Rhythmic constraints on stress timing in English

Fred Cummins[1] and Robert F. Port[2]

Suggested running title: Rhythmic constraints on stress timing

[1]Department of Linguistics, 2016 Sheridan Road, Northwestern University, Evanston, IL 60208
[2]Indiana University, Department of Linguistics and Cognitive Science Program, Bloomington, IN 47405

Abstract

Rhythm in speech is interpreted as the hierarchical organization of temporally coordinated prosodic units. This is a departure from conventional phonetic approaches which have focussed on isochrony. Speech Cycling, a novel experimental vehicle for investigating rhythmic constraints, is introduced. In a speech cycling task, subjects repeat a phrase together with a periodic stimulus and the temporal distribution of onset events (beats) is examined. Two speech cycling experiments are presented which probe the degree to which the relative durations of inter-stress intervals within a series of repeated phrases are independent. The data reveal the presence of strong rhythmic constraints on stress timing which have hitherto eluded experimenters. It is argued that these constraints are evidence for a task-specific dynamical system in which prominent events (stress beats) are constrained to occur at specific, predictable, phases of an enclosing cycle. The dynamical system is characterized by entrainment between metrical levels, a principle which underlies rhythmic coordination in other activities such as locomotion. It is demonstrated that the hierarchical nesting of metrical levels arises inevitably within a repetition task, and what is more, the construction of individual prosodic units depends to some extent on the resulting rhythmic pattern. By structuring an utterance so that prominent events lie at privileged phases of a higher-level prosodic unit, rhythm is seen as an organizational principle which has its roots in the coordination of complex action and its effect in the realm of prosodic structure.

1 Introduction

This paper presents experimental evidence that timing constraints based on rhythm are operative in the production of speech. Rhythm is viewed here as the hierarchical organization of temporally coordinated prosodic units. We show that certain salient events (beats) are constrained to occur at particular phases of an established period, and we develop an argument that the establishment of this period serves a coordinative function. This is a radical departure from the conventional treatment of rhythm in phonetics, which has been concerned primarily with the search for the degenerate rhythm of an isochronous series.

Numerous phoneticians have suggested that English speech tends to exhibit roughly equal inter-stress intervals (Jones, 1918; Pike, 1945; Abercrombie, 1967). The first instrumental search for isochrony in English was carried out by Classé (1939). Using the recently developed kymograph, he measured the intervals between the onsets of nuclear vowels of stressed syllables. Strict isochrony was only observable when the rhythmic groups had similar numbers of syllables with similar segmental content and carefully matched grammatical structure (1939, p.85). Subsequent studies have not been kinder to the isochrony hypothesis (Shen and Peterson, 1962; Dauer, 1983; Jassem, Hill and Witten, 1984; Crystal and House, 1990). This is so, despite the strong impressions listeners have that there is often a rhythmic basis to speech. Some have concluded that the perception of rhythmic regularity in speech is a construct, or illusion, probably based on the well known general tendency of the perceptual apparatus to regularize in the face of noise (Lehiste, 1977; Dauer, 1983; Benguerel and D'Arcy, 1986), and, specifically, to hear temporally patterned events as rhythmically structured whether they are or not (Woodrow, 1951; Fraisse, 1956).

In a comprehensive review of work until 1977, Lehiste (1977) listed some of the reasons why many studies had failed to document isochrony. The studies reviewed used a wide range of speech styles ranging over prose and verse. Researchers did not use a uniform methodological framework within which isochrony could be identified. For example, Shen and Peterson (1962) regarded each sentence as having one and only one primary stress, while O'Connor (1965) identified as many as 15 stress groups within a single limerick.

Another serious problem with early searches for isochrony is that it is never clear how much deviation from strict isochrony might be permissible, and perhaps required, for sentences to be perceived as regular. Benguerel and D'Arcy (1986) assessed the degree to which deviation from regularity, in the form of phrase-final deceleration, may actually be *necessary* for the perception of regularity.

Two further methodological issues might be raised. First, the measurement point, which for most researchers has been the onset of the nuclear vowel of the stressed syllable, needs some amending in the light of the discovery of the P-center ("perceptual center") or stress beat. Under the former name, Morton, Marcus and Frankish (1976) pointed out that series of syllables with isochronous acoustic onsets are not necessarily heard as isochronous. In common with Allen (1972), who used the term stress beat, they found that the perceived beat of a syllable was located near to the onset of voicing, but occurred earlier as a function of the length of the initial consonant cluster, and later as a function of the length of the syllable coda. Models which predict the location of the stress beat have since been proposed (Marcus, 1981; Scott, 1993). Though we develop and apply a variant of the model proposed in Scott (1993), P-center effects are an order of magnitude smaller than observed anisochrony in speech.

A second possibility which has received less attention is that some beats are silent, or marked only by additional pause length. Espoused by Abercrombie (1965,1967), this possibility should lead one to look for intervals, not only of length n, but $2n, 3n$ etc. This possibility is missed, for example, by Scott, Isard and de Boysson-Bardies (1985) who derived a measure of irregularity with which they look for evidence of isochrony. The measure they adopted is minimized by successive intervals whose ratios are $1:1$, and is increased by deviations from this ratio. The result is that successive intervals which are related as, e.g., $1.837:1$ are regarded as *more* regular than intervals related as $2:1$, as would be found in a perfectly isochronous series with one silent beat. Failing to consider silent stress beats could have led researchers to prematurely reject the isochrony hypothesis. If silent beats are admitted to the fold, it is obvious that speech rhythm may turn out to be more complex than the degenerate rhythm of a simple isochronous series.

In fact, we do not believe either of the latter two points are likely to cause a complete reinterpretation of the empirical work which has so far failed to document isochrony in English speech. Classé was right to observe that isochrony is likely to be very rare in unconstrained speech. But is does not warrant the present situation where some investigators have treated speech rhythm as if it were an empirically supported or supportable phenomenon (Couper-Kuhlen, 1993), while others, such as Benguerel and D'Arcy, have pointed out that perceptual regularity by no means entails regularity of the physical stimulus, and have suggested that speech rhythm might simply be unobservable (Benguerel and D'Arcy, 1986, p. 244).

Besides failing to agree on the basic units which might constitute a rhythm, these studies also missed something essential to rhythmic structure, that is entrainment or temporal constraint *across levels*. While isochrony constitutes rhythm by virtue of a single recurrent period, a hierarchical rhythmic structure is potentially much richer, with temporal constraints operative across levels.

The present approach to rhythm and structure in language is closer in spirit to the theory of metrical phonology, especially as cast in Liberman's original thesis work (1975). The original conception of the metrical grid was as an organizing principle which underlay diverse forms of complex behavior, including speech, music and dance. Liberman formulated this as as the Metrical Organization Hypothesis: "All temporally ordered behavior is metrically organized" (Liberman, 1977, p. 204). Subsequent work tended to emphasize language structure at the expense of coordinative principles underlying behavior (Liberman and Prince, 1977; Hayes, 1985). The divorcing of theory and experiment appears almost complete with Hayes' statement that the phonetic reality for the metrical trees of his thesis is essentially nonexistent (Hayes, 1985, p. 16), while experimental investigation of the claimed structures failed to support such basic notions as stress retraction (Cooper and Eady, 1986).

We believe the simple isochrony model of phonetics needs to be replaced, while the hierarchical organization claimed by phonologists needs experimental verification. We make two novel claims which we believe are required to arrive at a plausible model of speech rhythm. The first claim is that *rhythm is manifested as the temporal binding of events to specific and predictable phases of a superordinate cycle*. Formally, this is a claim that rhythm is hierarchical, and that elements low in the hierarchy will nest an integral number of times within higher elements. The phases at which events will occur arise directly from this integral nesting relationship; for example a nesting of two units within one larger one will produce events at phases 0.0 and 0.5 of the larger cycle.

The second claim to be made concerns the function of rhythm in speech. Our claim is that *rhythm in speech is functionally conditioned*. It emerges under just those speaking conditions in which a tight temporal coordination is required between events spanning more than one syllable. Linking disparate motor components together into a single temporal structure, or rhythm, greatly simplifies the problem of coordination among the many parts. In order to motivate this assertion, it will be necessary to look at the role played by rhythm in other forms of complex coordinated activity.

This view of the function of rhythm has been best developed in the study of inter-limb coordination, both in locomotion and in manual tasks (Schöner and Kelso, 1988a; Diedrich and Warren, 1995), and exemplifies the task dynamic approach to the control of action (Saltzman and Kelso, 1987; Meijer and Roth, 1988; Bingham, Schmidt, Turvey and Rosenblum, 1991). Within the task dynamic framework, the entire biomechanical apparatus involved in a particular action is understood to function as a single, task-specific, device (Kugler, Kelso and Turvey, 1980; Kelso, Saltzman and Tuller, 1986). From the myriad of complex individual components, a coherent action emerges as a low-dimensional system. The emergence of coherence is only possible because the components have been constrained to act together—they have been *coordinated* (in contrast to the *control* postulated in a motor program approach). In this sense, rhythm can be seen as a coordinative strategy; parts which together produce a rhythm are constrained in their relative timing, reducing the number of degrees of freedom of the system (Bernstein, 1967; Kugler et al., 1980). By way of analogy, consider the four wheels of a car, each of which could in theory point in any given direction. These four degrees of freedom are constrained by the chassis and steering assembly to effectively point along a single trajectory, which in turn is determined by the single degree of freedom of the steering wheel, thus greatly simplifying the control task of the driver.

Speech production requires the coordination of a huge number of disparate biological components, with intrinsic timescales ranging over about three orders of magnitude (glottal vibration, supralaryngeal articulation, respiration). Despite the 'slop' attributable to lazy articulation, rate variation, affective factors, etc., the final result is enormously detailed and rich in information. The coordinative principles which enable this feat are what we aim to uncover.

A productive strategy for studying the coordination of complex action systems can be found in the task dynamic approach to rhythmic coordination among the limbs (Kugler et al., 1980; Kelso, Holt, Kugler and Turvey, 1980; Schöner and Kelso, 1988a). The experimental strategy employed within the task dynamical approach to the study of action can be roughly summarized as the establishment of a strictly controlled task (setting up of boundary conditions), the discovery of a simple collective variable which succinctly captures the macroscopic properties of the system, and the experimental manipulation of that variable through a control parameter. A long series of studies by Kelso and coworkers has examined a model system which requires the rhythmic coordination of the limbs (Kelso, Southard and Goodman, 1979; Kelso, 1995). Within this paradigm, two fingers or hands (or arbitrary effectors) are wagged cyclically toward and away from the body's midsaggital plane. The wagging task sets very well defined boundary conditions for the action system. Given these constraints, the state of the entire movement system can be economically described by noting only the relative phase, ϕ, of the two cycling effectors, that is, the difference between the phases of the two fingers/hands. This collective variable can now be studied under various experimental manipulations. Most of the initial studies involved the experimental control of rate using a pacing metronome which was gradually stepped up. Thus a control parameter (rate) is used to influence the state of the complex system as described by a low dimensional collective variable (relative phase of the fingers or hands) within well defined boundary conditions (the wagging task).

The principal findings of this research are that subjects have a strong preference for a synchronous phase relation between the fingers or hands, where 'synchrony' means that the limbs move toward and away from the midline simultaneously, with no phase lag ($\phi = 0$). The anti-synchronous phase relation ($\phi = 0.5$), where both move left and then right, is less stable—but much more stable (small variance, insensitivity to perturbation) than any other phase angle between the limbs (Kelso and Kay, 1987; Kay et al., 1991). Furthermore, while both the synchronous and anti-synchronous patterns are stable at slower tempos, an increase in tempo eventually leads to a situation in which only synchrony is stable. Study of the stability properties of each production mode and of the transition between stable modes suggests the existence of an underlying dynamic which is parameterized by rate. The system exhibits two competing attractors at slower rates and a single attractor at fast rates. A detailed mathematical model which derives the observed stabilities and transition properties from two component oscillators (roughly, the fingers/hands) which are non-linearly coupled has been developed. It continues to generate rich predictions about the system behavior (Scholz, Kelso and Schöner, 1987; Scholz and Kelso, 1990; Kelso and Jeka, 1992).

One important variant of the basic experimental task is presented in Yamanishi, Kawato and Suzuki (1980), and, in slightly different form, in Tuller and Kelso (1989). In each case, subjects were asked to tap two fingers in time with an external stimulus—one flashing light per finger—with a fixed phase lag given between the two lights. The entire range of possible phase relations between the fingers can be probed over a series of trials by scanning the target through the range of possible phase lags. The goal was to see if the known stabilities of the autonomous movement system (the two fingers in the absence of an external pacing signal) were still evident when intermediate phase relations were specified by the environment. In Tuller and Kelso's experiment, the stimulus was present during tapping; in that of Yamanishi et al., subjects attempted to reproduce a phase lag from memory. The results, however, were remarkably similar. The most important findings were, first, that subjects tended to produce phases which were biased towards the known stable states of synchrony and anti-synchrony. That is, target phases were not reproduced accurately; rather, for targets close to 0.0 or 0.5 the produced phases tended to be between the targets and the stable states. Second, the variability of subjects' productions was lowest around these two values.

A central theme of the above work is that the rhythm (the fixed phase relationships between the components of the system) arises from the coupling between the components, while the coupling greatly simplifies the task of coordination among those components. The rhythmic constraints are thus evidence of the coordinative mechanism underlying a task specific device. Beyond this fundamental observation, the experimental procedure of Yamanishi et al. (1980) and Tuller and Kelso (1989) will be readily seen to be closely related to the experiments detailed herein. The importance of this particular type of experiment lies in the observation that the dynamics of the autonomous system (the movement system comprising two oscillating fingers or hands) are apparent even in the presence of the externally imposed timing signal. Observed phases gravitate towards attractors of the system and variance is lowest for targets near those attractors.

In the following work, our strategy will be to infer the attractor states of the speech production system in a given task from data obtained in the presence of an external timing signal. These attractors will be seen to correspond to readily identifiable rhythmic patterns. For this purpose a model task will be presented, the speech cycling task, which serves to establish a well-defined cycle at the phrase level.

2 General method: introducing speech cycling

We here introduce a novel experimental paradigm for the study of speech coordination which is analogous in many respects to the empirical work on coordination among the limbs described above. The basic strategy employed is to elicit repeated phrases in time with an auditory stimulus.

Manipulation of timing information in the stimulus is related to the time of specific events (beats) in the subject's speech.

In a speech cycling task, subjects are asked to repeat a short phrase along with an auditory stimulus. The simplest form requires subjects to repeat a short phrase together with a perfectly isochronous series of beeps. Under almost all circumstances, a "Harmonic Timing Effect" is observed, whereby the onsets of stressed syllables are found to lie at points within the overall phrase repetition cycle which divide the cycle into simple integer ratios (Cummins, 1997). This result is equivalent to the observation of Classé (1939) that a succession of closely matched phrases will, indeed, exhibit a regular rhythm. Beats at $\phi = \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}$ etc. of the cycle are regularly found. Any and all of these cases can be interpreted as examples of isochronous series if, as Abercrombie suggested, the series can include silent beats as well (Abercrombie, 1965).

An alternative interpretation which we propose, is that these simple divisions of the repeating cycle result from the emergence of a harmonic relationship between the cycle of the metrical foot and the phrase repetition cycle (PRC). By 'harmonic,' we mean the establishment of two periodicities which are related as simple multiples. This interpretation requires that we posit both the foot (operationalized here as the interval between stress beats) and the PRC as units in the production of speech under these conditions. As production units, they each have an intrinsic dynamic, and these dynamics can be mutually coupled (Port, Cummins and Gasser, 1995).

This dynamical interpretation of the speech cycling task suggests why isochrony should be more readily observable within the confines of a repetition task than in normal conversational speech. Repetition generates a stable cycle to which nested processes can (or must) entrain. The changing demands of unconstrained speech production do not allow this stability to persist, though it would emerge occasionally as the speech content permitted. In its simplest form, the speech cycling task provides a stable period, the PRC, within which nested and entrained periods can be seen to emerge.

A further development of this strategy, adopted here, is to provide two target tones in the repeating stimulus. The relative timing of one tone with respect to the other can be precisely varied, thus providing a target phase relationship for two speech events.

The foregoing has begged the question of measurement. Our algorithm for locating the "beat" of a syllable is given in the appendix. In essence, we examine the energy over the frequency range of the first two formants to identify the sonority rise at the onset of the nuclear vowel. A beat is defined as occurring halfway through this rise. This gives a consistent measure which can be largely automated[3]; it is similar to the P-center location algorithm of Scott (1993). An example of the beats determined using this procedure is given in Figure 1.

3 Experiment 1a: phrases as simple rhythmic structures

Consider the phrase "beg for a dime" repeated continuously. The phrase repetition cycle (PRC) is the interval from the beginning (or first beat) of one phrase to the beginning of the next. The monosyllables "beg" and "dime" each receive some degree of stress, so that we can identify two metrical feet within the PRC: one from "beg" to "dime" and another from "dime" to "beg". Under most circumstances, repeating this phrase will cause "dime" to fall halfway through the PRC, producing a reasonably isochronous series of beats. This first experiment probes the degree to which other,

[3] Code for beat extraction using this algorithm can be obtained on request from the first author at `f-cummins@nwu.edu`.

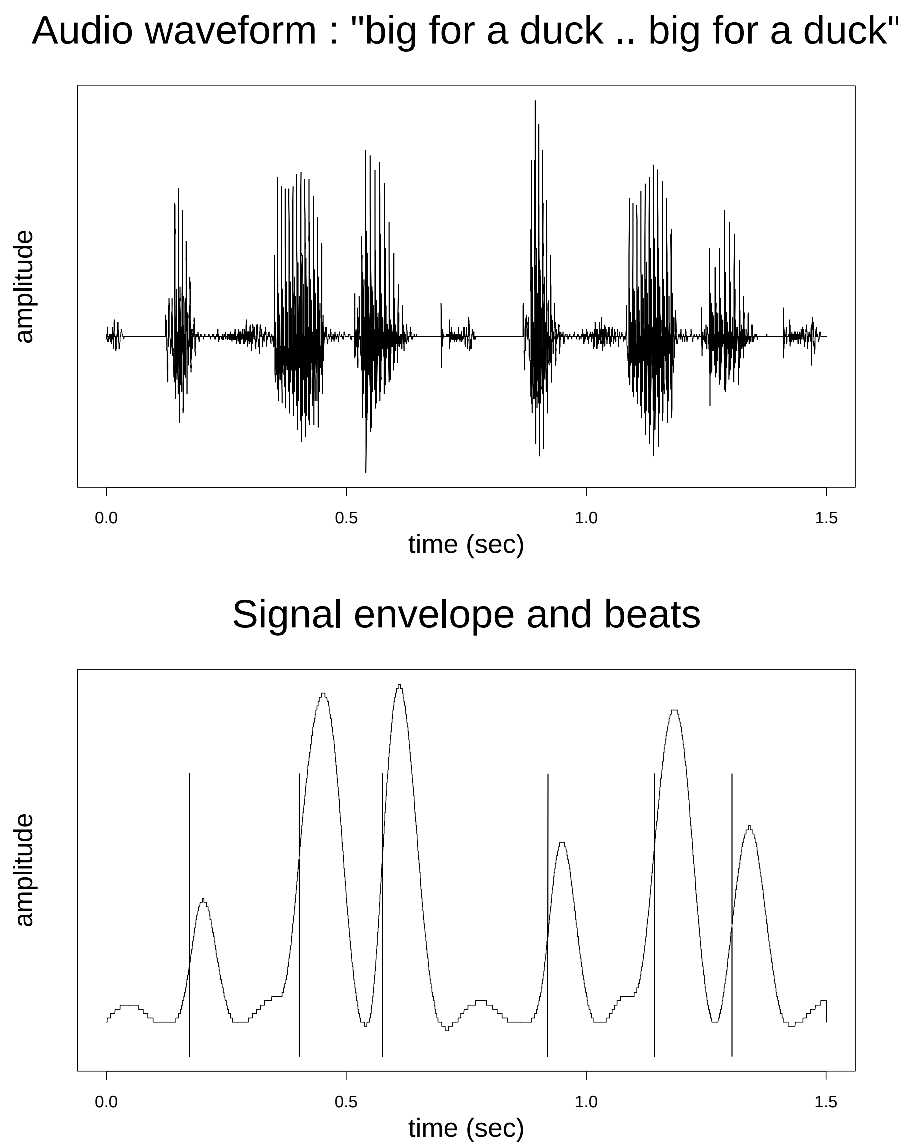

Figure 1: Beats extracted from the phrase "Big for a duck" repeated twice. Top panel: Acoustic signal. Lower panel: smoothed amplitude envelope and beats. For the task reported here, only the first and third beat of each group of three were considered. This utterance was obtained in the course of Experiment 2 herein, where the set interval between high and low beeps was 450 ms.

less rhythmical, patterns are possible. Following Tuller and Kelso (1989), subjects repeat a phrase together with an external pacing signal. As in Yamanishi et al. (1980), they also repeat the phrase after the stimulus is switched off.

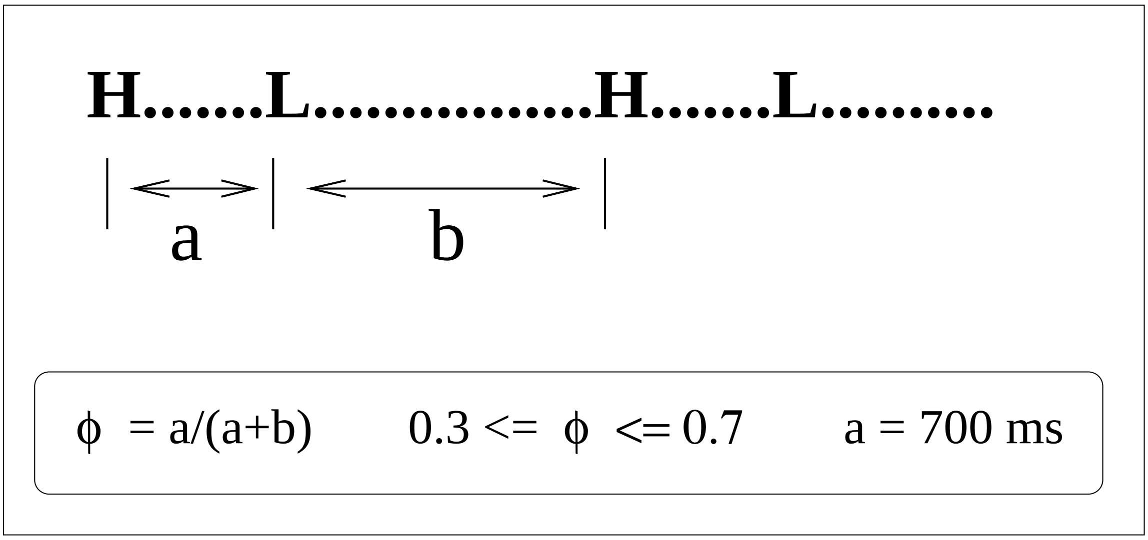

Figure 2: Stimulus used in Experiment 1. A succession of 14 pairs of alternating high (H) and low (L) tones are used. The interval from high to low tone, *a*, is fixed at 700 ms. The target phase, or relative time of occurrence of the low tone, is set by manipulating the interval from low to high tone, *b*.

3.1 Method

3.1.1 Stimuli

The stimuli in this experiment were sequences of 14 pairs of alternating short tones. The initial tone was a 1200 Hz tone of 50 ms duration which was sinusoidally ramped over its initial and final 10 ms to avoid transients. The lower tone was similar, at 600 Hz. The interval between the high and low tones was fixed throughout this experiment at 700 ms. The independent variable manipulated was the relative time of the low tone within the cycle defined by the high tones. This relative time is measured using a phase convention of $0 < \phi < 1$, so that a phase of 0.5 would mean that the interval between low and high was the same as that between high and low. For each trial, a target phase value (ϕ_{target}) was selected at random from a uniform distribution between 0.3 and 0.7. The low–high interval was then calculated such that the low tone now occurred at the target phase of the cycle whose end points were defined by the succession of high tones (See Figure 2). This gave a high–high cycle length within the range 1.0 sec (for a target phase of 0.7) to 2.333 sec (for a target of 0.3). Note that a different target is generated on each trial, and all targets between 0.3 and 0.7 are equally likely to occur. Stimuli were played at a self-selected comfortable listening level over headphones. The intensity of the last two pairs of tones were scaled down by factors of 0.66 and 0.33, respectively, so that the tones faded out rather than stopping abruptly.

3.1.2 Speech materials

A corpus of 30 short phrases with essentially identical prosodic structure was compiled. Each phrase was of the form *X for a Y*, where *X* and *Y* were each CVC words subject to the following constraints. The initial consonant was one of /b, d, g/. These were selected because the P-center literature suggests that the beat of a syllable is particularly close to the vowel onset when the initial consonant is a voiced stop (Fowler, 1977; Scott, 1993). We also wished to avoid partial devoicing of the vowel onset, as might occur with voiceless stops, as the measurement procedure outlined above is dependent upon voicing. The vowel was either short, from the set {ɛ, ɪ, ʌ}, or long, from the set {i, aɪ, eɪ, oʊ, u, aʊ}. The final consonant came from the set {p, m, f, v, t, d, n, s, z, k, g}. Long and short vowels were counterbalanced with the three initial consonants, and all words were real English words, though the resulting phrases were usually meaningless. This arrangement ensured that our results were not dependent on the narrow segmental makeup of the phrases. Sample phrases include "big for a duck" and "geese for a duke". A programming oversight resulted in a non-random pairing of phrases with targets, with the result that smaller target phases were more likely to be associated with long vowels in the final position. For this reason, no attempt is made to identify effects due to the segmental composition of the words.

3.1.3 Procedure

Four subjects took part, three female and one male, between 20 and 34 years of age. All were native speakers of regionally unmarked Midwestern dialect of American English. Subjects first filled out a standard questionnaire which established dialectical history, musical experience and experience with other rhythmic tasks such as juggling, ballroom dancing etc. The three female subjects were all accomplished musicians. The male subject was a non-musician.

On each trial, subjects were presented with one of the phrases on a screen. They were instructed to listen to the first two pairs of tones, and then to join in, repeating the given phrase in time with the tones, such that the first word of the phrase lined up with the higher of the tones, and the last word lined up with the lower of the tones. They were to attempt to line up their productions as accurately as possible with the stimulus. The stimulus consisted of 14 pairs of tones, so 12 repetitions (less any breath pauses, as described below) were obtained together with the stimulus. After the stimulus stopped, subjects continued to repeat the phrase for approximately another 12 repetitions, trying to maintain the timing pattern established together with the stimulus. They were then signaled to stop.

Initial pilots had shown that allowing subjects to breathe in an unconstrained fashion sometimes resulted in a small gasp-like breath being taken on every cycle, which introduced a bias in the timing of events within the cycle. For this reason, subjects were instructed to breathe in, then repeat the phrase, and to skip a whole cycle when they next needed to breathe. They were trained in this breathe—repeat—breathe—repeat task until they felt that they could concentrate again on the task at hand while fulfilling the breathing constraint.

Each subject completed 30 trials per session and 3 sessions on three different days. Each experimental session began with a practice run in which the breathe—repeat task was rehearsed. There were thus 90 trials per subject, with approximately 24 repetitions per trial, less breath pauses.

3.1.4 Measurement

Beats for the initial and final syllables of each phrase were automatically extracted using the algorithm given in the appendix, and then checked against a visual display of beat and amplitude envelope. Where automatic extraction failed (very few cases), measurements were made by hand using a display of both the amplitude envelope and the audio waveform. Phase measurements for the medial beat (*duck* in *big for a duck*) were computed with reference to the beats measured for the preceding and following phrase onset (in this case, *big*). All measurements were thus taken from speech and not

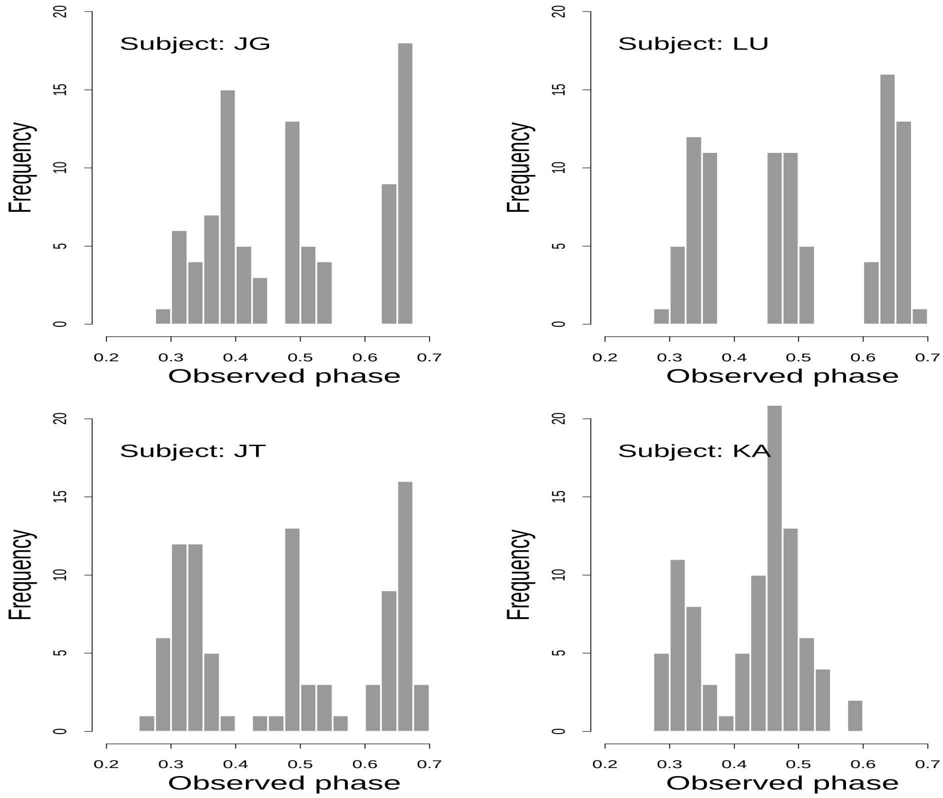

Figure 3: Experiment 1a. Distributions of $\hat{\mathbf{\Phi}}$ for phrase final syllables. Each data point is a trial median, with about 18 observations per trial. There are 90 trials per subject. Target phases were drawn from a uniform distribution between 0.3 and 0.7.

from the stimulus. No phase measurements were made for cycles during which subjects stopped to breathe. Initial examination of the time series data for each trial suggested edge effects for the first one or two repetitions, with stationarity reached thereafter. Occasionally, the stimulus cessation would also cause a particularly variable interval. For these reasons, the first two phase measurements both with and without the stimulus were discarded[4]. The task induced some speech errors. Where these were very obvious, the data were excluded, but the remaining data included a small number of apparent outliers arising from occasional dysfluencies during repetition. Because of the presence of outliers, the median phase per trial ($\hat{\mathbf{\Phi}}$) is used below rather than the mean.

3.2 Experiment 1a: results

3.2.1 Overall distribution of observed phases

The first and simplest hypothesis to be tested is that there are no rhythmic constraints on speech production. If this is so, subjects should be able to locate the onset of the final stress at any given point within the PRC. Their productions should thus mirror the distribution from which the targets were drawn. Figure 3 presents the distribution of $\hat{\mathbf{\Phi}}$, median phase per trial, for each subject. There

[4]This variability would have been interesting in its own right, had more than two points per trial been available.

are 90 trials per subject, with about 18 observations per trial median. In no case do subjects produce anything like the uniform distribution from which the targets are drawn. Rather, the histograms are markedly multimodal, with three clear modes in three cases and probably two in the fourth (subject KA).

3.2.2 Presence vs absence of stimulus tones

Within each trial, approximately half of a subject's productions were obtained together with the auditory stimulus and half were obtained after stimulus cessation, analogous to the procedures of Tuller and Kelso (1989) and Yamanishi et al. respectively. In a previous experiment (Cummins and Port, 1996b), we observed no difference in mean accuracy under these two conditions. This result was surprising enough to bear replicating, as we had initially hypothesized that subjects would be more accurate in their productions when the stimulus was present. In order to test this, two *inaccuracy* scores for each trial were computed by taking the median absolute difference between observed and target phases for repetitions produced along with the stimulus and the same median for the repetitions produced without the tones. This provides naturally paired observations which can be given a simple t-test. Table 1A lists differences, t-scores, degrees of freedom and p-values. Although our initial intuitions were that subjects should be more accurate when the stimulus was present, our negative findings from the previous experiment led us to use a two-tailed test. Indeed, the results are inconclusive, as subject JT appears more accurate with the stimulus and LU is more accurate without. Certainly there is no strong or consistent effect of presence of stimulus tones on median trial accuracy.

In similar vein, variability, expressed here as the interquartile range, can be computed for each half of a trial and a paired t-test done. This is summarized in Table 1B. Only Subject KA shows a significant effect, in producing more highly variable data when the stimulus is present.

In the light of these results, the data were pooled across stimulus condition, so that the median phase calculated over all repetitions within a single trial was used.

A. Median inaccuracy with and without stimulus				
Subject	mean difference	t	df	p
JG	-2.46e-3	-1.47	89	>0.05
LU	**2.88e-3**	**2.04**	**89**	**<0.05**
JT	**-8.63e-3**	**-3.72**	**89**	**<0.001**
KA	4.13e-3	1.62	89	>0.05
B. Interquartile range of inaccuracy with and without stimulus				
Subject	mean difference	t	df	p
JG	1.6e-3	1.64	89	>0.05
LU	1.28e-3	1.34	89	>0.05
JT	1.72e-3	1.70	89	>0.05
KA	**3.93e-3**	**2.74**	**89**	**<0.01**

Table 1: Experiment 1a. A: t-tests comparing median absolute inaccuracy within a trial for data collected with and without the stimulus. All tests are two-tailed. B: t-tests comparing interquartile ranges for inaccuracy with and without stimulus.

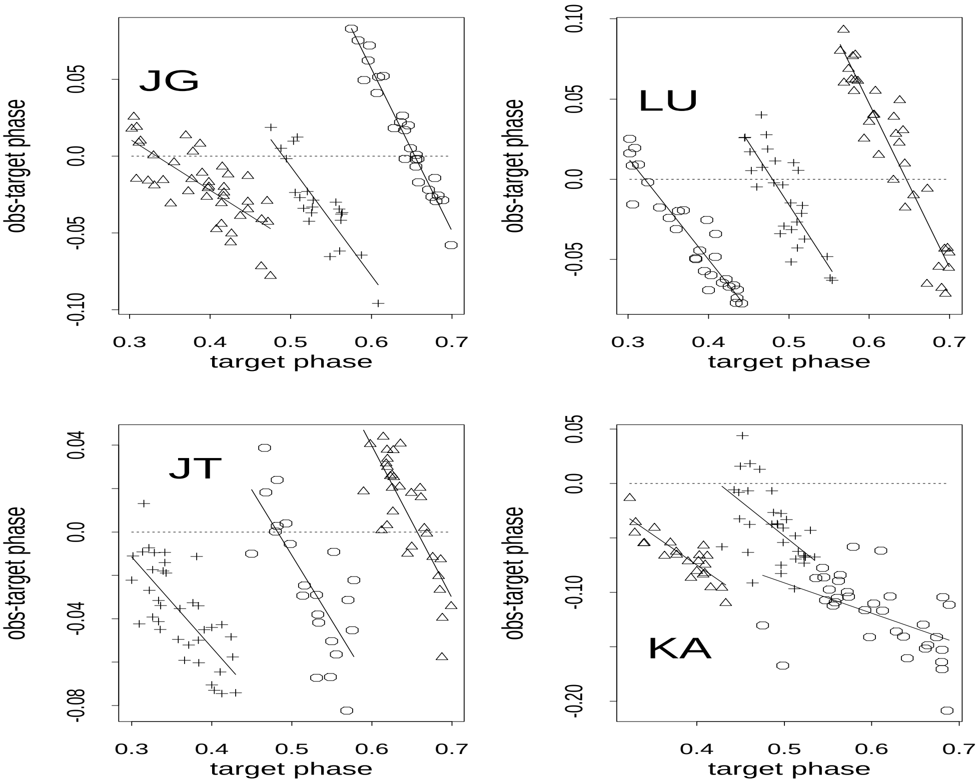

Figure 4: Experiment 1a. Difference between $\hat{\mathbf{\Phi}}$ and ϕ_{target} for each trial plotted as a function of the target phase. Perfect performance would yield data points on the line $y = 0$ (dotted line). Clusters have been automatically extracted from each data set and a local regression line fitted. Different plotting symbols have been used for data points within each cluster.

3.2.3 Data clustering

The histograms of Figure 3 suggest that each subject is producing a small number of distinct patterns, presumably as a function of target phase. Figure 4 shows the deviation of the median phase per trial from the target phase as a function of the target phase. Perfect performance would result in data points on the line $y = 0$. Instead, we see that the data cluster into distinct groups, corresponding approximately to the modes of the histograms. Within each cluster the data appear to be linearly related.

Clusters were estimated for each subject using the MCLUST model-based algorithm (Banfield and Raftery, 1993)[5]. This is a maximum-likelihood procedure in which it is possible to independently constrain cluster orientation, shape and volume to be alike (or freely variable) across clusters. As applied here, clusters were constrained to have similar elliptical shapes and areas, but their orientation was free to vary. This parameterization is consistent with local grouping around lines of unknown slope. The clustering algorithm proceeds by agglomeration, producing an optimal clustering for each

[5]Thanks to Adrian Raftery and Chris Fraley for advice and provision of up-to-date code for the clustering techniques applied herein.

value of 1–N where N is the total number of points. In order to select the best number of clusters, all choices of 2 to 5 clusters were examined for each subject and that clustering was chosen which maximized the weighted sum-of-r-squared values, computed as

$$\mathcal{R}_k = \sum_{i=1}^{k} w_i r_i^2 \tag{1}$$

where w_i is the proportion of data points in the ith cluster, and r_i^2 is the proportion of variance within that cluster which is accounted for by a linear fit. Furthermore, no clustering which produced any individual cluster of less than 5 data points was considered.

For the optimal clustering, Table 2 lists r^2, n and p for each within-cluster regression, and also gives the value of the intercept with the x-axis. All p values are highly significant. The local regression lines within each cluster are included in Figure 4.

sub	x-int	r^2	n	p	x-int	r^2	n	p	x-int	r^2	n	p
JG	0.333	0.553	41	<0.001	0.491	0.747	22	< 0.001	0.654	0.951	27	<0.001
LU	0.321	0.891	29	<0.001	0.48	0.669	27	<0.001	0.646	0.875	34	<0.001
JT	0.272	0.538	37	<0.001	0.482	0.566	22	<0.001	0.657	0.644	31	<0.001
KA	0.173	0.24	34	< 0.01	0.264	0.745	23	<0.001	0.424	0.31	33	< 0.001

Table 2: Regressions in Experiment 1a. For each data set clusters have been automatically extracted and local regression lines fitted. Listed here are values of the x-intercept, r^2, n and p for each within-cluster regression.

Each of the local regression lines has a negative slope. When the target phase lies below the x-intercept of the regression line, the phases produced are larger than the target, while targets larger than the x-intercept elicit smaller values. Within a cluster, then, all productions are biased towards the value of the x-intercept. This constitutes strong evidence for the existence of attractors located roughly at the x-intercepts, with all points of a cluster lying within one basin of attraction.

It can be seen that each of subjects JG, LU and JT exhibits three well-separated clusters. In 8 of 9 cases, these are located within 0.02 of the fractional values $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$. Only the lowest cluster of JT's data yields a different intercept (0.272). For subject KA (the only non-musician in the group) the data do not partition as neatly. The algorithm provided above finds three clusters, and while the associated regression lines all have negative slope, the intercepts do not permit simple interpretation. In particular, there is no evidence at all for the presence of an attractor above 0.5.

The data from subject JG provide an informative example of clusters which overlap in the x-direction. For a target phase within a small range (approximately 0.58–0.61) the system may be pulled toward one or other of two attractors.

3.2.4 Variance as a function of target phase

The task in this experiment is analogous to that of Tuller and Kelso (1989) and Yamanishi et al. (1980). In both those experiments, the investigators found that the variance in produced phase was lowest around the stable states of the system (synchrony and anti-synchrony) and higher elsewhere. Figure 5 plots the per-trial interquartile range of the produced phase as a function of the target

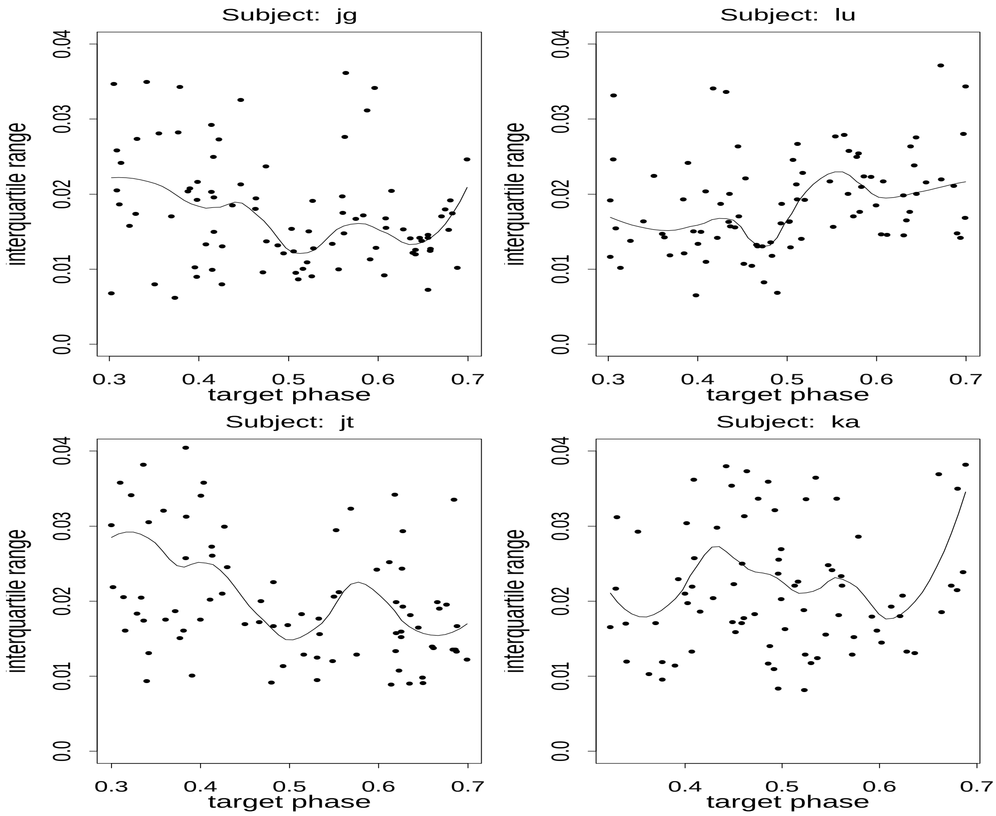

Figure 5: Experiment 1a. Interquartile range of accuracy within a trial as a function of target phase. For each subject, a locally weighted quadratic regression curve has been fitted using a weighted least squares fit on a neighborhood of 0.4 of the total number of data points.

phase. For each subject a smooth fit to the data has been derived using a locally weighted quadratic regression procedure. The method used is that of Chambers and Hastie (1992), where a fitted curve is built based on a least squares fit on a neighborhood (set here to the nearest 40% of the data points) of each predictor value.

For subjects JG, LU and JT, the variability is indeed at a local minimum at or around the attractors, with the largest effect visible for the attractor at 0.5. This closely resembles the pattern found in bimanual tasks where variance is lowest at relative phases of 0 and $\frac{1}{2}$. Subject KA's data again deviate from the pattern established by the other subjects, with apparent minima around 0.35 and 0.62.

3.3 Discussion of Experiment 1a

Experiment 1a managed to reveal a great deal of structure to the responses of subjects in this form of the speech cycling task. The experiment was designed with the goal of seeing just how rhythmically independent the metrical foot (taken here as the interval between two stress beats) can be from a higher level phrase repetition cycle which is imposed by the task. The data are quite striking. No subject comes even close to reproducing the uniform distribution from which target phases are drawn.

For three of the subjects, the evidence speaks strongly for the existence of three attractors. A precise numerical evaluation of these attractors is not possible, but each subject seems to have a strong fixed point near $\frac{1}{2}$, with one attractor on either side, at or near $\frac{1}{3}$ and $\frac{2}{3}$. A total of eight of the eleven regression lines fitted produce x-intercepts which are within 0.02 of these three fractional values. The interpretation of the three observed patterns as attractors is supported on several counts: the histograms demonstrate clearly that subjects are producing only a few discrete forms of behavior; the plot of accuracy as a function of target phase revealed that each of these stable behaviors can be associated with a particular target and that neighboring targets are biased towards these values; and finally, the relative stability of these patterns is emphasized by the drop in trial variance when the targets are close to the attractors.

A somewhat worrying aspect of these data is the fact that the one subject, KA, who differs greatly from the other three is the only male and non-musician in the group. While a very large gender difference seems highly unlikely, it is not at all implausible that musical experience plays a role in these results. The speech cycling task places rather unusual demands on the subject. These demands involve producing prescribed relative temporal onsets for events, and thus resemble musical tasks.

After they had completed three sessions, subjects were asked to informally report their impressions of the task and whether they thought they had accomplished the set goal of matching syllable to tone. All three musicians reported that they thought they had managed to match the syllable to the tone, even though their data show that this was not the case. Rather, they consistently produced three distinct patterns. Subject KA, on the other hand, reported having great difficulty with the task, and judged himself to have failed in the set task of matching syllables and tones. In order to balance our subject group for musical skills a further four subjects were recruited and a smaller version of the experiment was run.

4 Experiment 1b: effect of musical training

4.1 Method

The methods in this experiment were exactly as in the preceding, with the exception that a single session of 30 trials per subject was run. The four subjects recruited were two female non-musicians (AN and SP), one male non-musician (BS) and one male musician (JF). A male musician was included to check for a strong effect of gender. The subjects called "musicians" in this study report between 3 and 10 years of formal musical training, with between 8 and 29 years of experience playing a musical instrument, and all currently play at least one instrument and read written music fluently. None of the "non-musicians" currently play an instrument or are fluent readers. None has had any formal music training since high school, and all judge themselves to be non-musicians. A total of 4 musicians and 4 non-musicians took part in Experiments 1a and 1b.

4.2 Experiment 1b: results

As in the previous section, Figure 6 shows the overall distribution of median phases produced on each trial. Here, too, no subject comes close to reproducing the uniform distribution from which the targets were drawn. Subjects AN and SP (female, non-musicians) produce data much like the

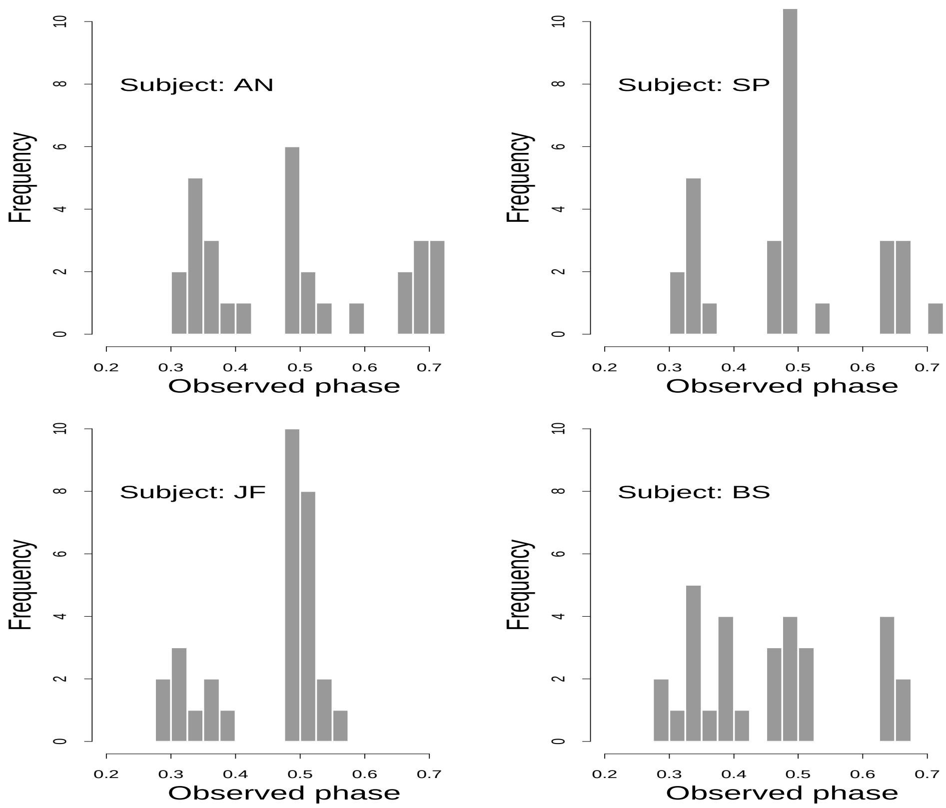

Figure 6: Experiment 1b. Distribution of $\hat{\mathbf{\Phi}}$ by subject. There are 30 trials per subject, and target phases come from a random uniform distribution between 0.3 and 0.7. Subjects AN, SP: female, non-musicians; Subject JF: male, musician; Subject BS: male, non-musician.

sub	x-int	r^2	n	p	x-int	r^2	n	p	x-int	r^2	n	p
AN	0.342	0.703	11	<0.01	0.501	0.941	9	<0.01	0.832	0.0571	9	>0.05
SP	0.324	0.629	8	<0.05	0.49	0.911	15	<0.001	0.656	0.573	7	<0.05
JF	0.229	0.446	9	<0.05	0.492	0.87	21	<0.001	—	—	—	—
BS	0.298	0.124	9	>0.05	0.407	0.996	5	<0.001	0.472	0.845	10	<0.001
BS	0.643	0.892	6	<0.01								

Table 3: Regressions in Experiment 1b. For each data set, clusters have been automatically estimated and a local regression line fitted. Listed here are values of the x-intercept, r^2, n and p for each regression.

three musicians of Experiment 1a. Subject JF (male, musician) produces a bimodal distribution, somewhat similar in appearance to that of KA of the last experiment. Finally, BS (male, non-musician) produces distribution in which two modes centered at about $\frac{1}{2}$ and $\frac{2}{3}$ are well defined, but the lower phases do not pattern as clearly.

Figure 7 plots $\hat{\mathbf{\Phi}} - \phi_{target}$ as a function of ϕ_{target}. Once more, clusters have been automatically

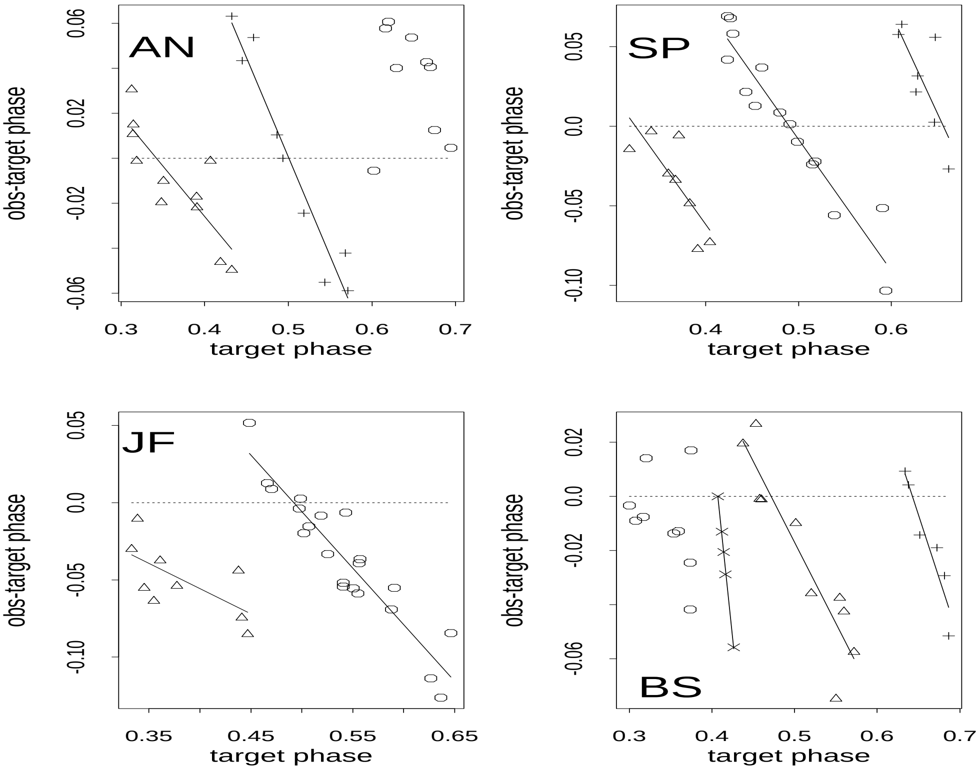

Figure 7: Experiment 1b. Distribution of $\hat{\mathbf{\Phi}} - \phi_{target}$ as a function of ϕ_{target}. Subjects AN, SP: female, non-musicians; Subject JF: Male, musician; Subject BS: Male, non-musician. As before, clusters have been estimated within each data set and a local regression line fitted. Only regressions for which $p < 0.05$ have been plotted.

identified for each subject and local regression lines fitted. The clustering algorithm used in the Experiment 1a was applied with the minor modification that the shape of the clusters was specified as more strongly ellipsoidal (ratio of major to minor axes was 5:1 in the first case and 20:1 in this case). This was necessary to preserve linear continuity with relatively sparse data (30 points per subject here, 90 in Experiment 1a). As before, the selection of the number of clusters was determined by the greatest sum of weighted r^2 given in Equation 1. Local regression lines have been added to Figure 7 where the p-value for the regression reaches significance at the 0.05 level.

From this figure it is clear that AN and SP are almost indistinguishable from subjects LU and JT in Experiment 1a, having attractors at values close to $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$. Subject AN has one outlier at around (0.61,-0.005) which has been classified as part of the rightmost cluster, and this point alone prevents the identification of a highly significant regression. Both AN and SP are non-musicians. Subject JF, by contrast, is a musician, yet his data pattern differently, in only having two clusters, with x-intercepts at 0.23 and 0.49. As with KA, there is no evidence for an attractor above 0.5. Finally, BS's data are interesting as he shows clear attractor structure at $\frac{1}{2}$ and $\frac{2}{3}$, but the cluster-based partitioning of the data is much less compelling for targets below about 0.45.

Table 3 provides values for the x-intercept for each local regression line. Again, all slopes are negative and this time 7 of the 10 regressions which reach significance produce intercepts within 0.02 of the three simple fractional values.

A. Median inaccuracy with and without stimulus				
Subject	mean difference	t	df	p
AN	-4.48e-3	-1.28	28	>0.05
SP	6.32e-3	1.09	29	>0.05
JF	4.98e-3	0.983	29	>0.05
BS	-4.82e-3	-1.13	29	>0.05
B. Interquartile range of inaccuracy with and without stimulus				
Subject	mean difference	t	df	p
AN	1.98e-4	0.10	28	>0.05
SP	**1.12e-2**	**3.00**	**29**	**<0.01**
JF	**4.00e-3**	**2.07**	**29**	**<0.05**
BS	4.90e-3	1.62	29	>0.05

Table 4: Experiment 1b. A: t-tests comparing median absolute inaccuracy within a trial for data collected with and without the stimulus. A total of 30 trials (in one case, 29) were run for each subject. All tests are two-tailed. B: t-tests comparing interquartile ranges for inaccuracy with and without stimulus.

Once more, t-tests were done to look for any difference in accuracy or variability between those repetitions obtained with the stimulus and those without. Table 4 lists the results. There were no significant differences in median inaccuracy score for data obtained with and without the stimulus. This replicates the finding that the presence of the stimulus does not increase the accuracy of subjects in performing this task. Two subjects were exhibited more variable performance in the presence of the stimulus.

4.3 Discussion of Experiment 1b

Experiment 1a was extended to permit equal sampling of two levels of musical skill. Our motivation was the surprising finding that one subject, KA, who was alone in being male and a non-musician, differed markedly from the three other subjects. The primary concern was that musicianship, or substantial formal musical training might greatly influence the results in a speech cycling task. These results demonstrate that non-musicians too may produce data which show clear evidence of a small number of attractors at phase values close to expected values.

This task elicits substantial inter-subject variability. The general rhythmic skill exhibited in speech cycling may be somewhat correlated with musical training, but may also be present in non-musicians (cf. AN and SP). A high degree of inter-subject variability is not very surprising given the nature of the task, which requires the alignment of speech and non-speech sounds at two independent points in the phrase. Previous researchers have reported subjects who are unable to perform simpler tasks which require alignment of non-speech stimuli (clicks, tones) with stresses in speech. Of 16 subjects required to judge as a "hit" or "miss" the alignment of a single click with a target syllable in a recorded phrase, Allen reported that only 3 were able to produce any coherent responses, one

sub	x-int	r^2	n	p	x-int	r^2	n	p	x-int	r^2	n	p
JT	0.31	0.393	8	>0.05	0.511	0.728	9	<0.005	0.655	0.916	11	<0.001
LU	0.3555	0.894	8	<0.001	0.522	0.841	10	<0.001	0.667	0.943	10	<0.001
SP	0.349	0.455	6	>0.05	0.548	0.841	22	<0.001	—	—	—	—
JF	0.378	0.583	6	>0.05	0.542	0.845	22	<0.001	—	—	—	—

Table 5: Regressions in Experiment 2. For each data set, clusters have been estimated and local regression lines fitted. Listed here are values of the x-intercept, r^2, n and p for each cluster.

of those 3 being the author himself (Allen, 1972). In tasks in which subjects attempted to align a series of alternating syllables so that a perceptually isochronous series resulted, both Seton and Scott reported that one subject failed to reach a satisfactory even rhythm (Seton, 1989; Scott, 1993).

5 Experiment 2: effect of speaking rate

One of the most important features of the model system studied by Kelso and colleagues is the qualitative change, or bifurcation, which the system undergoes as the rate of finger oscillation is increased. From being a system with two stable states, the system becomes monostable. The more complex case, where a target phase relation is environmentally specified and rate is simultaneously varied has not, to our knowledge, been systematically studied.

Maintaining the analogy between the studies of manual patterns and the speech cycling task, we can ask whether the attractor structure found in the above experiments depends on speaking rate. The location of the attractors and/or their number might conceivably change at faster rates. We thus undertook an initial, exploratory experiment at a faster rate.

5.1 Method

Methods employed were exactly as in Experiments 1a and 1b, with the following specific provisions. The fixed interval (interval a in Figure 2 above) was reduced from 700 ms to 450 ms. This yielded PRC periods ranging from 638 ms to 1500 ms. Four subjects were chosen, two from each part of Experiment 1a (JT, LU, SP and JF), giving two female musicians, one female non-musician and one male musician. A single experimental session was run, this time with 28 target phases evenly distributed between 0.3 and 0.705. Target orders were randomized within a session. The reason for this slight amendment to protocol was to ensure adequate sampling of all regions of the range 0.3–0.7 in as few trials as possible. Of the original stock of 30 phrases, 28 were randomly sampled for each session.

5.2 Experiment 2: results

Figure 8 shows the data clustering and local fits, computed exactly as in Experiment 1b, for all four subjects, while the x-intercepts and regression data are given in Table 5. It can immediately be seen that Subjects JT and LU exhibit almost no effect of rate. They each show the same attractor structure as in Experiment 1a. Subjects JF and SP each show evidence of two attractors only; just as for JF and KA in Experiment 1, there is no evidence for an attractor above 0.5.

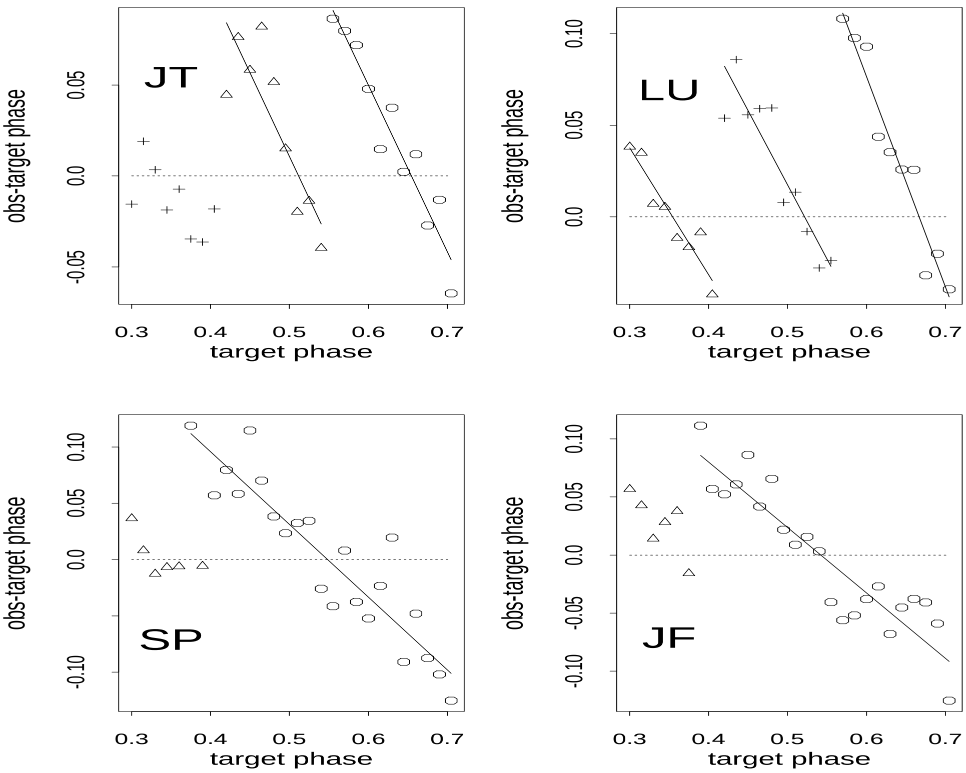

Figure 8: Experiment 2. Distribution of $\hat{\mathbf{\Phi}} - \phi_{target}$ as a function of ϕ_{target}. Subjects JT, LU: female, musicians; Subject SP: female, non-musician; Subject JF: male, musician. As before, clusters have been extracted and local regression lines fitted. Only regressions for which $p < 0.05$ have been plotted.

6 General discussion

The single most striking point about the data presented here is that subjects were never able to place the medial stress beat freely within the PRC. While there is a considerable degree of inter-subject variability, most of the subjects in this study showed strong preferences for 3 distinct phases. We can refer to these as the low, mid and high attractor. In all but one of the remaining cases, subjects exhibited a low and mid attractor, without a high counterpart.

The reader may demonstrate to herself the three distinct forms of organization documented here by reciting the phrase "beg for a dime" along with a series of evenly spaced beats, say finger taps. The low attractor is demonstrated by aligning "beg" and "dime" on successive beats and leaving the next (third) beat silent. The mid-attractor corresponds to the isochronous case: simply align "beg" and "dime" with alternating beats. The late attractor is demonstrated by introducing an artificial prominence on "for", and aligning each of "beg", "for" and "dime" with successive beats. The three patterns are illustrated using musical notation in Figure 9.

The evidence here is strongly supportive of the existence of attractors, i.e. privileged states of the system towards which the system will tend, and at which the system is most stable. The presence

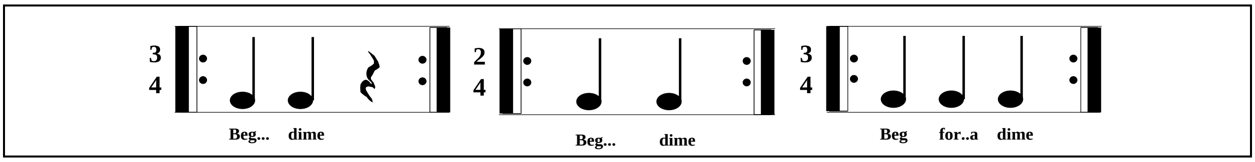

Figure 9: Musical notation illustrating the three stable patterns produced.

of attractors, in turn, entails the existence of a dynamic, which is readily seen to be both low-dimensional (our measurements are simple scalar phase values) and task-specific, as the PRC does not exist outside of the speech cycling task. The existence of a task-specific system which exhibits a low-dimensional intrinsic dynamic is evidence for a coordinative strategy in which relative phases are highly constrained, which is precisely the role we have suggested for rhythm in speech. In the remainder of this article, we will consider three themes of this argument: the case for the existence of a low-dimensional dynamic, the nature of the task-specific system and the role of rhythm in speech.

The presence of attractors and the existence of a low-dimensional dynamic. In Experiment 1a, we saw that subjects displayed only a few discrete behaviors in an experimental task where they were asked to place a stress beat at a point within a continuous range. They produced two or three phases reliably; intermediate targets yielded productions which were strongly biased towards the preferred phases. These preferred phases were often very close to values of $\frac{1}{3}$, $\frac{1}{2}$ or $\frac{2}{3}$, which divide the PRC into simple integral ratios—an effect we have called the Harmonic Timing Effect. The stability of these preferred phases was underscored by their reproduction at a much faster tempo (Experiment 2) and by the reduction in trial variance at the preferred values (Experiment 1a). Furthermore, although this form of the speech cycling task certainly requires a degree of skill, the effect is seen in both musicians and non-musicians (Experiment 1b). Together, these findings justify the interpretation of the preferred patterns as attractors of a dynamic system which is assembled by speakers in response to the demands of the experimental task.

In several cases, subjects did not show evidence of the attractor located at $\phi = \frac{2}{3}$. This attractor does appear to require a different prosodic structuring of the phrase than the other two (see Figure 9). For both the low and mid attractors, the phrase is divided into two feet, [*big for a*][*duck*]. The low attractor requires an additional silent, or implicit, foot, but does not require a restructuring of the phrase. The high attractor, in contrast, requires the division as [*big*][*for a*][*duck*], a completely different prosodic structure. The present data do not fully account for the observed distribution of this pattern. KA and JF did not produce this pattern at all, while SP produced it at the slower speaking rate, but not at the faster. For KA and JF the pattern may simply be a production strategy which they failed to discover; subjects were not given any feedback, nor were they told that they might produce a few discrete patterns. The case of SP suggests an additional hypothesis, which is that the number of feet within a phrase is constrained by speaking rate.

The nesting of integral numbers of feet within longer prosodic units could underlie the statistical tendency towards quantal timing observed by Fant and Kruckenberg (1996). In a close analysis of several text readings, they found a bias for inter-stress intervals which were multiples of a 500 ms unit. If there is a preferred interval of about this duration, prosodic reorganization may be a strategy employed by speakers to keep average foot length within some window of the preferred value. The inter-speaker variation we observed may thus partly reflect differences in preferred speaking rates; If KA and JF prefer slow rates, they may have been content to produce speech with a foot duration of 700 ms in Experiment 1. The possible interaction between prosodic structure and speech rate remains a rich and largely untapped research area.

The nature of the task-specific system. The current study is quite unlike previous attempts to apply the task dynamic approach to speech timing (Kelso et al., 1985; de Jong et al., 1993; Vatikiotis-Bateson and Kelso, 1993), which have typically studied the kinematics of individual articulators. While measurements of the position and velocity of, e.g., the jaw are relatively easy to obtain and provide a rich, continuous data stream, they are limited in scale because of the rate at which articulators cycle, which is typically once per syllable. They do not have periods co-extensive with the domains of English speech rhythm, that is, the metrical foot and larger units. In fact, no biomechanical part of the speech production system cycles at these timescales. For the study of rhythm at this level, then, no kinematic variable comparable to jaw position or velocity is available.

For this reason we have chosen to focus our attention on the acoustic signal. We defined the collective variable ϕ which is observable only once per cycle, and found that the distribution of ϕ is highly structured. The next question to be addressed is what kind of system could give rise to the observed data? While modeling work is still outstanding, an initial answer can be tentatively proposed.

If we consider the PRC and the metrical foot to each be non-arbitrary units in the production of speech under these conditions, then each can be said to have its own internal dynamic. As both are cyclic processes, the simplest dynamic we might propose is a second order oscillatory system. We have clearly demonstrated that the periods of the two processes are not independent, suggesting that the two systems are coupled. A further presumption is that the 'natural' period of the foot is smaller than that of the PRC. Under these very general conditions, almost any such system will rapidly settle into a state where the period of the more rapid oscillator (the metrical foot) is nested an integral number of times within the period of the slower (the PRC) (Thompson and Stewart, 1986; Glass and Mackey, 1988). Our collective variable ϕ can then be interpreted as a read out of the phase of the PRC at that point at which the metrical foot cycle has just restarted. This is entirely analogous to the mathematical technique of taking a Poincaré section of a system of two coupled oscillators. Further experimental work is required in order to tease out the nature of the two component oscillatory systems and the coupling function between them. An example of evidence pertinent to uncovering the nature of the underlying dynamic is the observation from Experiment 1a that for a range of target phases, subject JG produced two distinct patterns. This demonstrates that for some targets the system is potentially multistable. We have also previously demonstrated the presence of hysteresis in switching from one target to the other (Cummins and Port, 1996a).

The role of rhythm in speech. The speech cycling task introduced here stands somewhere on a continuum between reiterant speech and unconstrained speech; it does not eliminate as much phonetic detail as a reiterant speech task, but it contains a repetitive element. The simple act of phrase repetition appears to be sufficient to establish a tight temporal relationship between the phrase and its constituent metrical feet. Furthermore, the span of a metrical foot is itself partly determined by the rhythm established, as evidenced by the prosodic restructuring required for the high-attractor pattern.

In the presence of repetition the speech production system necessarily becomes coordinated, such that a higher level dynamic emerges within which the timing of subordinate processes are constrained. In this regard, speech rhythm is no different from the forms of coordination observed in other rhythmic activities (Ostry et al., 1983; Cummins and Port, 1996a). The reader may be familiar with the skill of patting one's tummy and rubbing one's head simultaneously. Once the skill has been acquired it may appear as if the two hands have become decoupled. If however one attempts to continuously alter the rate of one hand, the effect of coupling will rapidly be felt again. Collectively, the hands constitute a higher level system within which the timing of the individual component processes are highly constrained.

Rhythm in speech is a more complex phenomenon than simple isochronous models can represent. We have here demonstrated the hierarchical nesting of metrical levels arises inevitably within a repetition task, and what is more, the nature of the metrical levels depends to some extent on the resulting pattern. By structuring an utterance so that prominences (stresses, beats) lie at privileged phases of a higher-level prosodic unit, rhythm is seen as an organizational principle which has its roots in the coordination of complex action and its effect in the realm of prosodic structure.

7 Acknowledgments

Our work has benefited greatly from the thorough and insightful comments of Mary Beckman, an anonymous reviewer and the editors of this journal. Keiichi Tajima was a co-developer of the speech cycling methodology, and of many of our ideas on speech rhythm. We have also been helped by comments and discussion by Geoff Bingham, Ken de Jong, Gunnar Fant, Janet Pierrehumbert and Betty Tuller. The work was supported by a Research Incentive Dissertation Year Fellowship from Indiana University and a Summer Fellowship from the Cognitive Science Program at Indiana University, both to the first author.

References

Abercrombie, D. (1965). A phonetician's view of verse structure. In *Studies in Phonetics and Linguistics*, pages 16–25. Oxford University Press, London.

Abercrombie, D. (1967). *Elements of general phonetics*. Aldine Pub. Co., Chicago, IL.

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

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821.

Benguerel, A.-P. and D'Arcy, J. (1986). Time-warping and the perception of rhythm in speech. *Journal of Phonetics*, 14:231–246.

Bernstein, N. (1967). *The Coordination and Regulation of Movements*. Pergamon Press, London.

Bingham, G. P., Schmidt, R. C., Turvey, M. T., and Rosenblum, L. D. (1991). Task dynamics and resource dynamics in the assembly of a coordinated rhythmic activity. *Journal of Experimental Psychology: Human Perception and Performance*, 17(2):359–381.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*. Advanced Books and Software. Wadsworth and Brooks Cole, Pacific Grove, CA.

Cooper, W. E. and Eady, S. J. (1986). Metrical phonology in speech production. *Journal of Memory and Language*, 25:369–384.

Couper-Kuhlen, E. (1993). *English Speech Rhythm.* From the series *Pragmatics and Beyond.* John Benjamins, Philadelphia, PA.

Crystal, T. H. and House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, 88(1):101–112.

Cummins, F. (1997). *Rhythmic Coordination in English Speech: An Experimental Study.* PhD thesis, Indiana University, Bloomington, IN. Also Technical Report 198, Indiana University Cognitive Science Program.

Cummins, F. and Port, R. F. (1996a). Rhythmic commonalities between hand gestures and speech. In *Proceedings of the Eighteenth Meeting of the Cognitive Science Society*, pages 415–419. Lawrence Erlbaum Associates.

Cummins, F. and Port, R. F. (1996b). Rhythmic constraints on English stress timing. In Bunell, H. T. and Idsardi, W., editors, *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 2036–2039. Alfred duPont Institute, Wilmington, Delaware.

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

de Jong, K., Beckman, M. E., and Edwards, J. (1993). The interplay between prosodic structure and coarticulation. *Language and Speech*, 36(2,3):197–212.

Diedrich, F. J. and Warren, W. H. (1995). Why change gaits? Dynamics of the walk-run transition. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1):183–202.

Fant, G. and Kruckenberg, A. (1996). On the quantal nature of speech timing. In Bunell, H. T. and Idsardi, W., editors, *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 2044–2048, Wilmington, Delaware. Alfred duPont Institute.

Fowler, C. A. (1977). *Timing Control in Speech Production.* Indiana University Linguistics Club, Bloomington, IN. Reproduction of PhD Thesis, Dartmouth College.

Fraisse, P. (1956). *Les Structures Rhythmique.* Érasme, Paris.

Glass, L. and Mackey, M. C. (1988). *From Clocks to Chaos.* Princeton University Press, Princeton, NJ.

Hayes, B. (1985). *A Metrical Theory Of Stress Rules.* Garland Pub., New York, NY.

Jassem, W., Hill, D. R., and Witten, I. H. (1984). Isochrony in English speech: its statistical validity and linguistic relevance. In Gibbon, D. and Richter, H., editors, *Intonation, Accent and Rhythm*, volume 8 of *Research in Text Theory*, pages 203–225. Walter de Gruyter, Berlin.

Jones, D. (1918). *An Outline of English Phonetics.* W. Heffner, Cambridge, 1st edition.

Kay, B., Saltzman, E., and Kelso, J. A. S. (1991). Steady-state and perturbed rhythmical movements: Dynamical modeling using a variety of analytical tools. *Journal of Experimental Psychology: Human Perception and Performance*, 17:183–197.

Kelso, J. A. S. (1995). *Dynamic Patterns.* MIT Press, Cambridge, MA.

Kelso, J. A. S., Holt, K. G., Kugler, P. N., and Turvey, M. (1980). On the concept of coordinative structures as dissipative structures: II. Empirical lines of convergence. In Stelmach, G. and Requin, J., editors, *Tutorials in Motor Behavior.* North-Holland.

Kelso, J. A. S. and Jeka, J. J. (1992). Symmetry breaking dynamics of human multilimb coordination. *Journal of Experimental Psychology: Human Perception and Performance*, 18(3):645–668.

Kelso, J. A. S. and Kay, B. A. (1987). Information and control: A macroscopic analysis of perception-action coupling. In Heuer, H. and Sanders, A. F., editors, *Perspectives on Perception and Action*, chapter 1, pages 3–32. Lawrence Erlbaum Associates, Hillsdale, NJ.

Kelso, J. A. S., Saltzman, E., and Tuller, B. (1986). The dynamical perspective in speech production: Data and theory. *Journal of Phonetics*, 14:29–60.

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

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics and dynamic modeling. *Journal of the Acoustical Society of America*, 77(1):266–280.

Kidd, G., Boltz, M., and Jones, M. R. (1984). Some effects of rhythmic context on melody recognition. *American Journal of Psychology*, 97(2):153–173.

Kugler, P. N., Kelso, J. A. S., and Turvey, M. T. (1980). On the concept of coordinative structures as dissipative structures: I. Theoretical lines of convergence. In Stelmach, G. and Requin, J., editors, *Tutorials in Motor Behavior*. North-Holland.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5:253–263.

Liberman, M. (1975). *The Intonational System of English*. PhD thesis, M.I.T., Cambridge, MA. Published by Indiana University Linguistics Club, 1978.

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

Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics*, 30:247–256.

Meijer, O. G. and Roth, K. (1988). *Complex Movement Behavior: The Motor-Action Controversy*. Elsevier, Amsterdam.

Morton, J., Martin, S. M., and Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, 83:405–408.

O'Connor, J. D. (1965). The perception of time intervals. Technical Report Progress Report 2, Phonetics Laboratory, University College, London.

Ostry, D. J., Keller, E., and Parush, A. (1983). Similarities in the control of the speech articulators and the limbs: Kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology: Human Perception and Performance*, 9(4):622–636.

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

Port, R. F., Cummins, F., and Gasser, M. (1995). A dynamic approach to rhythm in language: Toward a temporal phonology. Technical Report 150, Indiana University Cognitive Science Program, Bloomington, IN.

Saltzman, E. and Kelso, J. A. S. (1987). Skilled actiions: A task dynamic approach. *Psychological Review*, 94:84–106.

Scholz, J. P. and Kelso, J. A. S. (1990). Intentional switching between patterns of bimanual coordination depends on the intrinsic dynamics of the patterns. *Journal of Motor Behavior*, 22(1):98–124.

Scholz, J. P., Kelso, J. A. S., and Schöner, G. (1987). Nonequilibrium phase transitions in coordinated biological motion: Critical slowing down and switching time. *Physics Letters A*, 123(8):390–394.

Schöner, G. S. and Kelso, J. A. S. (1988a). A synergetic theory of environmentally-specified and learned patterns of movement coordination. I. Relative phase dynamics. *Biological Cybernetics*, 58:71–80.

Schöner, G. S. and Kelso, J. A. S. (1988b). A synergetic theory of environmentally-specified and learned patterns of movement coordination. II. Component oscillator dynamics. *Biological Cybernetics*, 58:81–89.

Scott, D. R., Isard, S. D., and de Boysson-Bardies, B. (1985). Perceptual isochrony in English and French. *Journal of Phonetics*, 13:155–162.

Scott, S. K. (1993). *P-centers in Speech: An Acoustic Analysis.* PhD thesis, University College London.

Seton, J. C. (1989). *A Psychophysical Investigation of Auditory Rhythmic Beat Perception.* PhD thesis, University of York.

Shen, Y. and Peterson, G. G. (1962). Isochronism in English. In *Studies in Linguistics, Occasional Papers 9*, pages 1–36. University of Buffalo.

Thompson, J. M. T. and Stewart, H. B. (1986). *Nonlinear Dynamics and Chaos.* John Wiley and Sons, New York, NY.

Tuller, B. and Kelso, J. A. S. (1989). Environmentally-specified patterns of movement coordination in normal and split-brain subjects. *Experimental Brain Research*, 75:306–316.

Vatikiotis-Bateson, E. and Kelso, J. A. S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *Journal of Phonetics*, 21:231–265.

Woodrow, H. (1951). Time perception. In Stevens, S. S., editor, *Handbook of Experimental Psychology.* Wiley, New York, NY.

Yamanishi, J., Kawato, M., and Suzuki, R. (1980). Two coupled oscillators as a model for the coordinated finger tapping by both hands. *Biological Cybernetics*, 37:219–225.

Appendix A: beat measurement

The present work does not attempt to settle long outstanding questions about the exact location of a beat within a stressed syllable (Allen, 1972; Allen, 1975; Morton et al., 1976; Scott, 1993). It is well known that most subjects can reliably perform a range of tasks which seem to depend on the perception of an instantaneous beat, and while this beat is highly correlated with the onset of the syllable nucleus, it is displaced somewhat as a function of both the initial consonant(s) and the length of the rime (Allen, 1972; Morton et al., 1976; Fowler, 1977; Scott, 1993). In keeping with the work of Allen (1972) we refer to these events as stress beats, or simply beats, but we do not differentiate them from the P-centers sought by other researchers. The deviation of the perceived beat from the onset of a nuclear vowel is small compared to length of the average inter-stress interval, and so our results should be robust with respect to the details of the procedure outlined below.

The extraction of stress beats is done largely automatically, using an algorithm based on the work of Scott (1993). No claim is made that the beats extracted here are more accurate than those computed by any other algorithm. However, the algorithm presented here preserves most of the merits of existing procedures, and produces beats close to, but slightly after, the onset of syllabic voicing.

Speech is recorded digitally at 11025 Hz. The signal is bandpass filtered using a first order Butterworth filter centered at 1000 Hz and having a bandwidth of 600 Hz. Because of the shallow skirts of this filter, the net effect is to largely eliminate fricative noise and F0 energy, leaving energy in the formant regions intact. Informal trials with a range of filters indicated that any procedure which achieved these dual goals would suffice. The resulting signal is rectified (using absolute values) and smoothed heavily, usually using another first order Butterworth filter, this time as a lowpass filter with a very low cut off of about 10 Hz. The result of this stage is a smooth amplitude envelope. A beat is associated with every local rise in this envelope, and is defined as occurring at the point in time midway between the points where the local rise is 10% and 90% complete. This is similar to the heuristic used in Scott (1993), and serves to remove the effect of very gradual on- and offsets. Figure 1 illustrates beat placement for a sample phrase. The resulting beats are marked graphically in a display of the amplitude envelope, and are then checked visually for reasonableness and completeness. Both criteria are necessary as spurious beats may be detected, e.g. halfway through a diphthong, and beats may be missed, as when a syllable is partly or wholly devoiced. In the latter case, measurements are made by hand from the amplitude envelope, with the acoustic signal as a guide. No attempt is currently made to assign a relative strength to a beat, although this is, in principle, possible. Code is available on request from the first author (`f-cummins@nwu.edu`).