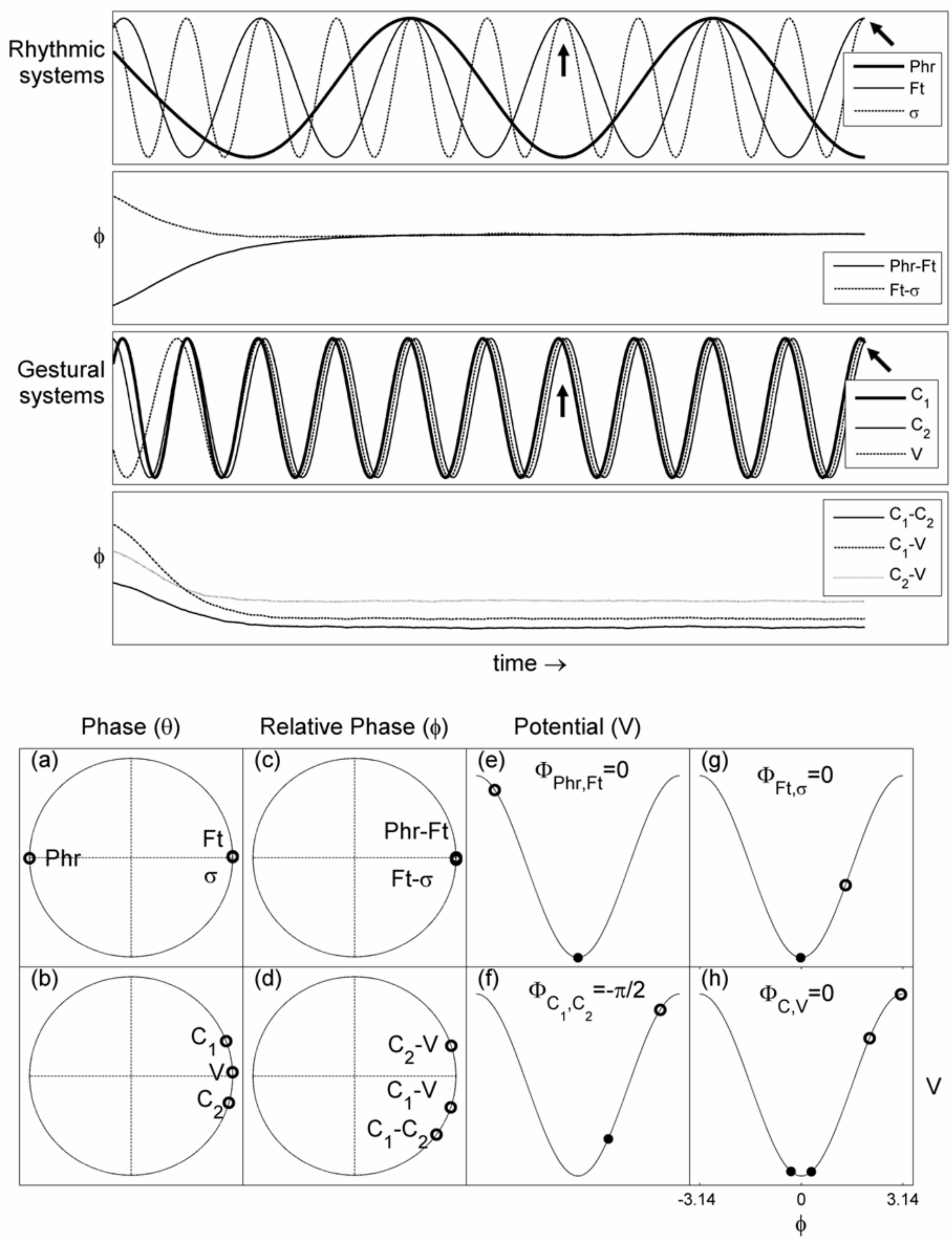


Fig. 8. Example simulation. Oscillator motions and relative phases are shown over time (top); arrows indicate points where relative phases are measured. Simulation-final phases (bottom, a-b) and relative phases (c-d) are shown on the unit circle. Initial and final relative phases are also shown in their corresponding potentials (e-h).

The simulation in this example ends when Phr reaches a minimum ($\theta = \pi$), and thus the Ft, σ, and V oscillators are at a peak. Fig. 8 (bottom) shows the final oscillator phases on the unit circle (a-b), where phase angle increases counterclockwise. Relevant simulation-final generalized relative phases are shown in (c-d). The initial and final relative phases are also shown in potentials. (e) and (g) show that rhythmic relative phases reach their stable fixed points at the minima of their corresponding potential functions. The potentials in (f) and (h) show that $\varphi_{C1\text{-}C2}$ and $\varphi_{C\text{-}V}$ do not attain their lowest possible energy states. The reason is that the gestural forces influencing each C system oppose one another: each C system experiences a force toward synchronization with V, but simultaneously experiences a repulsive force away from the other C. The final compromise corresponds to a balance between these forces. As expected, the φ_{CV} are equally displaced from their target ($\Phi_{CV} = 0$) in opposite directions.

To derive variability measures from the model, 500 simulations were run for each combination of parameters in Table 3 below. A random set of initial oscillator phases was generated for each simulation. Relative phases between C_1 and C_2, and between Phr and Ft, were extracted from each simulation only from points in time after $\varphi_{C1\text{-}C2}$ had stabilized. Furthermore, values were taken only from points when the Phr oscillator reached a minimum, which corresponds to the syllable *spa* (though similar results would be obtained for any point of time). In Fig. 8 two points in time where $\varphi_{C1\text{-}C2}$ would have been measured are indicated with arrows (note that each simulation ran longer than the example in the figure). Since there are several such points in each simulation, approximately 2500 φ measures were taken for each model.

Table 3. Standard deviations of φ_{PhrFt} and φ_{C1C2} from model simulations

Coupling	Coupling strength matrix A_{ij}	φ_{PhrFt} s.d.	φ_{C1C2} s.d.	Coupling	Coupling strength matrix A_{ij}	φ_{PhrFt} s.d.	φ_{C1C2} s.d.
(1) *no rhythm.-gest. weak Phr-Ft-σ*	$\begin{array}{cccccc} & 1 & & & & \\ 1 & & 1 & 0 & 0 & 0 \\ & 1 & & 0 & 0 & 0 \\ & & & & 1 & 1 \\ & & & 1 & & 1 \\ & & & 1 & 1 & \end{array}$	0.090	0.144	(2) *no rhythm.-gest. strong Phr-Ft-σ*	$\begin{array}{cccccc} & 2 & & & & \\ 2 & & 2 & 0 & 0 & 0 \\ & 2 & & 0 & 0 & 0 \\ & & & & 1 & 1 \\ & & & 1 & & 1 \\ & & & 1 & 1 & \end{array}$	0.046	0.151
(3) *weak rhythm.-gest. weak Phr-Ft-σ*	$\begin{array}{cccccc} & 1 & & & & \\ 1 & & 1 & 1 & 1 & 1 \\ & 1 & & 1 & 1 & 1 \\ & & & & 1 & 1 \\ & & & 1 & & 1 \\ & & & 1 & 1 & \end{array}$	0.100	0.124	(4) *weak rhythm.-gest. strong Phr-Ft-σ*	$\begin{array}{cccccc} & 2 & & & & \\ 2 & & 2 & 1 & 1 & 1 \\ & 2 & & 1 & 1 & 1 \\ & & & & 1 & 1 \\ & & & 1 & & 1 \\ & & & 1 & 1 & \end{array}$	0.047	0.125
(5) *strong rhythm.-gest. weak Phr-Ft-σ*	$\begin{array}{cccccc} & 1 & & & & \\ 1 & & 1 & 2 & 2 & 2 \\ & 1 & & 2 & 2 & 2 \\ & & & & 1 & 1 \\ & & & 1 & & 1 \\ & & & 1 & 1 & \end{array}$	0.098	0.089	(6) *strong rhythm.-gest. strong Phr-Ft-σ*	$\begin{array}{cccccc} & 2 & & & & \\ 2 & & 2 & 2 & 2 & 2 \\ & 2 & & 2 & 2 & 2 \\ & & & & 1 & 1 \\ & & & 1 & & 1 \\ & & & 1 & 1 & \end{array}$	0.046	0.088
(7) *weak rhythm.-gest. strong Phr-Ft-σ strong C_1-C_2-V*	$\begin{array}{cccccc} & 2 & & & & \\ 2 & & 2 & 1 & 1 & 1 \\ & 2 & & 1 & 1 & 1 \\ & & & & 2 & 2 \\ & & & 2 & & 2 \\ & & & 2 & 2 & \end{array}$	0.047	0.080	(8) *strong rhythm.-gest. strong Phr-Ft-σ strong C_1-C_2-V*	$\begin{array}{cccccc} & 2 & & & & \\ 2 & & 2 & 2 & 2 & 2 \\ & 2 & & 2 & 2 & 2 \\ & & & & 2 & 2 \\ & & & 2 & & 2 \\ & & & 2 & 2 & \end{array}$	0.047	0.067

By comparing models with relatively weak Phr-Ft-σ coupling (1,3,5) to those with relatively strong inter-rhythmic coupling (2,4,6,7,8), one can see in Table 3 that weaker coupling between rhythmic systems leads to greater variability in $\varphi_{Phr\text{-}Ft}$. Comparing degrees of rhythmic-gestural coupling strength between models (1,3,5), we see that φ_{C1C2} is most variable when no such coupling exists, is less variable when a weak coupling is added, and is least variable when rhythmic-gestural coupling is strong. The same holds for models (2,4,6). However, in comparing φ_{C1C2} between models (1,3,5) and models (2,4,6), where what differs is Phr-Ft-σ coupling strength, we see that stronger inter-rhythmic coupling has almost no effect on φ_{C1C2}. For this model to predict the experimentally observed association between rhythmic variability and gestural variability, an additional assumption must be made, namely that rhythmic-gestural coupling strength varies as a function of inter-rhythmic coupling strength. With respect to the model classes in Table 2, this means that $x_{1,2} = hb$, where b is the inter-rhythmic coupling strength and h is a constant of proportionality.

In general, the principle behind the observed patterns in these models is that random perturbations of phase velocity are reduced by coupling forces. Note that when intergestural coupling becomes stronger, as in models (7,8), the effect of rhythmic-gestural coupling on φ_{C1C2} becomes smaller. This occurs because the strongly-coupled gestural oscillators more quickly counteract fluctuations in phase velocity, in which case the driving forces from rhythmic systems contribute less to noise reduction. Variability patterns in the model can thus be seen to follow from the relative strengths of coupling and noise. If noise is relatively low, coupling strength has very little effect on variability of relative phase, because there is not much phase variability to begin with. As noise is increased, coupling plays a more important role in maintaining the stability of relative phase.

Although no attempt has been made here to model experimentally observed changes in synergistic relations across rhythmic tasks, one could accomplish this using the task dynamic model of multi-effector movements in Saltzman & Kelso (1983), Saltzman (1986), and Saltzman & Munhall (1989). This model employs an arbitrary weighting matrix to delegate components of motion to individual effectors. If distinct parameters govern coupling between rhythmic systems and the individual effectors associated with a gesture, then differences in rhythmic organization could alter synergistic relations. The details of this, however, remain to be worked out.

There are numerous ways in which the model presented above can be restructured or differently parameterized, some of which could be challenging—but not impossible—to distinguish experimentally. Regardless of which is correct, the importance of the model lies in its dynamical approach, particularly in its use of the concepts of coupling, synchronization, stability, generalized relative phase, and multi-timescale/multifrequency interaction. These concepts are general and will likely outlive major changes in specific instantiations of the model. They allow for a coherent understanding of an otherwise perplexing experimental observation: increased variability in intergestural timing co-occurring with increased variability in rhythmic timing. In model terms, this co-occurrence is observed because rhythmic and gestural systems are coupled and synchronize with each other, and moreover, the strength of such coupling varies according to the stability of rhythmic relative phase targets.

5. Conclusion and Future Directions

The experimental evidence reported here for interaction between rhythmic and gestural timing is important because conventional, non-dynamical models of speech production do not predict such effects. This evidence is seen in the co-occurrence of greater variability of timing between gestures and greater variability of timing between feet and phrases. Another finding, also not amenable to non-dynamical treatment, is that synergistic relations between effectors change systematically across rhythmic tasks. The variability patterns were modeled successfully using a multiscale system of coupled oscillators.

It is worth pointing out a glaring inconsistency between current dynamical models and what is more realistically occurring in cognitive planning systems. In all instantiations to date, the dynamical descriptions of planning systems employ limit cycles with constant amplitude parameters, which entails that the systems oscillate indefinitely. This is obviously false: oscillations must in reality be transient. Most models have used limiting fixed points of relative phase to make predictions about observables, but we must be open to the possibility that planning system amplitude variation occurs rapidly and that this has important consequences for what we observe. Moreover, in speech outside the lab, the "phrases" used here as fundamental timescales are not coherent entities, but rather composed of transiently active syntactic and semantic systems which presumably have their own complex dynamics.

Before future modeling efforts can address these issues, further investigation of interaction between rhythmic and gestural timing should be conducted. Some individual variation was observed in the present experiment, and understanding the sources and consequences of this variation is important. There are many ways in which the experimental design can be altered to probe rhythmic-gestural interaction from different angles, and experimental paradigms that refine the speech cycling task are needed. It is my hope that this study will spark future research into phenomena of interaction between speech rhythm and gesture.

Acknowledgements

This research would not have been possible without the generosity and guidance of Peter Watson and Ben Munson. Thanks to Keith Johnson, Rich Ivry, and Sharon Inkelas for conversations and discussions throughout the design and analysis of this experiment. Many thanks also to Hosung Nam for help in constructing the numerical model. Special thanks to Kim and Cade Tilsen.

Appendix A

Numerical simulations of the model were conducted using a 4th-order Runge-Kutta algorithm in Matlab. Each simulation ran for 100 seconds of model time, with 2^5 iterations per second. Initial phases of each oscillator were randomly taken from a uniform distribution on the interval [0, 2π]. Oscillator frequencies ω_i for the Phr, Ft, σ, C_1, C_2, V were 2π × [1, 2, 4, 4, 4, 4]. Phase and relative phase variables were wrapped on the unit circle; note that for clarity of conceptualization in Fig. 8 phases were depicted on the interval [-π, π]. All relative phase targets were 0, except for Φ_{C1C2}, which was set at 3π/2, or equivalently, -π/2. Each pair of oscillator frequencies ω_i and ω_j in Eq. (2) is divided by the lowest common denominator; in other words, for the purposes of defining a generalized relative phase, the frequencies of each oscillator are defined relative to the lower frequency oscillator in the pair. The Gaussian noise term η_i was modulated by the oscillator frequency and an amplitude parameter; hence the frequency terms in Eq. (1) can be rewritten $\omega_i\,(1 + a_\eta\,\eta_i)$. This reflects the idea that all oscillators should exhibit comparable variability over a single period, regardless of their frequency. The amplitude parameter a_η used in the simulations reported above was 0.2, which means that the standard deviation of the Gaussian noise was 0.4π radians. Substantially lower levels of noise lead to smaller effects of rhythmic-gestural coupling; larger levels of noise lead to larger effects, and much larger levels lead to non-stabilization of relative phases. The change of oscillator phase due to each coupling term was taken simply as the negative of the sine of the difference between relative phase and target relative phase, modulated by the coupling strength. It is common in more sophisticated models to use the inverse Jacobian to partition a change of relative phase into changes of phase in component systems—although not strictly necessary, the Jacobian strategy is more theoretically attractive because it distinguishes explicitly between change of oscillator phase and change of relative phase, which is more precisely what is described by the negative of the derivative of the potential.

References

Allen, G. (1972). The location of rhythmic stress beats in English: an experimental study, parts I and II. *Language and Speech*, 15:72-100, 179-195.

Allen, G. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3, 75–86.

Barbosa, P. (2002). Explaining Cross-Linguistic rhythmic variability via a coupled-oscillator model of rhythm production. *Proceedings of Speech Prosody 2002*.

Browman, C. & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, 45, 140-155.

Browman, C. & Goldstein, L. (1990). Tiers in articulatory phonology. *Papers in Laboratory Phonology I. Between the Grammar and Physics of Speech.* John Kingston/Mary E. Beckman, eds. Cambridge University Press, 341-376.

Browman, C. & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlee*, 5:25-34.

Cummins, F. & Port, R (1996). Rhythmic constraints on English stress timing. In *Proceedings of the fourth International Conference on Spoken Language Processing*, (H. T. Bunell & W. Idsardi (eds,), pp. 2036-2039. Wilmington, Delaware: Alfred duPont Institute.

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

Dauer, R (1983). Stress timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11, 51-62.

DeJong, K. (1994). The Correlation of P-center adjustments with articulatory and acoustic events. *Perception & Psychophysics*, 56:4, 447-460.

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. in *Action to Language via the Mirror Neuron System*, ed. Michael A. Arbib. Cambridge Univ. Press.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). *Dynamic Action Units Slip in Speech Production Errors*. Cognition, 103, 386-412.

Haken, H. (1983). *Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*, 3rd ed. New York: Springer-Verlag.

Haken, H. (1993). *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices*. New York: Springer-Verlag.

Haken, H., Kelso, J., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movement. *Biological Cybernetics*, 51, 347-356.

Haken, H., Peper, C., Beek, P., & Daffertshofer, A. (1996). A model for phase transitions inhuman hand movements during multifrequency tapping. *Physica D*, 90, 179-196.

Hoole, P. (1996). Issues in the acquisition, processing, reduction and parameterization of articulographic data. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation*, München, 34, 158-173

Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope. *Perceptual Psychophysics*, 43:1, 90-93.

Keith, W. & Rand, R. (1984). 1:1 and 2:1 phase entrainment in a system of two coupled limit cycle oscillators. *Journal of Mathematical Biology*, 20, 122-152.

Kelso, J. (1995). *Dynamic Patterns: the Self-Organization of Brain and Behavior*. MIT: Cambridge, Ma.

Kelso, J., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, 10:6, 812-832.

Kopell, N. (1988). Toward a theory of modeling central pattern generators. in *Neural Control of Rhythmic Movement in Vertebrates*. Cohen, A., Rossignol, S., & Grillner, S. (eds.), 369-413. John Wiley & Sons: New York.

Liberman, M. (1975). *The Intonational System of English*, Ph.D. dissertation, MIT, Garland Publishing Co., N.Y.

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry,* 8:3, 249-336.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, 83:5, 405-408.

Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. In *15th International Congress of Phonetic Sciences*, 3, 2253-2256.

O'Dell, M. (1995). *Intrinsic timing in a quantity language*. Licentiate Dissertation, Department of Linguistics, University of Jyväskylä, Finland.

O'Dell, M. & Nieminen, T. (1999). Coupled Oscillator Model of Speech Rhythm. In *Proceedings of the XIV International Congress of Phonetic Sciences,* San Francisco, 2, 1075-1078.

Pike, K. (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press.

Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). *Synchronization: A universal concept in nonlinear sciences.* Cambridge Press: Cambridge.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics*, 17, 175–192.

Port, R. (1986). Translating linguistic symbols into time. *Research into Phonetics and Computational Linguistics, Report No. 5*, Department of Linguistics and Computational Linguistics, Indiana University, Bloomington, Indiana.

Port, R., Cummins, F., & Gasser, M. (1995). A dynamic approach to rhythm in language: Toward a temporal phonology. *Technical Report No. 105, Indiana University Cognitive Science Program*, Bloomington, Indiana.

Port, R. (2003). Meter and Speech. *Journal of Phonetics*, 31, 599-611.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition,* 73, 265-292.

Repp, B. (1980). A range-frequency effect on perception of silence in speech. *Haskins Laboratories Status Report on Speech Research*, SR-61, 151-165.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: a preliminary model. *Experimental Brain Research Series*, 15, 129-144. Springer-Verlag: Berlin.

Saltzman, E. & Kelso, J. (1983). Skilled actions: a task dynamic approach. *Haskins Laboratories Status Report on Speech Research*, SR-76, 3-50.

Saltzman, E. & Kelso, J. (1987). Skilled actions: a task-dynamic approach. *Psychological Review*, 94, 84-106.

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

Saltzman, E. & Byrd, D. (1999). Dynamical simulations of a phase window model of relative timing. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey, (Eds.). *Proceedings of the XIVth International Congress of Phonetic Sciences*, 2275-2278. New York: American Institute of Physics.

Saltzman, E. & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19, 499-526.

Saltzman, E. & Byrd, D. (2003). The elastic phase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31, 149-180.

Saltzman, E. Lofqvist, A., Kay, B. Kinsella-Shaw, J. & Rubin, P. (1998). Dynamics of intergestural timing: a perturbation study of lip-larynx coordination. *Experimental Brain Research*, 123, 412-424.

Saltzman, E., Lofqvist, A., & Mitra, S. (2000). “Glue” and “clocks”: intergestural cohesion and global timing. In M. Broe and J. Pierrehumbert (eds.) *Papers in Laboratory Phonology*, 5, 88-101. Cambridge, UK: Cambridge University Press.

Scott, S. (1993). P-centers in Speech: An Acoustic Analysis. *Unpublished doctoral dissertation*, University College London.

Scott, S. (1998). The point of P-centres. *Psychological Research*, 61, 4-11.

Sternad, D., Turvey, M., & Saltzman, E. (1999). Dynamics of 1:2 coordination: sources of symmetry breaking. *Journal of Motor Behavior*, 31:3, 224-235.

Winfree, Arthur T. (1980). *The Geometry of Biological Time*. Springer-Verlag: New York.