UC Berkeley
UC Berkeley PhonLab Annual Report

Title
Syllable Timing in Repetition Disfluency

Permalink
https://escholarship.org/uc/item/1xb4v4qj

Journal
UC Berkeley PhonLab Annual Report, 2(2)

ISSN
2768-5047

Author
Tilsen, Sam

Publication Date
2006

DOI
10.5070/P71xb4v4qj

eScholarship.org Powered by the California Digital Library
University of California

# Syllable Timing in Repetition Disfluency

Sam Tilsen
University of California, Berkeley

**Abstract.** This paper investigates rhythmic patterns in 3-cycle repetition disfluency (e.g. *We went to the park and$_1$ and$_2$ and$_3$ watched the birds*...). Results show a harmonic timing effect: in a slow-speech set of corpus data, the distribution of *and$_2$* phases relative to the *and$_1$* to *and$_3$* interval is trimodal ($\chi^2$ = 20.1, $p < .005$), with modes approximating low-order harmonic ratios (1/3, 1/2, 2/3). A task-dynamic coupled-oscillators model can represent the structure of the observed distribution as attractor-structure in a potential function, accounting for why harmonic modes are observed. Furthermore, the model can treat the ratio of oscillator amplitudes as a control parameter representing speech-rate; as this parameter is varied, higher-order modes become unstable, explaining why trimodality is not observed in a fast-speech subset of the data. In contrast to previous experimental findings of harmonic timing, the effects observed here operate on a smaller time scale and cannot be attributed to an external stimulus or self-entrainment.

## 1. Introduction

This investigation presents evidence for rhythmic coordination of syllables in 3-cycle repetition disfluency (e.g. ...*I went to the store and$_1$ and$_2$ and$_3$ I bought some milk*...). Though there are many issues regarding when and why speakers produce such repetitions, we will focus on a somewhat different matter—the timing (or, *phase*) of the second syllable (*and$_2$*) relative to the first (*and$_1$*) and third (*and$_3$*). We will see that in slow-speech conditions, the distribution of *and$_2$* phase is trimodal. This pattern can be modeled as emerging from a dynamical system of coupled oscillators. To understand these arguments, we should first consider what is meant by "rhythmic coordination".

### *1.1. Background on rhythmic coordination*

"Coordination" in this context implies the ordering or positioning of events into specific temporal patterns. We can conceptualize the process whereby events are coordinated in two distinct ways: either as involving *control*, in which an agent manipulates the relations between events, or as *self-organizing*, in which the patterns emerge from the dynamics of the system itself.

What does it mean to say that some pattern is "rhythmic"? The literature on rhythm in speech contains several different approaches, each emphasizing different aspects of the concept of RHYTHM, listed below:

- *regular recurrence of events*:
    - isochronous intervals between events (*strict isochrony*)
    - intervals between events perceived as regular (*perceptual isochrony*)

- *prominence inequality*: differences in relative salience between events (e.g. όσ, σό )

- *hierarchic structure*: grouping of events into larger, more complex events
    - grouping by spatial geometry
    - grouping by temporal geometry

Regular recurrence of events in time is the most central aspect of RHYTHM. The other two aspects, prominence inequality and hierarchic structure, depend upon some form of recurrence, but not vice versa. To see this, consider that a beat does not necessarily involve complexes of events or the alternation of more and less prominent events. In many languages spontaneous speech-events can be *perceived* to re-occur regularly, but equal temporal intervals from one event to the next are rarely observed.

Prominence inequality is the basis for a class of phonological formalisms involving the representation of prominence asymmetries with a metrical grid. In these schemes, the prominence of stress/accent-bearing units—usually syllables—is represented by the presence or absence of prominence marks on successively higher grid levels (e.g. Liberman 1975, Liberman & Prince 1977). A number of researchers do not focus on prominence inequality as the most important aspect of rhythm; some rhythmic typologies aim to characterize languages according to their isochronous units (e.g. stress-timing and syllable-timing (Pike 1945), and mora-timing (Port, Dalby, & O'Dell 1987)), rather than the relative prominence between units.

Hierarchic structure, in the traditional phonological sense, involves the abstract grouping of sub-syllabic events (segments, moras) into syllables, and the grouping of syllabic events into higher units: feet and phrases. In metrical grid approaches, hierarchic constituency is represented by placing brackets around groups of prominence marks on syllable and foot grid levels; in metrical tree approaches, branching-tree structures represent hierarchic constituency. All of these approaches use spatial geometry to directly represent hierarchical relations.

A different conception of hierarchic rhythmic structure has emerged in dynamic systems (or, *task-dynamic*) approaches to speech rhythm and the perception of rhythm (Saltzman & Kelso 1987; Cummins & Port 1998; Large & Jones 1999). In these views, the grouping of speech events entails temporal coordination of those events. In other words, hierarchic structure implies specific timing patterns of the grouped subunits. This allows for predictions about when certain events should tend to occur in relation to other events—predictions which are absent in abstract phonological conceptions of hierarchic structure.

Such approaches are called "task-dynamic" because they posit that the system responsible for coordination can be usefully described in terms of the dynamics of a "task variable". The task variable is often the phase of some target movement or event with respect to some other movement or event. The temporal dynamics of this variable are visualized with the aid of a potential field; in this way a spatial geometry is used to indirectly represent temporal geometry. In our case, the primary task-dynamic variable will be the phase of $and_2$, the second syllable in a threepeat. We will see that the dynamics of this variable are analogous to those observed in investigations of phrase and foot timing conducted by Cummins and Port (1996, 1998), in which an interesting phenomenon known as the "harmonic timing effect" is observed.

*1.2 Harmonic timing*

Using a *speech cycling task*, Cummins and Port (1996, 1998) found evidence for constraints on the timing of stressed syllables in English. In the 1998 version of their experiment[1], subjects were played a high-low two-tone metronome pattern (H-L-H-L...), in which the interval between the two tones was fixed at 700ms; the base period (H-L-H) was varied from 1000 ms - 2330 ms, hence varying the target phase of the L tone from 0.30 to 0.70 (see Fig. 1 for illustration). The phase in this case is the ratio of the duration of the interval between the H and L tones to the duration of the base interval between H tones. Subjects were instructed to repeat a two-stress phrase (always of the form: *X for a Y* , e.g. *big for a duck*) so that the first stressed word is aligned with the H tone and the second stressed word is aligned with the L tone. After 12 repetitions, the two-tone metronome pattern was stopped, and subjects continued repeating the phrase, trying to maintain the stimulus rhythm.

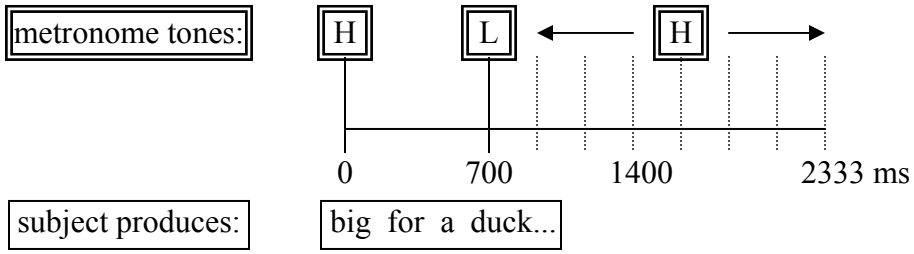


Fig. 1. S*peech cycling task from Cummins and Port (1998)*. The interval between a high and low tone (H and L) was fixed at 700 ms, while the base period of repetition was varied randomly and uniformly such that the L tone occurred from phase 0.30 to 0.70. Subjects were instructed to produce the stressed syllables of the phrase in alignment with the metronome tones.

With this experimental procedure, Cummins and Port discovered a *harmonic timing effect:* subjects were biased to produce the second stressed syllable of the phrase (actually, the P-center of this syllable; cf. section 2.4) at phases close to integer ratios: 1/3, 1/2, 2/3—i.e. 0.33, 0.50, 0.66. In other words, despite the fact that the distribution of target phases was uniform from 0.30 to 0.70, subjects were biased toward producing low-order integer-ratio phases even when the target phase was less harmonic, e.g. 0.40 or 0.60. The phases 0.33, 0.50, and 0.66 are thus special; the timing of the second stressed-syllable in a repeated two-stress phrase is attracted to these "harmonic" phases.

Interestingly, there was inter-subject variability in whether two or three of the special phases were utilized. For some subjects, the distribution of produced phases only had modes at the 1/3 and 1/2 phases, other subjects had modes only at the 1/2 and 2/3 phases, and still others exhibited all three. Most subjects showed no difference in accuracy (deviation from the target) between the with- and without-tone production conditions. Another important finding was that when the H-L interval was shortened to 400ms, some subjects no longer used the higher-order modes (1/3 and 2/3).

*1.3 Task-dynamic model of the harmonic timing effect*

Why are subjects in the speech cycling task biased to produce harmonic phases, rather than matching equally well all target phases? Cummins and Port, following synergetic and task-dynamic approaches to motor coordination (Haken, Kelso, & Bunz 1985; Saltzman & Kelso 1987), argue that the tendency to produce integer phase ratios (1/3, 1/2, 2/3) can be modeled with a dynamical system of two coupled oscillators. Such a model can describe behavioral patterns which result from periodicity on multiple scales.

As described in Port (2003), the oscillators produce pulses cyclically, and the pulses act as attractors for syllable beats. In the speech-cycling, one oscillator corresponds to the lower-frequency phrase repetition cycle (a "phrase oscillator") and the other corresponds to the higher-frequency foot repetition cycle (a "foot oscillator"—here a foot is treated as the interval between stress beats).

The state of this system can be described by one task variable, relative phase φ, which is the difference between the phases of the oscillators. Stable relative phases can be represented as valleys in a potential function. Eqs. (1a) and (1b) are the potential functions for systems of oscillators with 1:2 and 1:3 frequency coupling. Fig. 2 depicts graphs of these functions. The derivative of the potential with respect to phase is the negative of the first derivative of phase φ (Eq. (2); cf. Haken 1983; Haken, Kelso, & Bunz 1985; Kelso 1995 for further details).

(1a) $V(\varphi) = -\cos \varphi - \cos 2\varphi$ (1b) $V(\varphi) = -\cos \varphi - \cos 3\varphi$

(2) $d\varphi/dt = - dV/d\varphi$

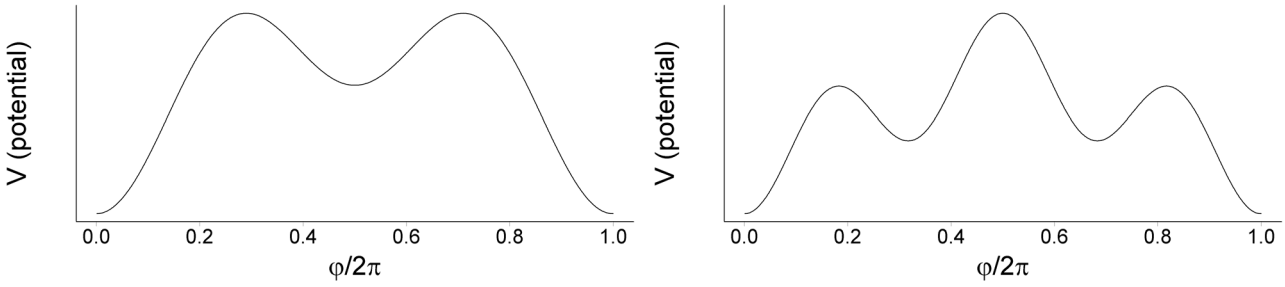


Fig. 2. *Potential functions for 1:2 and 1:3 frequency-locked oscillators*. Left: potential function for 1:2 coupling. Right: potential function for 1:3 coupling. Valleys correspond to attractors, i.e. stable modes of coordination. Note that axes map onto circles, so that phase 0 = phase 1.

This model was inspired by one Haken, Kelso, and Bunz (1985) developed to describe bimanual coordination (known as the HKB model). The rhythmic patterns described by the HKB model came from experiments in which subjects were instructed to repeatedly flex their fingers or wrists either simultaneously (in-phase) or in syncopation (anti-phase). The reader may wish to replicate one of the main findings right now. Begin by syncopating finger flexions of the left and right hands at a comfortable rate. Slowly increase the rate of repetition. Continue increasing the

rate until no longer able to achieve syncopation. The reader should have observed that around the frequency where syncopation becomes difficult, relative phasing of the fingers becomes more variable, and beyond that frequency, only the in-phase mode of coordination is stable.

This phase transition from two stable modes of coordination to just one stable mode is nicely described by a coupled-oscillators model. We need only to add amplitude parameters *a* and *b* to the potential function for a 1:2 coupled oscillator system (Eq. 3) to model this phase transition. We can then treat the ratio *b*/*a* as a control parameter that explicitly represents the amplitude of the 2nd harmonic and by hypothesis represents the effects of rate of repetition. Fig. 3 shows how the attractor structure changes as the control parameter is varied. As *b/a* is decreased from 1, the 0.5 phase mode becomes less stable, and beyond *b/a* = 0.25, becomes unstable (cf. HKB (1985) for further details).

(3) $V(\varphi) = - a \cos \varphi - b \cos 2\varphi$

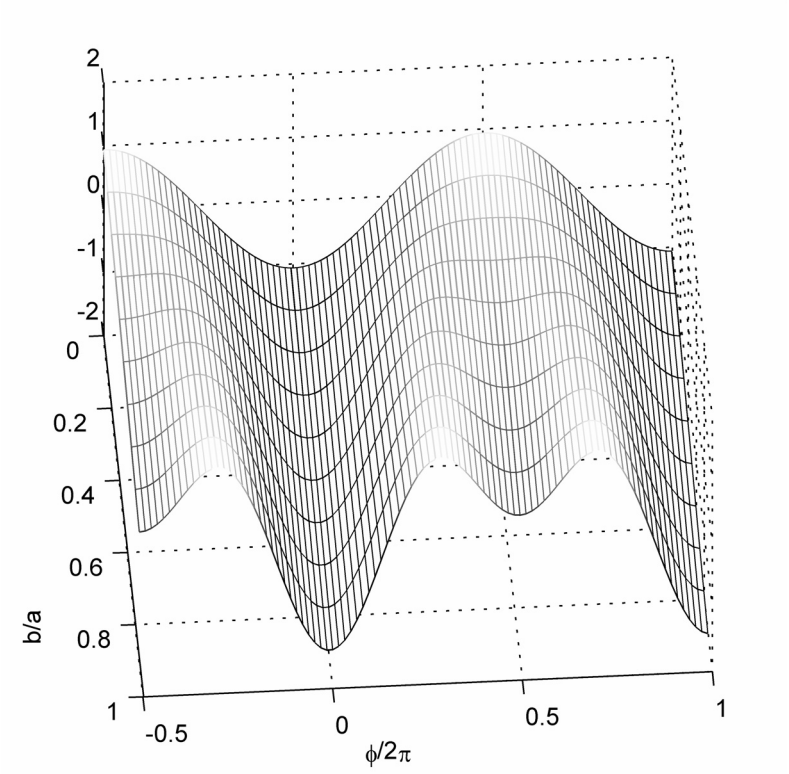


Fig. 3. *Hypothesized effect of speech-rate on the stability of anti-phase coordination*. As the control parameter *b/a* is decreased, the anti-phase attractor in the potential function becomes less stable. For *b/a* < 0.25, the attractor is unstable.

Let us return for a moment to the speech-cycling results. Cummins and Port posited that each attractor corresponds to a distinct metrical parsing of the phrase. The 0.50 attractor corresponds to a two-foot metrical parsing: [big for a] [duck]; the 0.66 attractor corresponds to the three-foot parsing [big] [for a] [duck]; and the 0.33 attractor corresponds to a the three-foot metrical parsing: [big for a] [duck] [.], in which the third foot is a "silent, or implicit, foot".

The key idea behind this analysis is to associate concentrations of density in the distribution of observed phase relations (the "rhythmic patterns") with attractors in a dynamical system model (the "rhythmic coordination"), and furthermore to associate the attractors with metrical "structures". The existence of these rhythmic patterns argues for a coordinative system that is biased to produce them. This approach thus reconceptualizes hierarchical metrical structure as the dynamics of rhythmic coordination system.

Another issue regards how are phrase and foot oscillators are instantiated in the cognitive system. Do they arise from the motor system, or from the perceptual/attentional system? Large and Jones (1999) argue for the perceptual view, proposing a theory of attentional dynamics in which rhythmic structures arise from internal oscillations associated with sensory and attentional systems. These internal oscillations produce pulses, which in turn generate temporal expectancies; hence rhythmic motor production follows indirectly from rhythmic perception: the motor system serves the expectancies by aligning the most salient part of some event with the pulses of an internal oscillator.

One alternative possibility is that such oscillators are instantiated in both perceptual and motor systems. Whether harmonic timing effects can be observed in conditions where no perceptual entrainment has occurred may bear upon these issues. Another possibility is that the expectancies or biases are continuous, rather than discrete or pulse-like events.

Although the phrase-foot harmonic timing effect is robust in experimental settings, the effect has not been observed in spontaneous speech. Outside of the lab, people rarely repeat a word, foot, or phrase once, much less several times. This and other factors such as variable syllable structure, lexical and phrasal accents, and semantic and pragmatic influences confound detection of the effect between successive phrases in spontaneous speech. There is, however, one situation in which word repetition is not extremely rare: repetition disfluency.

### *1.4 Background on repetition disfluency*

Repetition disfluency is a deviation from normal speech patterns (or, a "speech error") in which a word or short phrase is repeated without an emphatic intent (1). Various researchers have interpreted the cause of repetition in different ways. Maclay & Osgood (1959) grouped repetition with "hesitation phenomena," which included filled pauses, unfilled pauses, and lengthening (2-4), patterns which frequently co-occur (5). Hieke (1981) argued that hesitation phenomena are strategic devices to allow for more planning time. Howell and Au-Yeung (2002) treat repetition as a "fluency failure" that occurs when speech planning does not keep pace with speech execution.

| | |
|---|---|
| (1) *I I* expected something else *and and* that was... | (repetition) |
| (2) I *uh* expected something else and *uh* that was... | (filled pause) |
| (3) I .. expected something else and .. that was... | (unfilled pause) |
| (4) [*aiiii*] expected something else [*ææennd*] that was... | (lengthening) |
| (5) [*aiiiii*] uh .. I I expected something else and uh and... | (combination) |
| | |
| (6) I .. the expectation that something else will happen.... | (false start/deletion) |
| (7) I .. we expected something else. | (substitution) |
| (8) I expected something else .. to hear something else. | (insertion) |

Other research on disfluency has been inspired by feedback control-system models of the speech planning system. Levelt (1983) distinguished between *covert repairs* and *overt repairs*, based upon whether there is an indication in the utterance of what was problematic in the original plan. In these terms, (1)-(5) exemplify covert repairs, while (6)-(8) are classified as overt repairs. The motivation for this distinction arose from a model of speech planning in which an internal monitor detects problems with a speech plan and triggers repairing. According to the *main interruption rule* (Levelt 1983), the flow of fluent speech stops immediately upon detection of a problem. False starts (6) occur frequently, but substitution (7) and insertion (8) are much less prevalent in spontaneous speech (Shriberg 2001).

*1.4.1. Factors in the distribution of hesitation disfluency*

There are both structural and non-structural factors in the distribution of repetition and other hesitation disfluencies. Non-structural factors include discourse context, speaker-style, and speech-rate. Shriberg (2001), comparing corpora of human-to-human speech and human-to-computer speech, found that repetitions, filled pauses, and deletions occur primarily in human-to-human speech; this indicates that these phenomena serve some purpose in the speaker-hearer interaction. Shriberg also found speaker-specific variation in the relative frequency of repetition vs. deletion: speakers can be classified as "repeaters" or "deleters," according to which type of speech error they produce more often. However, this speaker-specific factor has not been conclusively separated from speech-rate: Shriberg also found a correlation between faster speech and a higher proportion of deletion, and between slower speech and relatively more repetition.

Structural factors that influence the distribution of hesitation include syntactic position, syntactic complexity, phonological complexity, and lexeme frequency. Sentence- and turn-initial hesitation is in general more common than sentence or turn-medial hesitation (Maclay & Osgood 1959), and repetition occurs more often preceding constituents of higher complexity (Clark & Wasow 1998). More frequent words are more likely to be repeated. Function words are repeated more frequently than content words (Maclay & Osgood 1959), and this cannot be attributed to their syntactic position or frequency (Clark & Wasow 1998). Hesitation is also more likely before constituents of higher phonological complexity.

*1.4.2. Phonetic correlates of disfluency.*

The phonetic effects of disfluency are most evident near the interruption point (Shriberg 2001), which is the point in time when fluent speech is interrupted. Some of the manifestations of disfluency reported by Shriberg are listed below:

- *Lengthening* of rhymes or syllables that precede and follow the interruption point. Lengthening can occur during or before a repeated form, and can sometimes be the only surface manifestation of disfluency. The pitch contour associated with lengthening due to disfluency is normally flat or slowly falling (2001: 161).
- *Creaky-voice or laryngealization, word cutoffs, and diplophonia* are language-specific or speaker-specific phonetic patterns occurring just before the interruption point.
- *Abnormal coarticulatory patterns* are sometimes observed, such that the form preceding interruption anticipatorily coarticulates (from most to least common): 1) nothing, 2) a fluent continuation, 3) a planned but not produced word, or 4) its own repetition. Shriberg

notes that in a small number of cases "speakers must be planning to repeat while they are still producing the first instance of the [repeated] word" (2001: 163).

- *Unfilled pauses* may or may not be present.

### *1.4.3. Causes of repetition*

Perspectives on the causes of repetition can be divided into two classes: *discourse-functional* perspectives, which posit functional or pragmatic explanations for repetition; and *planning-systemic* perspectives, which resort to models of speech-planning for explanatory factors. For the most part, these perspectives are complementary.

Discourse-functional perspectives involve the speaker-hearer interaction in some way. For example, Hieke (1981) viewed repetition as a stalling strategy for the speaker to buy more planning time, and Levelt (1983) viewed repetition as a signal to the hearer that the speaker needs more planning time. Clark and Wasow (1998) outlined the *continuity hypothesis*: "all other things being equal, speakers prefer to produce constituents with a continuous delivery" (206), either because 1) speakers are considering the processing tasks of their addressees, or 2) speakers want to make an impression of fluency upon their addressees, for social reasons.

Planning-systemic perspectives attribute the cause of disfluency to some aspect of the cognitive speech planning system. In the Levelt (1983) model, the planning system involves internal planning, monitoring, and repairing processes, which are conceptualized as a feedback control-system. Occasionally errors arise in the planning process, and disfluencies result from the need to repair those errors. For example, Kolk and Postma (1993) formulated their *covert repair hypothesis* within a speech planning model in which speakers are biased to produce major constituents because of the planning system itself, rather than pragmatic factors. Similarly, Clark and Wasow (1998) argued for a major-constituent planning bias inherent in the planning system, and also found evidence for a *complexity hypothesis*, which holds that hesitation is more frequent before more complex constituents, because constituent complexity presumably induces a greater proportion of planning errors.

A different brand of systemic perspectives are *activation-based* accounts, which attribute repetition to lingering activation of a preceding form. One such model is EXPLAN (Howell & Au-Yeung 2002). This model treats planning (PLAN) and execution (EX) of speech as independent processes occurring in parallel, and considers all "fluency failures"—including stuttering, abnormal lengthening, deletions, substitutions, filled and unfilled pauses, and repetition—as failures of the planning process to keep pace with the execution process. In particular, repetition and pausing occur when planning lags behind execution, in which case the speaker can retrieve the plan of a word recently used (the activation of which is lingering) and execute it again (Howell & Au-Yeung 2002).

## *1.5. Subclassification of 2-cycle repetitions*

Most (and maybe all) previous work on repetition disfluency has focused on 2-cycle repetition (i.e. *...and$_1$ and$_2$...*).[2] Phonetic and distributional properties of both the 1st and 2nd cycles have been used to motivate discourse-functional and planning-systemic subclassifications of repetitions. We will later consider whether these classifications extend to the 2nd cycle in 3-cycle repetitions, and whether they can account for the observed rhythmic patterns.

### *1.5.1. 1st cycle-based classification: hesitation-anticipating vs. hesitation-non-anticipating*

One way of classifying 2-cycle repetitions is based upon characteristics of the 1st cycle. Clark and Wasow (1998) distinguished between "states" of the speaker during the utterance of the 1st cycle: either the speaker has anticipated the interruption, or has not. In this view there are thus two distinct types of 1st cycles: those produced with the expectation of upcoming hesitation, and those produced without the expectation of hesitation. Essentially the distinction is about when—relative to the production of the 1st cycle—the speaker anticipates planning trouble.

Clark and Wasow observed that hesitation-anticipating[3] 1st cycles usually have longer durations and fuller vowel qualities than hesitation-non-anticipating 1st cycles. With repetitions of the lexeme "the", for instance, the form [*θi*] occurs more frequently in hesitation-anticipating 1st cycles than the otherwise more common, reduced form [*θə*]; [*θi*] is also repeated more frequently than [*θə*] in these cases (1998: 227). Lengthening and pre-pausing are typical of a hesitation-anticipating 1st-cycle, while post-pausing/filling and the absence of lengthening are typical of a hesitation-non-anticipating 1st-cycle.

### *1.5.2. 2nd cycle-based classification: stalling vs. retracing.*

Another way of classifying repetitions is based upon the discourse function of the 2nd cycle. Hieke (1981) identified two functionally distinct types of repetition. He posited a class of *prospective repeats*, which anticipate continued hesitation and "stall" to give the speaker planning time; and also a class of *retrospective repeats*, which function to reestablish fluency after pauses and other hesitations by "retracing" to the preceding constituent. Retraces function to create cohesion between the units of interrupted constituents. Retraces tend to be of normal duration and are not followed by pauses, while stalls are usually of abnormally long duration and may be followed by a filled or unfilled pause.

### *1.5.3. Combination of classes 1st and 2nd cycle-based classifications.*

Given the distinction between hesitation-anticipating and hesitation-non-anticipating 1st cycles and the distinction between stalling and retracing 2nd cycles, there are four logical possibilities for subclassification of a 2-cycle repetition, shown in Table 1. These subclassifications make use of both a discourse-functional perspective (the pragmatic intention of stalling vs. retracing in the 2nd cycle) and a systemic perspective (the time course of hesitation, hesitation-detection, and utterance of the 1st cycle in speech-planning/production).

| | |
|---|---|
| **CLASS 1**<br>*anticipating*<br>*& retracing* | **CLASS 2**<br>*anticipating*<br>*& stalling* |
| **CLASS 3**<br>*non-anticipating*<br>*& retracing* | **CLASS 4**<br>*non-anticipating*<br>*& stalling* |

Table 1. *1st and 2nd cycle-based subclassification of 2-cycle repetition*. Each class represents one of the four logically possible combinations of classes based upon the 1st and 2nd cycle distinctions described above.

Duration and pause profiles of these classes are schematized in Fig. 4. The contrast between classes 1, 2 vs. 3, 4 exemplifies the difference between prototypical hesitation-anticipating and hesitation-non-anticipating 1st cycles: the former has a lengthened 1st cycle, with a possible pause before that cycle; the latter exhibits normal duration of the 1st cycle, with no pause preceding. The contrast between classes 1,3 vs. 2,4 exemplifies the difference between prototypical retracing and stalling 2nd cycles: retraces are of normal duration and are not followed by a pause; stalls are lengthened and may be followed by a pause.

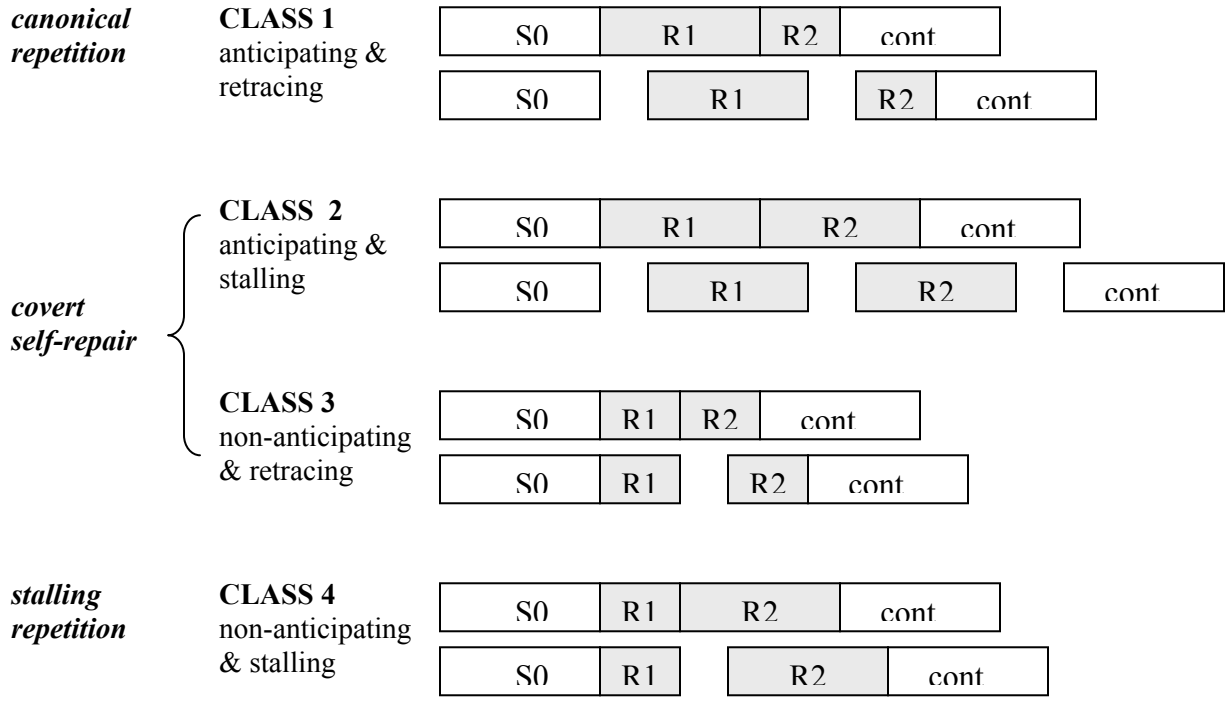


Fig. 4. *Schematization of syllable durations and pauses in the subclassification of 2-cycle repetition*. The lengths of the boxes represents duration, and a space between the boxes represents a pause. "...S0" stands for the context preceding the repetition, and "cont..." stands for the continuation following the repetition. Within a given class, the difference between two rows is whether the possible pauses are shown.

Plauché and Shriberg (1999) used hierarchical clustering to analyze 800 2-cycle repetitions of the function words "I" and "the" from the Switchboard corpus. The features in their analysis were syllable and pause durations, the ratio between syllable durations, pitch values and slopes, and non-modal voicing. From this analysis they identified three primary clusters, two of which—*canonical repetition* and *stalling repetition*—clearly correspond to the hypothesized classifications shown in Fig. 4. The third cluster seems to occupy a middle ground between classes 1 and 4. One possible explanation for the failure for all four predicted classes to emerge in the cluster analysis is that the hierarchical clustering algorithm produced an amalgam of two classes. The clustering occurred over many dimensions, and so there is no guarantee that some of the less relevant dimensions washed out a distinction between classes 2 and 3 in more relevant dimensions.

## 2. Method

### *2.1. Motivation and analogy to dynamic rhythm setting*

Previous research on repetition disfluency has not considered the *timing* of repeated syllables and has dealt almost exclusively with 2-cycle repetitions. This investigation analyzes timing in 3-cycle repetitions ("threepeats") of the English function words "and" and "I". Studying threepeats allows us to extract a relative phasing measure loosely analogous to the one employed by Cummins and Port (1998) to find a harmonic timing effect. In this analysis, the primary measure of relative phase is determined by the temporal location of the P-center of the second syllable (R2) relative to the interval defined by the P-centers of the first and third syllables (R1 and R3).

To some extent, the phase of R2 ($\varphi_2$) is not directly analogous to the phase of the stressed syllable in the Cummins and Port (1998) speech-cycling task. First, the event associated with R2 in a threepeat is in a substantial way identical to those immediately preceding and following, which is not the case in the speech-cycling task. Second, the events in 3-cycle repetitions occur on a smaller time-scale: whereas the base repetition period in the speech-cycling task ranged from 1000 ms to 2330 ms, the range of periods defined by R1 and R3 in threepeats covers shorter durations, from about 350 ms to 1200 ms. Third, in this case there are no external stimuli or task instructions: the patterns observed are entirely spontaneous. Fourth, the period of repetition is not defined by a repeated interval; in this case the interval from R1 to R3 occurs only once.

### *2.2. Data collection.*

Tokens containing 3-cycle repetitions of *and* & *I* were collected from the Switchboard-1 corpus of human-to-human telephone conversations. These conversations are spontaneous speech in which volunteers were given a topic for discussion with a stranger. Non-standard forms of the lexemes *and* & *I* were accepted, including [*aə*] for /*ai*/; [*ən*], [*æ̃*], [*n*] for /*ænd*/; as well as any variants bearing minor coarticulation with surrounding segments. The corpus contains approximately 3 million words, spoken by over 500 speakers from every major dialect of American English.

The lexemes *and* & *I* were chosen because these words exhibit the highest frequency of threepeating in the corpus. There were around 115,000 1-cycle tokens of each of these forms, with slightly more *I* than *and*. The chance of any word in the corpus being *and* or *I* is rather high, about 7.7%. The chance of these forms being repeated once is about 3.9%. There were 213 threepeats of *and*, 362 of *I*—thus the chance of either form being threepeated is about 0.25%, and over the whole corpus the chance of finding an *and* or *I* threepeat is rather small: 0.02%, i.e. approximately two times out of every ten thousand words.

| **Corpus size = ~3,000,000** | **and** | **I** | **both** |
|---|---|---|---|
| **Total freq.** | 112,536 (3.75 %) | 118,436 (3.95 %) | 230,972 (7.70 %) |
| **1-cycle** | 103,061 (3.44 %) | 107,794 (3.59 %) | 210,855 (7.03 %) |
| **2-cycle** | 4,364 (0.15 %) | 4,716 (0.16 %) | 9,080 (0.30 %) |
| **3-cycle** | 213 (0.01 %) | 362 (0.01 %) | 575 (0.02 %) |
| **4-cycle** | 27 (< 0.00 %) | 31 (< 0.00 %) | 58 (< 0.00 %) |

Table 2. *N-cycle token frequency for "and" & "I" in the Switchboard-1 corpus*

Despite the relative infrequency of threepeating, one should not consider the phenomenon unworthy of investigation. Threepeats occur only slightly less frequently with other function words like *a*, *the,* and *uh*. Future investigations should analyze these forms as well.

*2.3. Data normalization*

A sizeable number of tokens (56%) were excluded from the analysis. The reasons for these exclusions are listed below, and Table 3 shows frequency counts for both forms.

- *Inspiration (@)*. (11%) Tokens with inspiration intervening between cycles were excluded, since one can reasonably argue that such activity will distort measurement of inter-syllabic temporal relations.
- *Segmental anticipation*. (17%) Anticipatory coarticulation of a segment(s) belonging to the onset of the continuation sometimes occurs on even the 1st and 2nd cycles of a 3-cycle repetition (e.g. ...*and and*[s] . *and someone*...). Although not collected in the original sample, the production of a syllable or word from the continuation often occurs between cycles. These cases can be viewed as anticipation on a larger domain, yet sometimes the anticipated material does not end up being uttered in the continuation, making that material seem more like a planning error. These phenomena are worth further consideration, but are possible confounds to the rhythmic measures pursued here. If the domain of an anticipation was judged to be segmental or greater, the token was excluded.
- *More than 3 cycles*. (10%) For the current study, these tokens have been excluded, but they deserve more attention in the future.
- *Transcription and other problems* (18%) Sometimes the transcription was judged incorrect, resulting in a token not having 3-cycles of repetition in actuality. Other problems included transient noises, low sound-levels, and missing files.

| | & & & | @ | SEGMTL. ANTICIP. | > 3-CYCLE | TRANSCR. PROBLEM | OTHER | TOTAL |
|---|---|---|---|---|---|---|---|
| ***and*** | 126 (.1) | 24 (2.2) | 32 (3.3) | 32 (1.5) | 22 (.0) | 30 (.3) | 266 (38%) |
| ***I*** | 215 (.0) | 23 (1.3) | 84 (2.0) | 36 (.9) | 37 (.0) | 42 (.2) | 437 (62%) |
| **TOTAL** | 341 | 47 | 116 | 68 | 59 | 72 | 703 |
| | | | | | | p = .04 | $\chi^2 = 11.7$ |

Table 3. *Observed frequency and chi-square contribution of excluded token classes by form*. "& & &" = 3-cycle repetition retained for analysis; "@" = inspiration. Numbers in parentheses represent contribution of cell to chi-square value.

Chi-square analysis comparing the frequencies of the various classes of discarded tokens between *and* & *I* shows that threepeats of *and* & *I* differ with respect to the reasons for which they were sometimes excluded from analysis ($X^2 = 11.7$, p=.04). One major contributor to the difference is the higher relative frequency of inspiratory activity with *and* compared to *I*. There is probably a syntactic explanation for this: *and* can be used as a sentential discourse connector, and there is a greater likelihood of inspiration inter-clausally. Another major contributor is the higher relative frequency of segmental anticipation encountered with *I* compared to *and*. This arises from a tighter morphosyntactic affiliation (and hence greater fusional relation) between subject pronouns and subsequent auxiliary and modal auxiliaries (e.g. *I'll*, *I'm*, *I'd*). A third difference is that *and* threepeats are more likely to be part of a 4-cycle repetition. Despite these differences we will assume that—with respect to R2 phasing in 3-cycle repetition—*and* & *I* threepeats are to a sufficient extent commensurate.

*2.4. P-center estimation.*

Up to this point I have neglected an important question: how exactly do we define the temporal location of the "event" that corresponds to a syllable. This is not trivial, since a syllable is naturally an interval in time, rather than a point. Fortunately, previous research has provided a useful way to estimate this event, known as the "P-center", or "beat". A naive view might posit that syllable onsets (whether acoustic or articulatory) are the relevant events; alternatively, one could propose that amplitude peaks are good candidates. However, several experimental paradigms have shown that neither onsets nor amplitude peaks are the best choices for coordinated events.

Predicting where the beat of a syllable occurs is far from trivial. The first modern experimental paradigm to address this issue was *finger tap alignment* (Allen 1972, 1975), in which subjects aligned finger taps with syllables they produced. Allen found that subjects tapped their fingers somewhere close to the onsets of the vowels of stressed syllables; he called the temporal location of the corresponding finger tap the "production-center" of a syllable.

A different experiment, called a *dynamic rhythm setting task* (Morton, Marcus, and Frankish 1976), showed that there is also a perceptually salient center of a syllable near the vowel onset. In this task, a synthesized base syllable (A) is repeated at periodic intervals, and subjects use a knob to adjust the timing of another synthetic syllable (B), until they feel the beats of all syllables are regular (see Fig. 5). Initially the location of (B) is randomly chosen to occur somewhere between 20-80% of the base period.

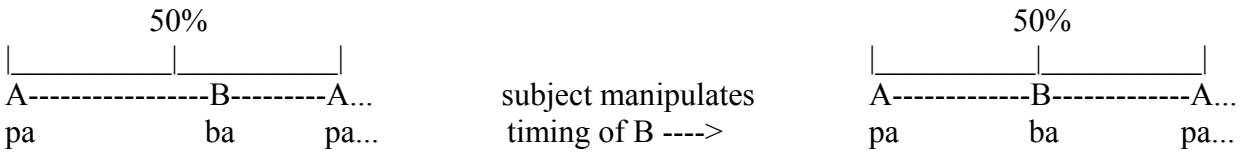


Fig. 5. *Dynamic rhythm setting task*. A synthesized base syllable (A) is repeated at periodic intervals. The subject adjusts the timing of a second syllable (B) with a knob until they feel the beats of A and B are regular. Initially the location of (B) is randomly chosen to occur somewhere between 20-80% of the (A)-(A) period.

In these experiments the salient instant or "center" of a syllable falls somewhere near the onset of the vowel. However, Morton et. al. (1976) found that the exact location of the center is influenced by the presence and duration of onset and coda clusters. The size of this effect is on the scale of approximately 5-20ms. The researchers called this center-point the "perceptual moment of occurrence," or "P-center," since the task involves only the perception of isochrony.

The notion of a "center" of a syllable implies an interesting idea: the relevant dimension in which a syllable-center can be defined is not real-time, but rather some perceptual- or articulatory-time that is not directly measurable and must be non-linearly mapped to real time. Furthermore, the center of the syllable in this dimension might be conceptualized as its midpoint (median), center of gravity (mean), peak (mode), or some other type of center. If one knows the mapping between the acoustic/perceptual/articulatory dimension and the P-center dimension, as well as the nature of the center itself, one should be able to predict the location of a P-center from an acoustic/perceptual/articulatory signal.

This expectation has led a number of researchers to attempt to develop algorithms to approximate P-centers from acoustic signals. For example, Howell (1988) used the amplitude envelope of a syllable to predict its P-center; Pompino-Marschall (1989) used a gammatone filterbank (which approximates auditory nerve responses) and a nonlinear function of energy events defined by thresholds in syllable constituents; Scott (1993) used the energy in a specific bandwidth of the spectrum, and Cummins and Port (1998) used a variation on this last method where the bandwidth was 700-1300Hz—this last method is the one employed in the current study.

While the above approaches treat P-centers as perceptual, Fowler (1979) argued that they are directly associated with gestural events. There is evidence that the exact location of P-centers exhibits some inter-subject variation; this can be taken to indicate that the phenomenon is articulatory if one assumes that speakers exhibit more gestural than perceptual differences. de Jong (1994) provided evidence that P-centers are associated with a complex of acoustic/perceptual and articulatory events. There are other forms of evidence for the psychological reality of P-centers, notably that deviations from regular timing of P-centers are noticed by pre-babbling infants (Fowler, Smith, & Tassinary 1986).

Because of their perceptual and articulatory salience, P-centers provide decent temporal estimates of syllabic-events that are coordinated, regardless of whether that coordination is conceptualized as controlled or self-organized.

*2.4.1. Labeling and P-center estimation procedures*

In this study, each threepeat token retained from the exclusion process was hand-labeled on the basis of visual and auditory cues, using Praat (speech analysis software). The hand-labeling produced an interval tier (Fig. 6) defining the following:

- (S0): stressed syllable preceding the 1st cycle of repetition
- (W0, W0+): unstressed syllables following S0, if present.
- (R1, R2, R3): the 1st, 2nd, and 3rd syllables in the threepeat.
- (P1, P2, P3, P4): pauses occurring before their respective syllables (defined here as a silent interval longer than 100ms which cannot be attributed to a stop consonant)
- (W4, W4+): unstressed syllables following R3, if present.
- (S4): stressed syllable following R3.

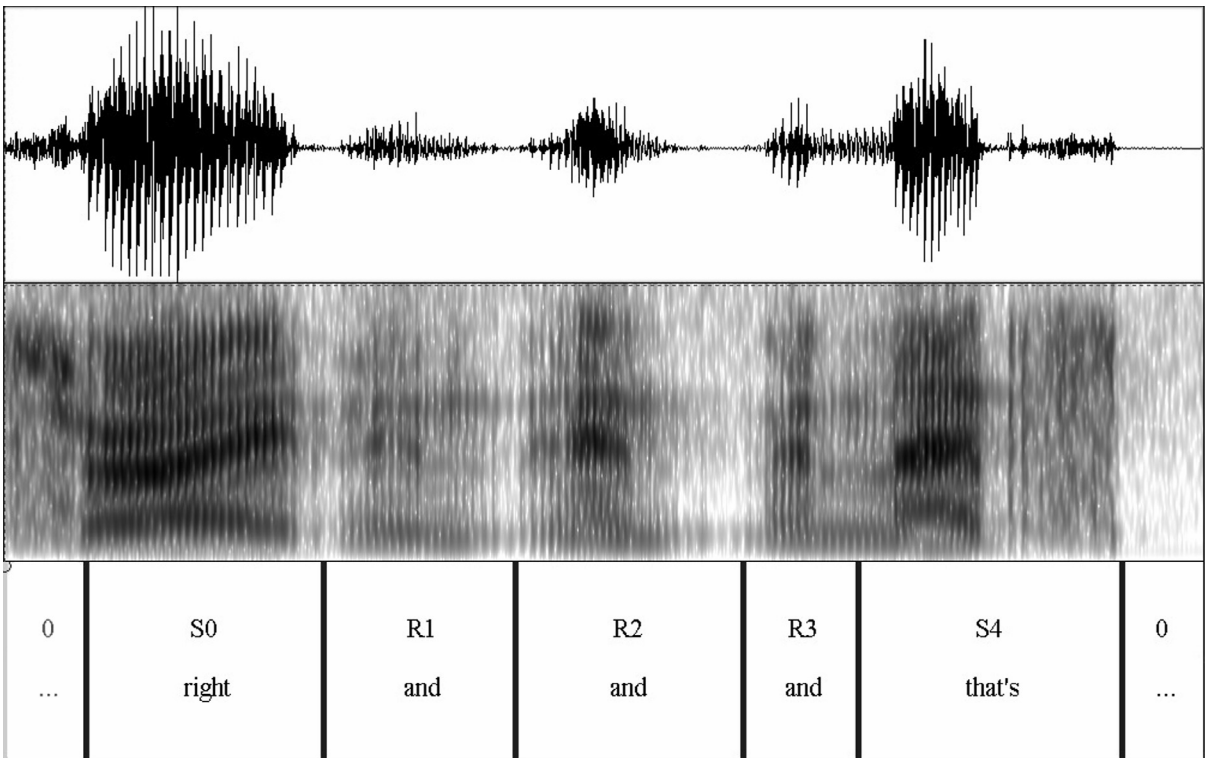


Fig. 6. *Example of "and" threepeat with labeled intervals.* Top: speech signal. Middle: spectrogram. Bottom: text labels (words added in figure for convenience). From the utterance: "that's right and and and that's been really um way it's been for me for the last five or six years tha[t]- i returned a TV set to be repaired but the problem was that uh we had a a lightning storm and" (sw3891B-ms98-a-0025).

The interval labels aided in the automatic extraction of P-centers with a Matlab script. This extraction process followed the procedure described in Cummins and Port (1998). The acoustic signal is filtered with a first-order Butterworth bandpass filter from 700Hz - 1300Hz.

This eliminates fricative noise and energy from F0 (and attenuates energy from low F1). The signal is then rectified (using absolute value) and smoothed with another Butterworth filter, which in the present study was a fourth-order lowpass Butterworth filter with a 10 Hz cutoff. The resulting band-limited signal is thus relatively smooth, allowing for consistent measurements to be taken.

The estimate of P-center that will be considered in this report is the energy-rise centerpoint,[4] which is defined as the point in time corresponding to the first midpoint in signal amplitude between the local minimum and maximum signal amplitudes. The energy-rise center is an approximation of the time of maximum velocity of the band-limited energy function at vowel onset.

The filtered signal contours for most instances of *and* & *I* follow smooth trajectories from energy minimum (or silence) to energy maximum (Fig. 7). Hence the measurements cannot be seen as artifacts of the hand-labeling or P-center estimation. Furthermore, because these forms are vowel-initial, the effects of consonantal onsets on P-center locations are absent.

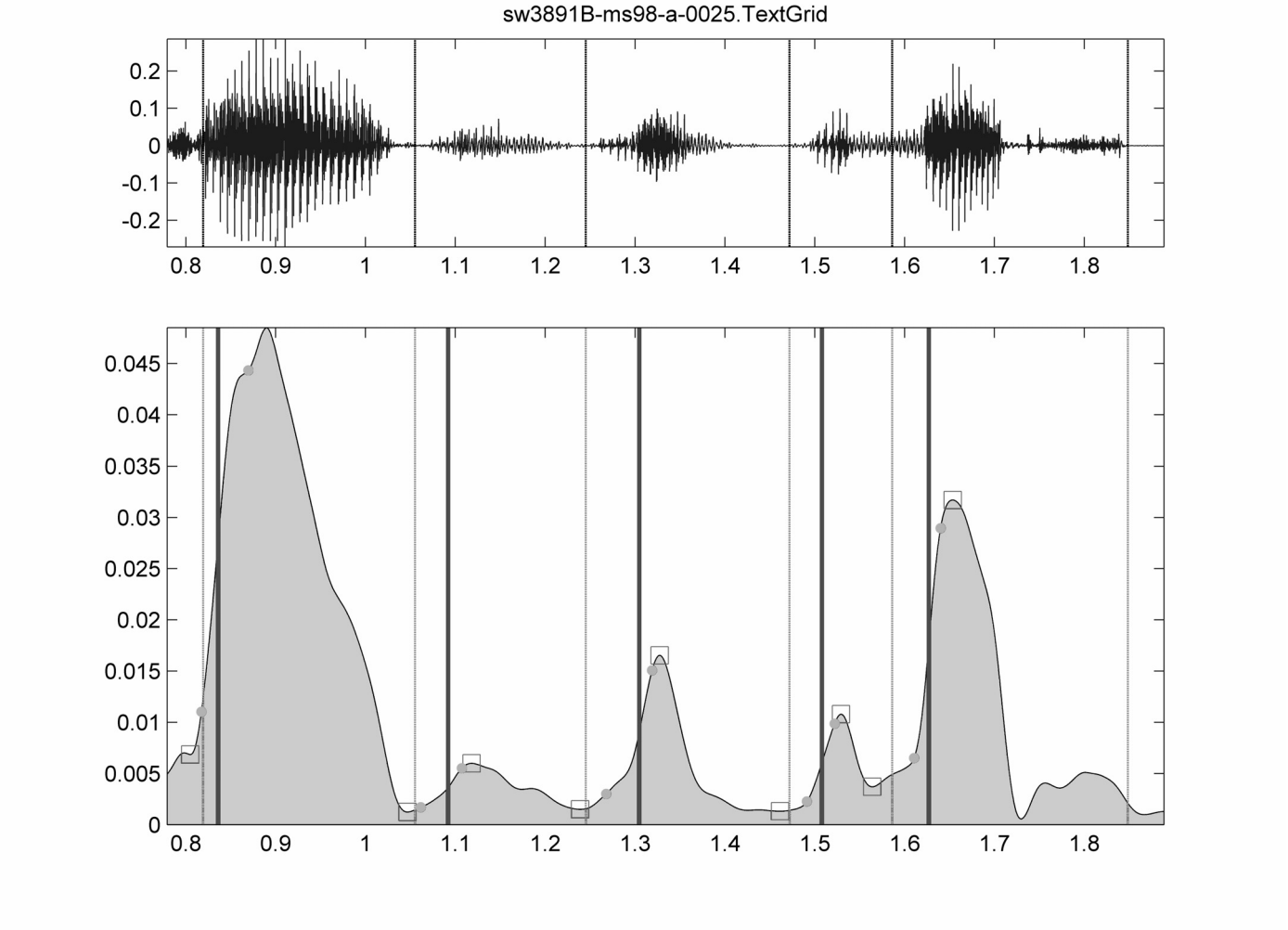


Fig. 7. *P-center locations and energy envelope of token from Fig. 6*. Thick lines indicate p-centers, thin lines indicate interval labels, boxes indicate local energy maxima and minima, dots indicate 10% above minima and 10% below maxima.

A handful of very brief and reduced R1 and R3 cycles (~8) exhibited more linear contours, rather than peaks and valleys. This was also sometimes the case for W0 and W4 cycles. If a contour had a mostly positive slope throughout the labeled interval, the energy maximum was located closer to syllable offset, introducing a negligible degree of noise in the data. If a contour had a mostly negative slope throughout the labeled interval, the onset was chosen to represent the location of the P-center.

2.5. *Transformation to phase*

Schematically, the relevant R2 phase information is initially measured in a temporal dimension (Fig. 8, top). This is then converted to a representation in a phase dimension (Fig. 8, bottom) by normalization of the duration between the p-centers of R1 and R3 (pR1 and pR3) to 1. This normalization is applied to all tokens, allowing the phase of R2 ($\varphi_2$) to be compared across utterances.

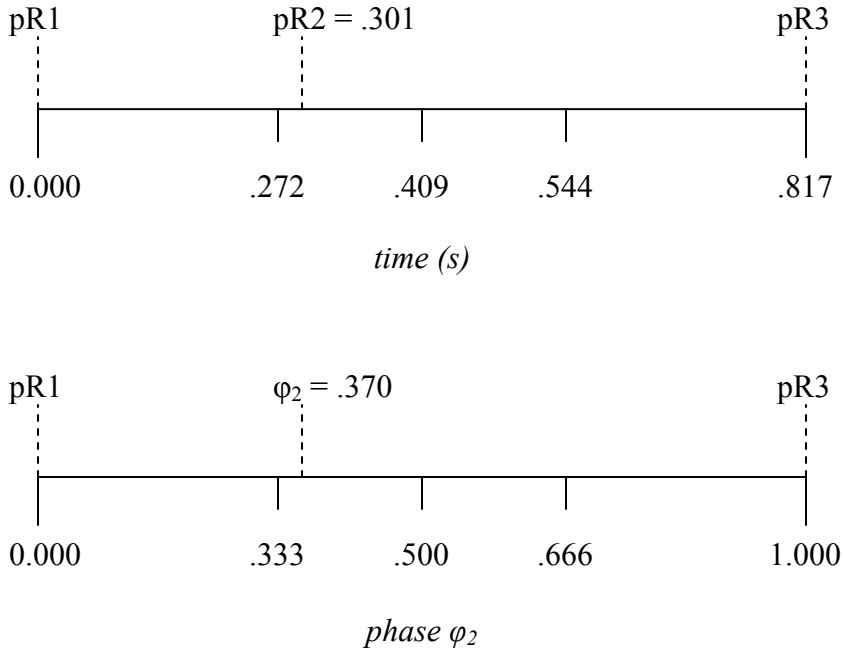


Fig. 8. *Relation between time and phase*. Top: location of the p-center of R2 in a temporal dimension. Bottom: phase of the p-center of R2 in a phase dimension.

## 3. Results

### 3.1. *Predictions of the null model*

Before examining the data, let us consider a naive hypothesis about the factors governing the distribution of our task-dynamic variable, $\varphi_2$. What do we expect the distribution of $\varphi_2$ to look like? For a null hypothesis, we might assume that the sole factor is some sort of "repulsive force" that operates between adjacent syllables. The effect of this force is to act against any two syllabic P-centers occurring too close together in time; the repulsive force should not be understood as an inviolable constraint, but rather as interacting with other repulsive forces. Let us further assume that the strengths of the repulsive forces exerted upon R2 by R1 and R3 are equivalent, and that forces exerted by non-adjacent syllables S0, W0, W4, and S4 can be ignored. Fig. 9 represents this null model:


$$rF_{R1R2} = rF_{R2R3}$$

Fig. 9. *Schematization of the null model*. rF = repulsive force. R1 = 1st-cycle, etc.

If the repulsive forces $rF_{R1R2}$ and $rF_{R2R3}$ are equal, the beat (P-center) of R2 is expected to occur exactly halfway between the beats of R1 and R3, i.e. $\varphi_2$ should be .50. We can also assume that there is some noise in the system, so $\varphi_2$ will be distributed normally around a mean of .50.

### *3.2. Observed $\varphi_2$ distributions*

#### *3.2.1. $\varphi_2$ distribution for all data*

The distribution of $\varphi_2$ for the entire dataset is shown below in Fig. 10. Its shape appears to be consistent with the null model prediction that R2 phase should be normally distributed. The mean phase, however, is .455, which is different from the expected .500 ($t = -8.08$, $p < .001$)—we will have more to say on this difference in subsequent sections.

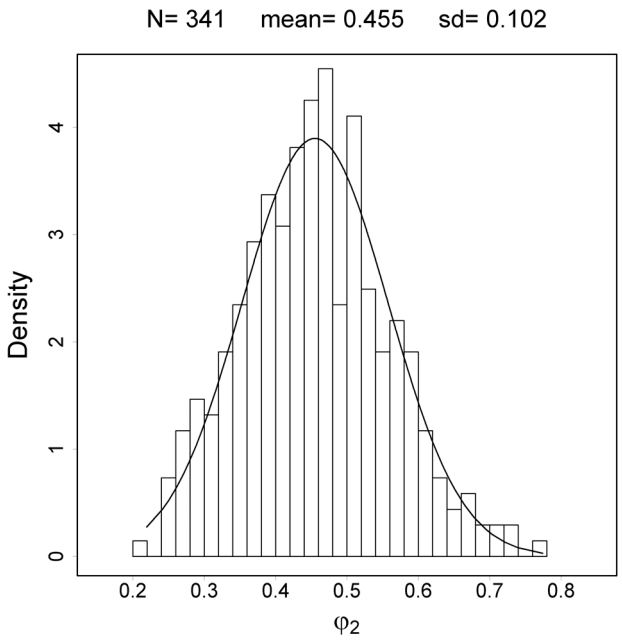


Fig.10. *Observed R2 phase distribution*. Histogram of observed $\varphi_2$ with normal curve superimposed (number of tokens = 341, mean $\varphi_2$ = .455, standard deviation = .102)

3.2.2 *$\varphi_2$ distributions in fast and slow speech-rate subsets*

Though the null model suffices for predicting the shape of the $\varphi_2$ distribution for the entire dataset, when we consider a slow-speech subset of the data, an intriguing pattern emerges. There are good reasons to speculate that the duration of the interval from pR1 to pR3 might interact with $\varphi_2$. To wit, in the HKB model of bimanual coordination, changes in rate of repetition lead to a phase transition in the stability of anti-phase coordination; furthermore, Cummins and Port (1998) observed a reduction for some subject in the use of 1/3 and 2/3 rhythmic modes with a shorter repetition period in the speech cycling task. The duration of the interval from pR1 to pR3, though not exactly analogous to the repetition periods in those tasks, can serve as a proxy for rate of repetition or "speech-rate" in threepeats.

Fig. 11 shows the distribution of pR1-pR3 interval durations. Here one can clearly see that the distribution is skewed rightward. The dashed line represents the median value of interval durations. We will use the median value to separate the dataset into "fast-rate" and "slow-rate" subsets, although the results are not qualitatively different if the mean is used instead.

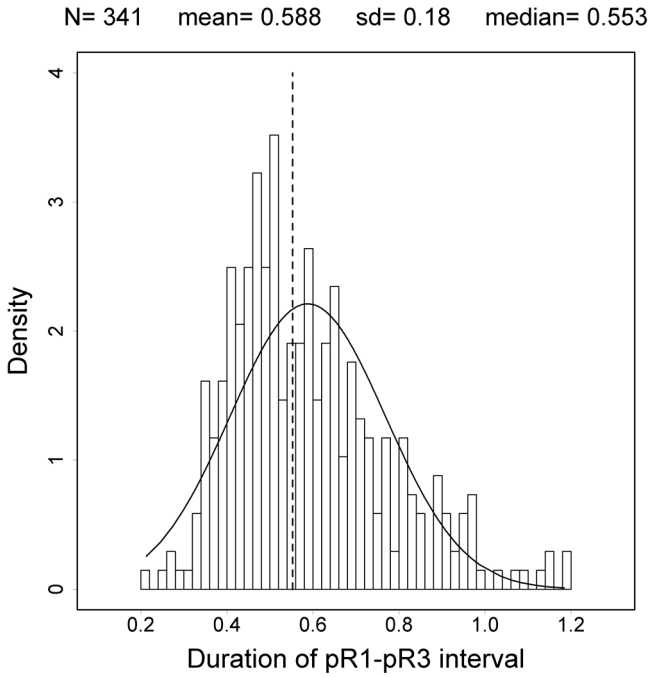


Fig. 11. *Distribution of pR1 to pR3 interval durations.* Histogram of observed pR1-pR3 durations with normal curve superimposed (number of tokens = 341, mean = .588, sd = .18). Dashed line represents median value (.553).

Fig. 12 shows the $\varphi_2$ distribution in fast-speech (left) and slow-speech (right) subsets. In the fast-speech subset, the shape of the distribution is expectedly normal. In the slow-speech subset, however, the shape of the distribution is quite remarkable. Rather than a single mode, there appear to be three modes. This trimodality is the central finding of this paper.

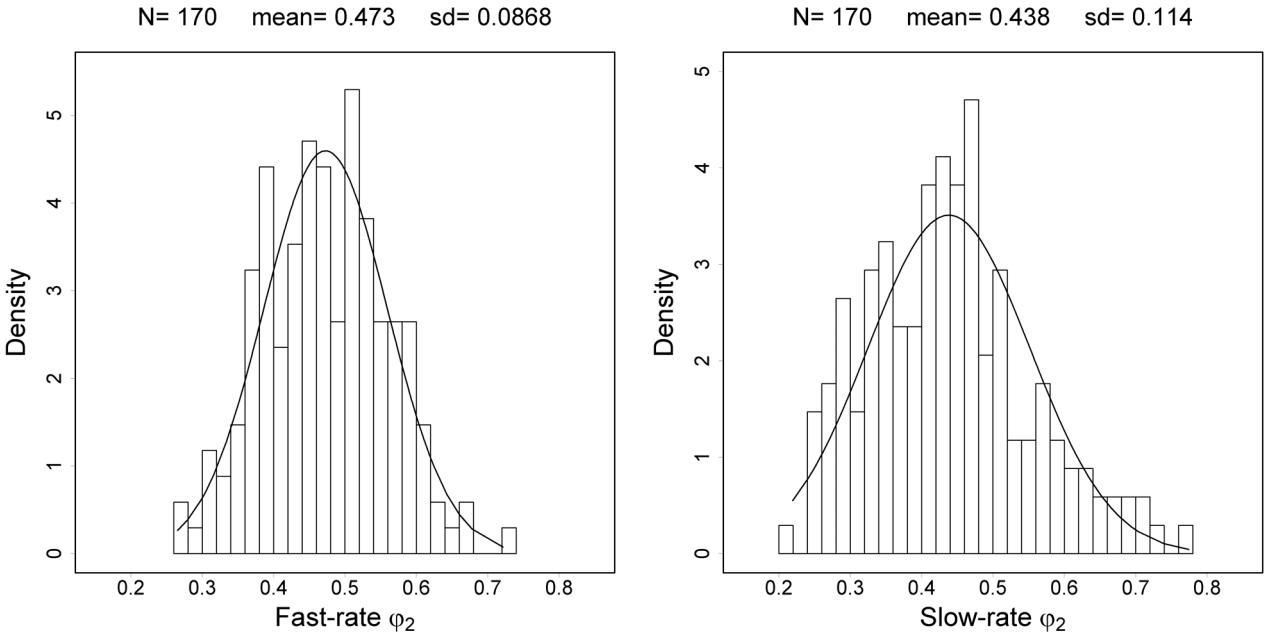


Fig. 12. *R2 phase distributions in fast- and slow-speech subsets*. Left: histogram of fast-speech $\varphi_2$ with normal curve superimposed (number of tokens = 170, mean = .473, standard deviation = .087); Right: histogram of slow-speech $\varphi_2$ with normal curve superimposed (number of tokens = 170, mean = .438, standard deviation = .114).

Note that mean R2 phases in the fast-rate and slow-rate subsets are different ($t = -3.16$, $p<.002$). We will address the reason for this later. More to the point, the populations exhibit different degrees of variance ($F=1.71$, $p<.001$). In the following section we will see that absolute deviation from mean phase is to some extent predictable from the duration of the pR1 to pR3 interval.

3.2.3. *Correlation of $\varphi_2$ deviation and rate*

Fig. 13 shows the correlation between the absolute value of $\varphi_2$ deviation and pR1-pR3 interval duration. The dashed line represents the linear regression (intercept = 0, slope = .13). Analysis of variance on the two variables shows that they are positively correlated ($R^2 = .63$, $F(1,340) = 590.7$, $p<.001$).

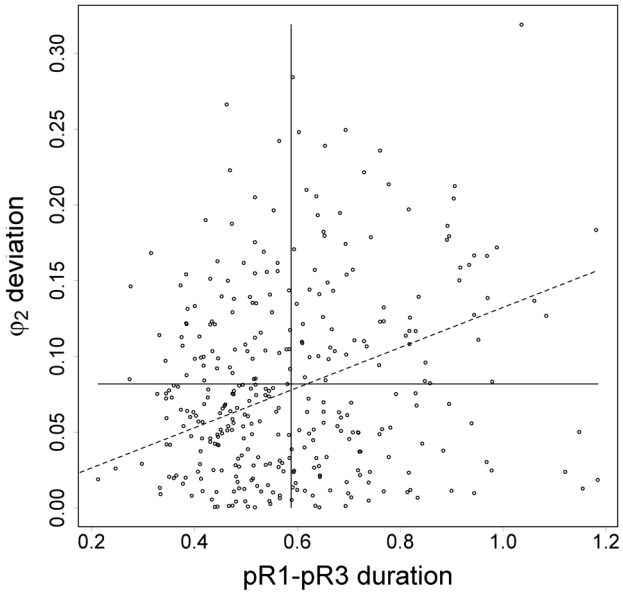


Fig. 13. *Linear regression of R2 phase deviation as a function of pR1-pR3 interval duration*. The solid lines correspond to the means of absolute phase deviation and pR1-pR3 duration. The dashed diagonal line represents the linear regression (intercept = 0, slope = .13).

The null hypothesis can predict this relation between $\varphi_2$ deviation and speech-rate if we additionally stipulate that the noise is dependent upon speech-rate. The relation can also be seen as a consequence of Weber's Law: the timing of longer intervals correlates with greater variability. However, the trimodality of slow-speech $\varphi_2$ cannot be understood in these terms.

3.2.4. *Trimodality of slow-speech $\varphi_2$ distribution.*

The shape of the slow speech-rate $\varphi_2$ distribution is patently non-normal, and we can quantify our confidence in this observation. The question is whether a single Gaussian probability density function is a better model of the distribution than a mixture model of Gaussians. By "better model," I mean a model that 1) minimizes the difference between the observed phases and model-predicted phases, 2) is not fitting only noise in the data, and 3) is not fitting a small group of outliers.

A single gaussian has parameters for the mean ($\mu$) and standard deviation ($\sigma$) of the density function. For our slow-rate $\varphi_2$ the parameters are $\mu = 0.438$ and $\sigma = 0.114$. Thus the parameter space of the single-gaussian model is two-dimensional.

It is possible to compare a single gaussian model with a model involving a mixture of gaussians. This is accomplished by considering the single gaussian model as a special case of a mixture model. In the present circumstance, we are interested in a mixture model of three gaussians. The full parameter space consists of eight parameters: $\mu_1$, $\mu_2$, $\mu_3$, $\sigma_1$, $\sigma_2$, $\sigma_3$, $p_1$, $p_2$. The latter two parameters determine the contribution of each gaussian to the mixture density (the third gaussian has contribution $1-p_1-p_2$). The single gaussian is thus a special case where $\mu_1 = \mu_2$

$= \mu_3$, which can be expressed equivalently by setting either $p_1$ or $p_2$ to 1, so that only one of the gaussians contributes to the density.

Maximum likelihood estimation can be used to compute an estimate of the parameters of the mixture model. This was accomplished with a numeric minimization algorithm designed for non-linear equations (implemented with "nlm()" in the R stats package; the function help file cites references Dennis and Schnabel (1983) and Schnabel, Koontz, and Weiss (1985)).

The minimization algorithm treats the observed data as fixed parameters while varying the model parameters. Each step of the algorithm computes a value for the negative of the log likelihood of the mixture model. The mixture model parameters are adjusted from step to step according to whether their adjustment leads to a further minimization of the negative of the log likelihood. When the gradient of the function across steps is sufficiently small, the algorithm stops. The parameters at this point represent the parameters of the mixture model that, given the observed data, maximize the log likelihood of the model.

Because of the high-dimensionality of the mixture model parameter space and because of the non-linearity of the model, the algorithm is sensitive to initial parameter estimates and step sizes. It is thus useful to run the algorithm recursively, fixing various sets of parameters on successive runs. The procedure that was followed was first to obtain initial parameter estimates from visual inspection of the histogram of phase data; then on the first run of the algorithm, the simplification was made that the standard deviations of the three component gaussians were equal. Output parameter estimates were then fed back into a second run of the algorithm in which means were fixed and standard deviations were allowed to vary. The output parameter estimates of this second stage were sufficient for statistical significance.

To test the null hypothesis that $\mu_1 = \mu_2 = \mu_3$ against the alternative that $\mu_1 \neq \mu_2 \neq \mu_3$, the likelihood ratio test can be used. Let $l(.)$ denote the log likelihood; then $2l(\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, p_1, p_2) - 2l(\mu_1 = \mu_2 = \mu_3 , \sigma_1 = \sigma_2 = \sigma_3)$ converges to $\chi^2_{r-q}$, where $r$-$q$ is the degrees of freedom, $r$ being the dimension of the mixture model, $q$ the dimension of the single gaussian model. Thus the ratio between the log likelihood estimates can be evaluated for significance against the chi-square distribution.

Fig. 14 (left) shows the histogram of slow-rate $\varphi_2$. The single gaussian model (dashed line) and 3-gaussian model (solid line) are also shown. Fig. 14 (right) shows the composition of the trimodal model. The component gaussians have means of .331, .459, and .637, with comparable standard deviations. We can be reasonably confident that the mixture model meets the three criteria listed above ($\chi^2 = 20.1$, df = 6, $p < .005$). The same cannot be said for fast-rate $\varphi_2$.

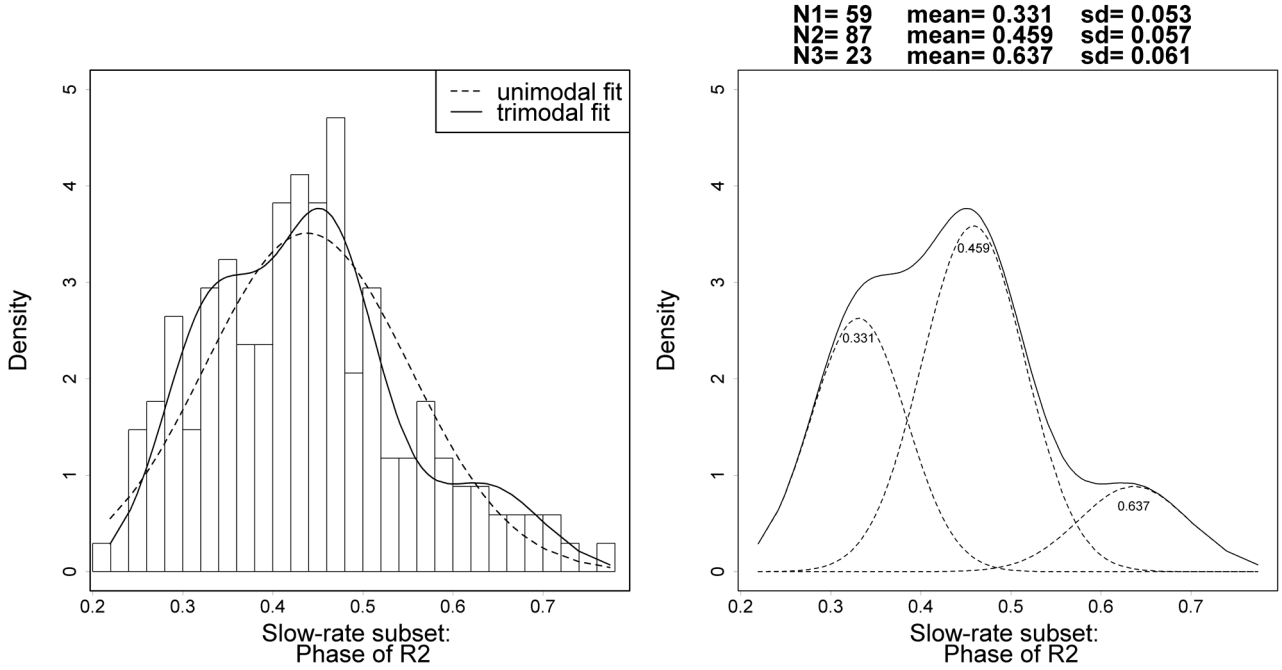


Fig. 14. *Trimodality of observed* $\varphi_2$ *in the slow speech subset.* Left: histogram of observed phases in the slow-rate subset, with uni-modal (dashed line) and tri-modal (solid line) density model. Right: internal composition of the 3-gaussian mixture model (means = .331, .459, .637; standard deviations = .053, .057, .061; percentages of density = 35%, 52%, 13%)

The null model cannot account for the observation of trimodality without some additional stipulations. The existence of three modes in the distribution tempts us to posit some sort of categorial factor. A paramount question here is whether the values of the means of the component gaussians are in themselves significant. Clearly they are not extremely different from the harmonic phases (.333, .500, .666) in the task-dynamic coupled-oscillators model of Cummins & Port (1998) and Port (2003). In the next section, we will consider an adjustment to the phase measures, which to some extent improves the correlation between the observed phase modes and the harmonic ones.

*3.3. Speech-rate slowing*

Recall that for the entire dataset, the mean $\varphi_2$ is .454, different from the predicted .500. Within the confines of the null model, this difference could be accounted for by positing that the repulsive forces between R2 and R3 are stronger than those between R2 and R1. This is represented schematically in Fig. 15.

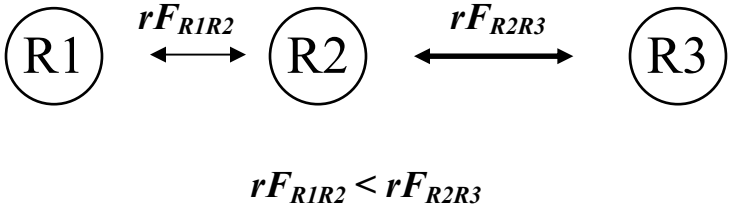


Fig. 15. *Null model with unequal repulsive forces*. R1 = 1st cycle of repetition, etc. rF = repulsive force.

Unequal repulsive forces, however, are an entirely stipulative explanation for the data. An explanation that is more in line with research on repetition disfluency points to a different cause: a local decrease in speech-rate, which is evidenced by lengthening and pausing. Shriberg (2001) reported evidence of lengthening of repeated forms in 2-cycle repetition, and Clark and Wasow (1998) and Hieke (1981) observed unfilled pauses in various situations.

Syllable durations and cycle durations (inter-p-center-intervals) in the present data also support the idea that speech rate often slows to some extent during repetition disfluency. First, the mean duration of the R2 is greater than the mean durations of R1 and R3 (Fig. 16, left). Durations of R2 syllables are on average longer than those of the corresponding forms in fluent speech. Second, mean pR2-pR3 cycle duration is greater than mean pR1-pR2 and pR3-pS4 cycle durations (Fig. 16, right). These observations hold true when *and* and *I* are considered separately as well.

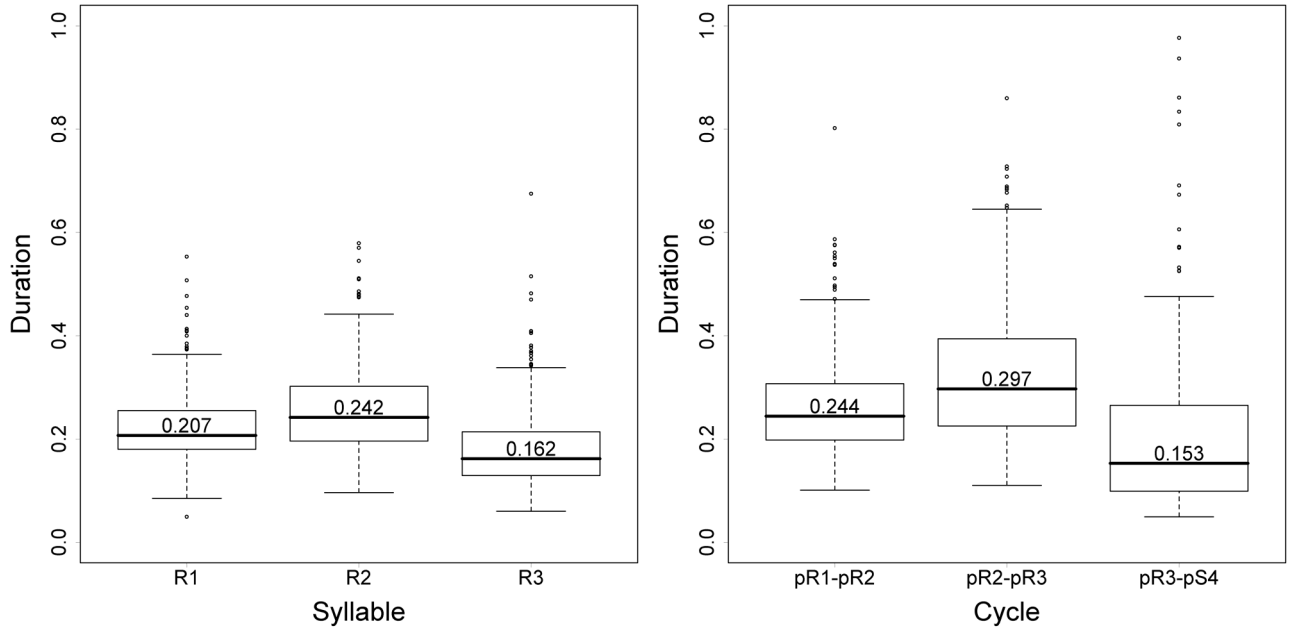


Fig. 16. *Mean syllable and cycle durations*. Left: syllable durations of R1, R2, and R3. Right: cycle durations (inter-p-center- intervals) of pR1-pR2, pR2-pR3, and pR3-pS4. Bold lines in each boxplot represent median values, which are labeled.

Furthermore, the relative frequencies of pauses across locations are also evidence of a local slowing of speech-rate in repetition disfluency. Pauses are defined here as silent intervals longer than 100ms not attributable to stop consonants. Fig. 17 shows that P2 and P3 (pauses before R2 and R3) are more frequent than P1 and P4 (pauses before R1 and S4).

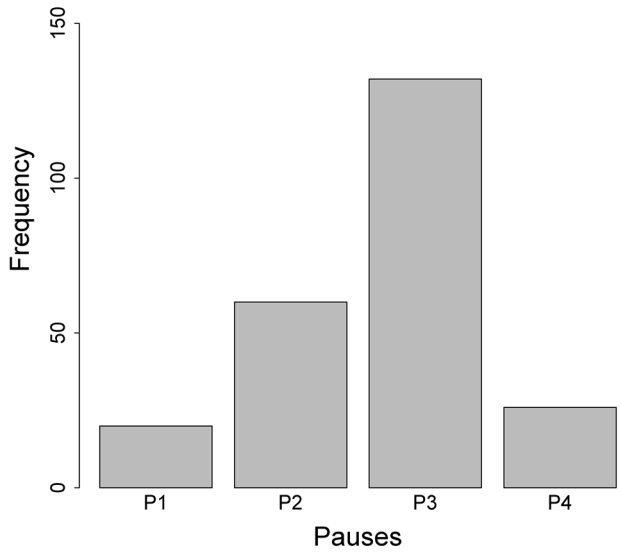


Fig. 17. *Frequency of pauses*. P1 = a pause before R1; P2 = a pause before R2, etc.

The relatively higher frequency of P3 vs. P2 suggests that speech-rate slowing normally continues between R2 and R3. Since R3 syllable and cycle duration is relatively brief, the implication is that normally there is an abrupt increase in speech-rate with R3. Thus the extent to which speech-rate slows during threepeats normally reaches a maximum somewhere in the R2-R3 interval. There is more to be said on this matter, but for now let us consider how to deal with speech-rate slowing.

3.3.1. *Phase adjustment for slowing*

If speech-rate indeed slows during a threepeat, then our measurement of $\varphi_2$ is likely to be distorted. Here we should note that there is no objective measure of "speech-rate," which depends upon both the type of event measured and the size of the window of measurement. For syllable-beats, a large window produces a smoother and more meaningful rate contour but overlooks local fluctuations in rate; a smaller window captures local fluctuations in rate but produces a more jagged, less generalizable, contour. If the window is so small that some inter-onset-intervals are longer than the window itself, then the measurement of rate is better-off being a measure of successive syllable durations, or of inter-onset-intervals, which take pauses into account (cf. Large & Jones 1999 for an inter-onset-interval approach to musical rhythm). Even better are the intervals between p-centers, the "inter-beat-intervals", which reflect a cognitively more salient event.

Furthermore, syllable shape, stress, and intonation interact with syllable duration and beat-spacing, normally obfuscating patterns in inter-beat-intervals of fluent speech. In the present study, however, these complicating factors are substantially avoided, and thus the inter-beat-interval measurement is somewhat more useful.

Regarding stress, there is no reason to believe that any of the cycles here bear lexical stress, because they are function words. Regarding intonation, although no investigation of pitch

contour during repetition was conducted, there appears to be little change in pitch across any given token. There are no cases in which a cycle of repetition is endowed with contrastive or emphatic discourse stress or focus. (A number of tokens exhibit some degree of laryngealization, which complicates pitch measurement; future research should investigate patterns of laryngealization in 3-cycle repetition disfluency.)

Regarding consistency in syllable shape, the issue is more complicated: although in a given token the citation forms of R1, R2, and R3 are identical (i.e. /*ai*/ or /*ænd*/), there is considerable variation in the pronounced forms (e.g. *i*: [*ai*], [*aə*]; *and*: [*ænd*], [*ən*], [*æ̃*], [*n*]) between tokens, though much less so within a given token. Future work should also investigate patterns of reduction in 3-cycle repetition disfluency.

Because of these complications, no precise measure of speech-rate is possible. Yet the inter-beat-interval measure can serve as an indirect and slightly noisy correlate of speech-rate. The idea behind this is that speech-rate can be conceptualized as a continuous variable influencing the temporal distribution of syllable events.

If speech-rate indeed slows during a three-peat, then compared to the null model with a constant speech-rate, what would we expect to be different in the observed distribution of R2 phases? Assuming that the change in speech-rate is linear during the pR1-pR3 interval, then the duration of the interval between pR2-pR3 should tend to be longer than the interval between pR1-pR2. This difference would bias $\varphi_2$ to be smaller than 0.500, which is exactly what we observe.

A linear transformation to adjust for slowing speech rate can be applied to the data, so that the mean is .500. This is just the linear transformation $T(\varphi_2) = \text{factor}_{\text{rate\_adj}} * \varphi_2$, where the rate adjustment factor is $.500 / \mu(\varphi_2)$. The effect of this transformation is to provide a distribution in which the mean of the transformed phase, $\varphi_2^T$, is exactly .500.

Fig. 18 below shows the histogram (left) and trimodal model (right) of the slow-rate subset $\varphi_2^T$. While the shape of the distribution is the same, the locations of the side peaks are slightly different: in addition to the mean peak at 0.504, the second and third peaks occur at 0.364 and 0.699—these values to a small extent better match the harmonic ratios .333, .500, and .666 than those corresponding to the untransformed distribution; the transformation is useful because the middle mode carries the most density, hence better centering this mode provides a better correspondence of density concentrations and harmonic ratios on .500 phase.

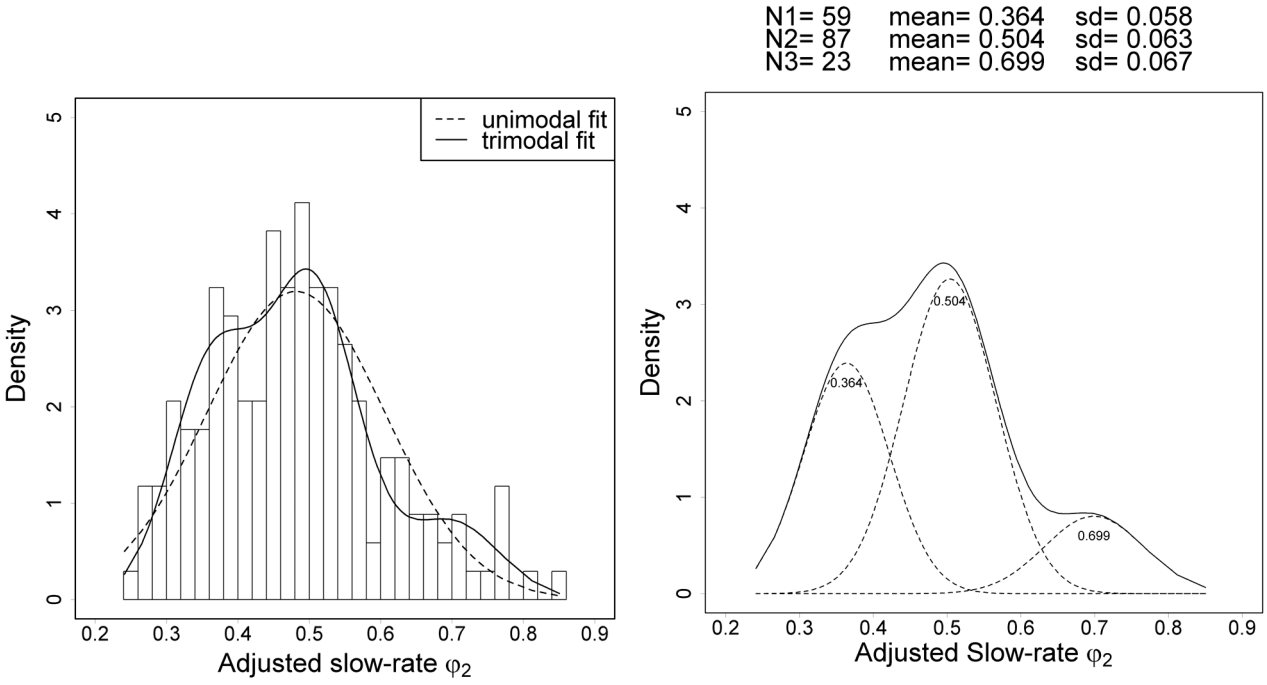


Fig. 18. *Trimodality of observed* $\varphi_2$ *in transformed slow speech subset.* Left: histogram of transformed observed phases in the slow-rate subset, with uni-modal (dashed line) and tri-modal (solid line) density model. Right: internal composition of the 3-gaussian mixture model (means = .364, .504, .699; standard deviations = .058, .063, .067; percentages of density = 35%, 52%, 13%)

## 4. Discussion

Why is there trimodality in the distribution of observed phases in the slow-rate subset, and why do the values of the modes approximate low-order harmonic ratios? Why is this trimodality absent in the fast-rate subset? In this section, we will consider how a coupled oscillator model similar the one used in Port (2003) can explain these findings, and explore a specific prediction this model makes about the timing of the syllables following threepeats. We will also discuss what previous functional or systemic accounts of repetition disfluency have to say about this phenomenon and return to the issue of local speech-rate slowing.

### 4.1. *Task-dynamic coupled-oscillators model of rhythm in threepeats*

One major difficulty in finding evidence for harmonic timing of stress-feet in spontaneous speech is that there is almost no repetition at the phrase level. In order to measure the relative phases of stressed syllables within a phrase cycle, the speech cycling task (Cummins and Port 1996, 1998) establishes the periodic repetition of a phrase for subjects. Instead of examining repeated phrases, the current study has examined the repetition of syllables in 3-cycle repetition disfluency, which occurs fairly commonly in spontaneous conversational speech.

The findings reported above are not precisely analogous to finding of harmonic timing in the speech-cycling task, for several reasons. First, the time-scale of the "base period" in threepeats is on the order of 400-1200 ms, while the time-scale of the phrase repetition period is

on the order of 1000-2300 ms. Second, the "base period" in speech-cycling corresponds to a repeated interval, but in our case the period R1-R3 is not repeated. Third, syllable-events in threepeats are not equivalent to stressed syllables in phrasal repetition, nor is a threepeat exactly analogous to a phrase.

Despite these differences, a very similar finding of harmonic trimodality is observed in the slow-rate subset of data. Because of theses similarities and differences, one might say in that slow-speech threepeats we observe *a* harmonic timing effect, rather than *the* harmonic timing effect. In other words, harmonic timing appears to operate on multiple time scales.

Furthermore, the harmonic trimodality we observe in threepeats is suggestive of the same sort of coupled-oscillators model Cummins & Port (1998) and Port (2003) use to understand rhythmic patterns in speech-cycling tasks. In the Port (2003) development of the model, subjects are believed to adopt one of two possible frequency-locking ratios between a pair of coupled oscillators, either 1:2 locking or 1:3 locking, represented in potential field equations (1) and (2) below. Recall from section 1.3 that these potential functions describe phase attractors at 0, 0.5, 0.33, and 0.66. The task-dynamic variable in the model represents the relative phase of the first and second stressed syllables of the phrase.

| 1:2 locking | 1:3 locking |
|---|---|
| (1) $V(\varphi) = -\cos \varphi - \cos 2\varphi$ | (2) $V(\varphi) = -\cos \varphi - \cos 3\varphi$ |

Another way to think about this model is to consider the case of three coupled oscillators, as described by the potential in equation (3). We can endow each oscillator with an amplitude parameter. Recall that one of the important control parameters of the HKB two-oscillator model is the ratio of the amplitude parameters (section 1.3). As this ratio is varied, the system undergoes a phase transition between one and two stables modes of relative phase.

In the three-oscillator system below, let us assume that the 2$^{nd}$ and 3$^{rd}$ harmonic oscillators are mutually competitive, which can be represented by the relation in equation (4). The "choice" between either 1:2 or 1:3 mode-locking can then be viewed as a phase transition, by taking either $k_2/k_1$ or $k_3/k_1$ as a control parameter. As this parameter is varied, the dynamics change from three stable modes to two stable modes, or vice versa. This transition from three stable modes to two stable modes is shown in Fig. 19.

(3) $V(\varphi) = k_1 - \cos\varphi - k_2 \cos 2\varphi - k_3 \cos 3\varphi$

(4) $k_2 + k_3 = 1 \qquad k_2, k_3 > 0$

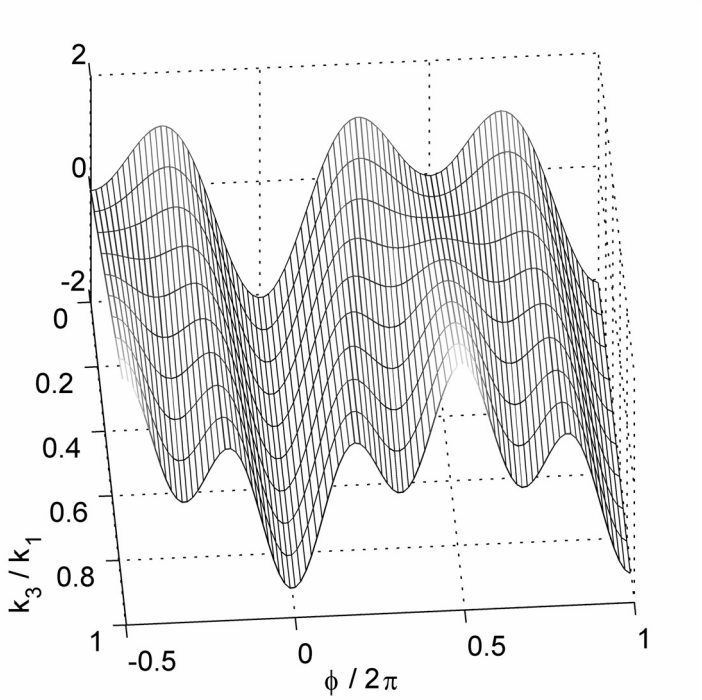


Fig. 19. *Phase transition from 3 stable modes to 2 stable modes*. Potential space corresponding to equation (3). As the parameter $k_3/k_1$ is decreased, the stable modes 0.33 and 0.66 lose their stability and the unstable mode 0.50 becomes stable.

In other words, rather than saying that people adopt either the dynamics of (1) or (2) in any given utterance, we can say that both 1:2 and 1:3 modes of frequency locking are intrinsic to the system and are both present to some degree. Further, the normal behavior of the system is such that 1:2 and 1:3 frequency locking are competitive. We can thus conceptualize rhythmic events as biased toward the harmonics of a superordinate cycle, even if we do not know whether the 1/2 or 1/3 harmonic will dominate at any given time.

This model accounts for the observation that, in the slow-rate subset of data, the values of the modes of the phase distribution are close to harmonic ratios. The values are what they are because the dynamic system underlying the production of syllables biases them toward harmonic attractors of the phase space.

Furthermore, the model accounts for why the trimodality is absent in the fast-speech subset: a phase transition in the dynamics has occurred such that 1:3 coupling is no longer stable. The higher-order attractors are more prone to instability as cycle-rate increases (Haken, Kelso, & Bunz 1985). Conversely, slower rates allow the higher-order attractors to exert stronger influences. These ideas follow from the hypothesis that the control parameter $k_3/k_1$ interacts monotonically with speech-rate.

4.2. *The coupled-oscillators model of threepeats and the timing of continuation*

How do we evaluate the coupled-oscillators model? What sorts of testable predictions does it make? Let us begin to answer these questions by hypothesizing that the syllable following the threepeat (S4), which is usually the continuation of fluent speech, is governed by the same rhythmic system as the three-peat. In other words, knowledge of the phasing-mode with which $\varphi_2$ is associated might tell us something about the timing of subsequent syllables. Consider Fig. 20 below; our general hypothesis, which we wish to test, is that the particular mode R2 is associated with influences which mode S4 is associated with.

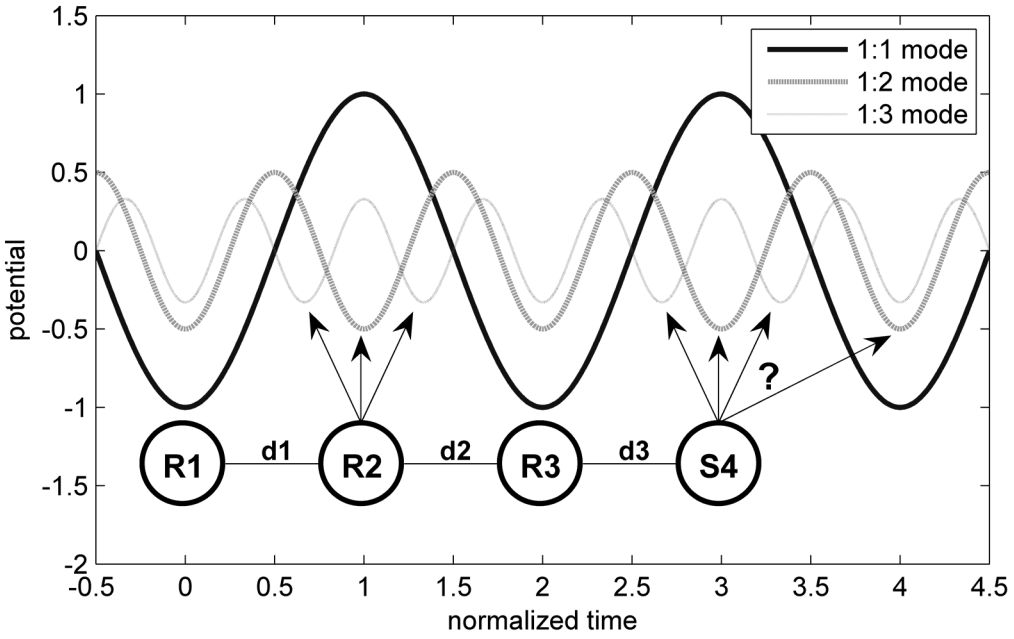


Fig. 20. *Schematization of hypothesized periodicities underlying $\varphi_2$ and S4 timing.* Valleys represent attractors in the potential functions. Time is expressed in normalized units. Mode-association of R2 is known, mode-association of S4 is predicted. d1 = pR1-pR2, i.e. the inter-p-center interval between R1 and R2, etc.

This hypothesis is motivated by the expectation that the coupled periodicities inferable from the threepeat are intrinsic to the speech-production system—thus we expect those periodicities to influence syllable-timing after the threepeat. More specifically, if R2 is associated with (or, closest to) the 1:2 mode, then S4 should occur closer to the corresponding 1:2 attractor 2 cycles later. Likewise, if R2 is associated with the 1:3 mode (i.e. is closest to phase 1/3 or 2/3), then S4 should be biased to occur closer to an attractor in the 1:3 mode. Because the presence of a W4 (unstressed syllable preceding S4) might interfere with S4 timing, we will consider S4 only in tokens with no W4 (n=233).

To test this hypothesis, we use the density parameter estimates from the trimodal fit of slow-rate $\varphi_2$ obtained in section 3.2.4 to separate the entire dataset into three groups; each group

consists of tokens whose $\varphi_2$ value is associated with one particular mode of the trimodal distribution. Then for each token we determine which of the set of predicted d3 durations is closest to the observed d3. The set of predicted d3 can be a function of either d1 or d1+d2—either way, these intervals tell us what the harmonic values of d3 are predicted to be given our assumption. We then count the best-fitting d3 values for all the tokens and perform a chi-square analysis on the count data (Table 4).

The chi-square analysis reveals that $\varphi_2$ mode does indeed interact with timing of S4 ($\chi^2 = 33.0$, $p < 0.000$). At face value, this is quite remarkable. Without the coupled-oscillators model there is no intuitively obvious reason why $\varphi_2$ association with 1:2 or 1:3 modes should interact with d3 duration. If we use the d1+d2 duration to predict d3 modes, the results are qualitatively unchanged and still significant ($\chi^2 = 19.3$, $p < 0.000$). Note that the differences are still significant when all S4 are considered ($\chi^2 = 8.6$, $p = 0.035$), but not when C4 (next syllable, stressed or not) are considered.

| | | **CONTINUATION (S4 w/o W4) TIMING** | | | | |
|---|---|---|---|---|---|---|
| | | **1/3 MODE** | **1/2 MODE** | **2/3 MODE** | **1/1 MODE** | **TOTAL** |
| **OBSERVED $\varphi_2$** | **1:2 MODE** | 111 (3.4) | 30 (4.1) | 28 (2.5) | 21 (1.4) | 199 (58%) |
| | **1:3 MODE** | 93 (6.3) | 2 (7.6) | 4 (4.7) | 4 (2.6) | 142 (42%) |
| **EXPECTED $\varphi_2$** | **1:2 MODE** | 132 | 21 | 4 | 4 | |
| | **1:3 MODE** | 72 | 11 | 11 | 9 | $\chi^2 = 33.0$<br>$p < 0.000$ |

Table 4. *Contingency table of S4 timing mode by $\varphi_2$-mode*. Observed and expected counts are shown. Numbers in parentheses indicate contributions of cells to chi-square value.

Upon closer inspection of Table 4, however, one can see that our predictions are not completely in accord with the data. The first thing we should note is that S4 is much more likely to be biased toward the 1/3 attractor than any of the other modes, regardless of $\varphi_2$ mode. This presumably reflects the observation that a greater proportion of R3 behave, from a functional perspective, as retraces (cf. section 1.5.2); because inter-beat-intervals become longer during threepeats, and because many R3 mark a return to fluent rates, d3 will tend to be shorter then d1.

We predicted that the continuation (S4) of a threepeat should tend to be closest to an attractor corresponding to the $\varphi_2$-mode. To a partial extent we can see that this is indeed the case: 1:2-mode $\varphi_2$ have fewer 1/3-biased d3 than expected, and 1:3-mode $\varphi_2$ have more 1/3-baised d3 than expected.

An unexpected finding is that 1:2-mode $\varphi_2$ occur with proportionately more 2/3-biased d3 than 1:3-mode $\varphi_2$—this is the opposite of what was predicted. Another somewhat perplexing pattern is that 1:2-mode $\varphi_2$ occur with relatively more 1/1-biased d3 than 1:3-mode $\varphi_2$. Since

both 1:2 and 1:3 timing modes are harmonics of 1:1 timing, neither should be associated with 1/1-biased d3 to a greater extent.

It is not clear exactly how to interpret these results. If there is one, the generalization to be made here is that 1:3-mode $\varphi_2$ biases S4 to be attracted to 1/3, while 1:2-mode $\varphi_2$ biases S4 to occur later. It is possible to view this effect as a rhythmic "echoing," such that 1:3 phasing of $\varphi_2$ is echoed in the timing of d3, regardless of whether $\varphi_2$ belongs to the 1/3 or 2/3 attractor.

### 4.3. *Discourse-functional and systemic views of repetition disfluency*

Now let us consider whether the discourse-functional and planning-systemic approaches to repetition disfluency described in sections 1.4 and 1.5 can account for the results presented above. These approaches underlie the two-fold subclassification of 2-cycle repetitions described in section 1.5.3. In this section we will see if we can plausibly interpret the harmonic trimodality of R2 phase in threepeats as arising from functional principles or systemic organization.

One way we might explain the observed patterns is by classifying R2 as stalling or retracing, or as hesitation-anticipating or hesitation-non-anticipating. For threepeats, the distinction between stalling and retracing certainly applies to R3, which may either be of normal or lengthened duration and may or may not be followed by a pause before the continuation. R2, however, does not seem likely to constitute a retrace. This is evidenced by its durational characteristics: R2 is normally the longest syllable in a threepeat. More abstractly, the continuation of a threepeat, by definition, never follows R2, so at most one might consider R2 to be "defective" retrace or part of a "retracing complex".

Likewise, the distinction between anticipating and non-anticipating R1 does not seem very well suited to R2 in threepeats, simply because the production of R2 by definition implies that hesitation has been anticipated. Thus classifying R2 seems like the wrong approach.

Another way we might explain the observed patterns is by considering how the classification of R1 and R3 in threepeats might interact with R2 phase. In order to evaluate this, we will classify R1 and R3 according to criteria and then compare the mean R2 phases across classes. To construct a subset of tokens that can reasonably be classified as hesitation-non-anticipating, the criteria of unlengthened R1 and no P1 (i.e. no preceding pause) can used in combination. For hesitation-anticipating tokens, either a lengthened R1 *or* a P1 should motivate inclusion in the subset (P1 is relatively infrequent and not a necessary condition for hesitation-anticipation). Likewise, the retracing subset can be defined by having both unlengthened R3 and no P4, and the stalling subset by having either lengthened R3 *or* a P4.

Table 5 shows that the subclassification does to some extent interact with the phase of R2. Hesitation-non-anticipation biases the phase of R2 earlier, but only when R3 is characteristic of stalling. Likewise, retracing biases the phase later, but only when R1 is characteristic of an anticipating cycle. In the other cases, the means are very close to the mean of the entire dataset (0.454).

| | CRITERIA | | | HESITATION-NON-ANTICIPATING | | | HESITATION-ANTICIPATING | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | R1 < MEDIAN | NO P1 | BOTH | R1 > MEDIAN | P1 | EITHER |
| | | N= | | 170 | 321 | 157 | 171 | 20 | 184 |
| | | | MEAN $\varphi_2$ | .443 | .455 | .441 | .467 | .462 | .467 |
| RETRACING | R3 < MEDIAN | 170 | .467 | | | | | | |
| | NO P4 | 315 | .455 | | | | | | |
| | BOTH | 169 | .467 | | | .451 (N=86) | | | .483 (N=83) |
| STALLING | R3 > MEDIAN | 171 | .443 | | | | | | |
| | P4 | 26 | .457 | | | | | | |
| | EITHER | 172 | .444 | | | .429 (N=71) | | | .454 (N=101) |

Table 5. *Comparison of mean R2 phase across hesitation-non-anticipating vs. hesitation-anticipating contexts and retracing vs. stalling contexts.* The "both" and "either" columns refer to the subset of data meeting both the R1 and P1 criteria or either of those criteria.

Despite these systematic interactions between R2 phase and the discourse and systemic subclassification systems and R2 phases, the sizes of the effects are not large enough to imply that specific modes of the R2 distribution are attributable to the 1st cycle or 2nd cycle distinctions. In other words, the effects are not categorical or even semi-categorical.

Furthermore, the functional and systemic perspectives—without additional stipulations—do not make any predictions about observing harmonic ratios of 0.33, 0.50, and 0.66 in our distribution. Even if such perspectives could predict three modes of categorical behavior, why these modes would approximate these values remains mysterious. It could be merely a coincidence that the modes observed correspond relatively well to these specific ratios. However, this "coincidental" explanation seems unattractive.

The observation that speech-rate interacts with the extent to which we observe three modes is also not amenable to analysis from speech-planning or discourse-functional perspectives; nor is our observation that R2 mode interacts with the timing of the continuation S4. Since we are unable to interpret these phenomena using traditional perspectives on repetition disfluency, the dynamical system model is a useful tool for understanding them.

*4.3. Speech-rate slowing in repetition.*

In order to account for the difference between the mean of observed $\varphi_2$ (.454) and the model-predicted harmonic phase .500, section 3.3 hypothesized that speech-rate normally slows during threepeats. In the monitoring-repairing perspectives on repetition disfluency, the cause of

speech-rate slowing, or "hesitation", is that the speaker suddenly becomes aware of a problem in their speech-plan. We showed earlier that correcting for this slowing by a linear transformation approximates the harmonic modes reasonably well. In actuality the slowing probably occurs more non-linearly and varies token-by-token, but a linear model suffices to produce acceptable correspondence between observed and predicted phasing modes.

This simple linear model of slowing can offer insight into the phonetic patterns of both 2-cycle and 3-cycle repetition. Let us assume that speech-rate is constant in the immediately preceding context, and that when hesitation is anticipated, speech-rate slowing and its effects (lengthening, pausing) become more likely. These assumptions follow from the distinction between hesitation-anticipating vs. hesitation-non-anticipating 1st cycles, proposed by Clark and Wasow (1998). Also, let us assume that the distinction made by Hieke (1981) between retracing and stalling 2nd cycles is such that retracing corresponds to a rate acceleration or return to normal/baseline rate, and stalling represents further deceleration in rate.

With these assumptions, a model incorporating local linear speech-rate slowing during repetition disfluency can make more coherent our understanding of the phonetic consequences of the subclassifications of 2-cycle repetitions and 3-cycle repetitions. Fig. 21 shows the four subclasses of 2-cycle repetitions, along with a representation of when and the extent to which hesitation has slowed speech-rate in each case (note that if this were a representation of speech-rate the curves would be inverted).

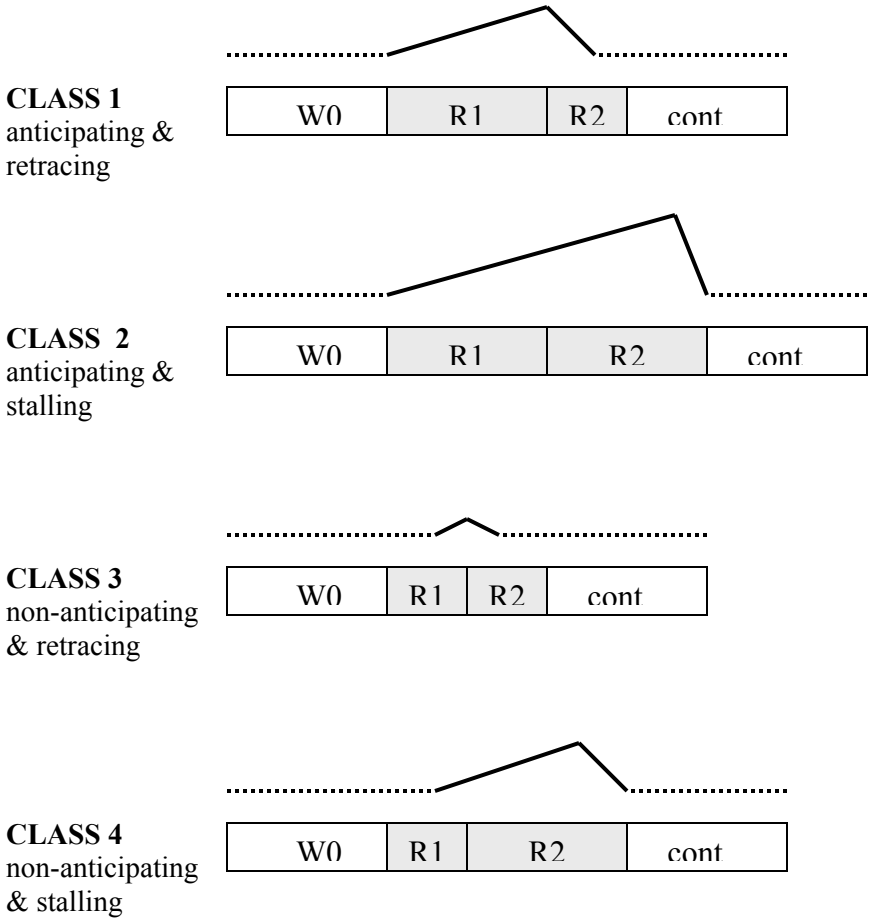


Fig. 21. *Speech-rate and the subclassification of 2-cycle repetitions.* Boxes represent syllable durations, lines represent the extent to which hesitation has slowed speech-rate.

In the hesitation-anticipating cases (classes 1 and 2), anticipation, or the beginning of slowing, precedes or coincides with R1. If the hesitation is anticipated very early, a pause before R1 is possible. In contrast, in hesitation-non-anticipating cases (classes 3 and 4), speech-slowing occurs towards the end of the production of R1, and a pause preceding R1 is unexpected. The idea here is that pausing before R1 and lengthening of R1 only occur when hesitation has to some extent been anticipated before R1.

Similarly, in the retracing cases (classes 1 and 3), speech-rate accelerates (i.e. the speaker returns to fluency, when planning problems no longer cause slowing) just before or near the beginning of R2. No pause following R2 is expected. In contrast, in the stalling cases (classes 2 and 4), speech-rate continues to decelerate until the continuation. A pause before the continuation may or may not occur, depending upon exactly when a repaired speech-plan becomes available.

If we assume that speech rate decelerates and accelerates once in the R1-R3 period, then we can make the prediction that there will be few examples like the one profiled in Fig. 22, where there is a temporary acceleration during R2 followed by a deceleration in R3. Is this prediction borne out in the data? Of the 341 tokens in the dataset, only 12 (3.5%) exhibit R2_P3

durations that are less than both R1_P2 and R3_P4 durations, which means that the deceleration-acceleration-deceleration pattern is indeed quite rare. Moreover, only 4 tokens whose pR1-pR3 durations are greater than the mean exhibit this profile, and in most of the 12 anomalous tokens the durational difference responsible for the acceleration during R2 is on the order of only 50ms, which presumably corresponds more closely to a constant rate rather than a changing one.

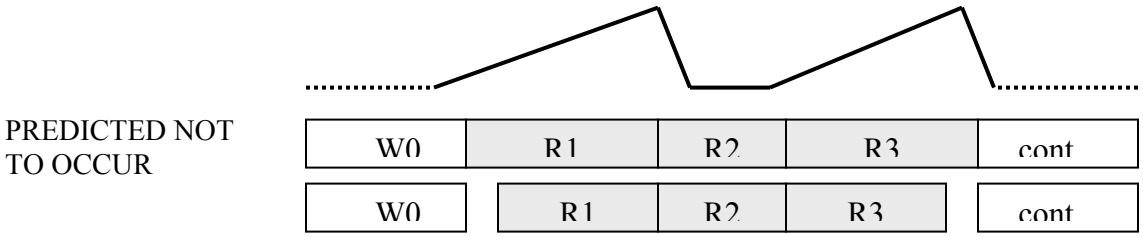


Fig. 22. *3-cycle repetition profile expected not to occur*. Boxes represent syllable durations, lines represent the extent to which hesitation has slowed speech-rate.

How can speech-rate slowing be integrated with the coupled oscillator model, so that the finding of harmonic trimodality and previously observed durational and pausing patterns can be understood in the same framework? One solution in the spirit of the dynamic systems approach would be to model speech-rate as a driving force on the oscillators. Without illustrating the details (still in development), the idea here is that a speech-rate driving force can accelerate or decelerate the oscillators that influence the timing of syllable production, thus relocating the attractors of the potential functions.

## 5. Conclusion

The results reported above provide evidence for rhythmic coordination of syllabic gestures in 3-cycle repetition disfluency. This raises the question of why such coordination is not readily observed in fluent speech. One possible explanation is that fluent speech does indeed employ the same sort of coordinative system, but that other systems such as stress and intonation, morphology and syntax, as wells as variability in syllable shape and lexical content, add aspects of temporal control that mask the underlying periodicities of the system.

One way to conceptualize this is to posit a distinction between motor-intrinsic rhythmic dynamics and motor-extrinsic rhythmic dynamics. The motor-intrinsic rhythms are described by the dynamics of the coupled-oscillator model, and apply to motor behaviors in general. These dynamics may arise from the multitudes of neurons in less-frontal motor areas and inter-connected subcortical structures like the basal ganglia and thalamus. Here the idea of self-organization comes into play: individual neurons exhibit oscillatory dynamics, and the oscillatory dynamics we observe on the scale of behavior arise from the interactions of those neurons. There is indeed neurological evidence that points to this understanding. The dynamics of electrical field potentials in certain brain regions can be correlated with behavioral dynamics. To wit, Kelso et.al. (1991, 1992) were able to correlate phase-shifts in the event-related field potentials with behavioral phase shifts from syncopation to synchronization in manual coordination.

In contrast, the motor-extrinsic dynamics are not very rhythmic and are not well-described by the dynamics of the coupled-oscillator model. Instead they may be better described by control-system speech-planning models, and they may arise from neurons in more frontal areas, especially the pre-frontal and pre-motor cortices, which are associated with the speech-planning "executive" and higher-level linguistic systems. In other words, syntactic and lexical linguistic systems introduce non-harmonic temporal dynamics to speech.

The evidence for coupled-oscillations presented herein also bears on the issue of whether harmonic timing arises from the perceptual system, the articulatory system, or both. Because there were no external periodic stimuli in this spontaneous conversational speech, it is unlikely that the phenomenon is perceptually-driven. In addition, the fact that the harmonic timing occurs on a different time-scale than the harmonic timing in speech-cycling tasks (450-1200 ms vs. 1000-2300 ms) argues for oscillatory dynamics of syllables in addition to feet and phrases.

Acknowledgements

Keith Johnson and Ian Maddieson for many helpful and insightful comments and conversations during various stages of this research; Elizabeth Shriberg for assistance in the use of the Switchboard corpus; Greg Hather and Cathy Tuglus for statistical consultation; and to Kim Tilsen.

[1] In the 1996 version of the task, speakers tried to repeat a short phrase (e.g. *take a pack of cards*) in time to a regularly repeated sequence of synthesized versions of the first and last stressed words in the phrase (e.g. *take . cards*). The period of the first synthesized word was fixed at 1.5 seconds, and the phase of the second (the target phase) was varied along eight values from .30 to .65 of the base 1.5s period. The stimulus was stopped after seven repetitions, and subjects continued repeating the phrase seven times, then paused for three seconds, and then produced seven more repetitions, trying to reproduce the relative timing of the stimulus. Note that the example phrase is more likely to be interpreted as a three-stress phrase than phrases of the form *X for a Y*.

[2] Note that "2-cycle repetition" should not be confused with "2 cycles *of* repetition," which implies 3-cycle repetition. The use of "cycle" in this context is novel. It makes sense to use this term because the model we are pursuing associates the initial form and subsequent repeats with modes of an oscillating system.

[3] This terminology is mine. Clark and Wasow (1998) used the terms "preliminary commitment" and "premature commitment"; I chose to employ the terms "hesitation-anticipating" and "hesitation-non-anticipating" because these more transparently convey the basis for the distinction.

[4] The use of the energy-rise centerpoint differs slightly from the procedure followed by Cummins and Port (1998). They used the energy-rise midpoint, which is the midpoint between points in time when signal amplitude reaches 10% of maximum and 90% of maximum. I chose to use the centerpoint because the midpoint seemed to skew p-centers too early for fricative and voiced onset consonants. In most other cases the two methods produce nearly identical results, the differences being around 5-30 ms, which is not enough to introduce procedural artifacts.

References

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

Clark, H. and Wasow, T. (1998). *Repeating Words in Spontaneous Speech*. Cognitive Psychology, 37, 201-242.

Cummins, F. and Port, R. (1996). *Rhythmic Constraints on English Stress Timing*. Unpublished.

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

DeJong, K. J. (1994). *The Correlation of P-center Adjustments with Articulatory and Acoustic Events*. Perception & Psychophysics, 56 (4), 447-460.

Dennis, J.E. Jr. and Schnabel, R.B. (1983). Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Englewood Clis, NJ: Prentice-Hall

Fowler, C. (1979). *Perceptual centers in speech production and perception*. Perception and Psychophysics. 25. 375-386.

Fowler, C. A., Smith, M. R., & Tassinary, L. G. (1986). *Perception of syllable timing by prebabbling infants*. Journal of the Acoustical Society of America, 79, 814–825.

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

Hieke A. E. (1981). *A Content-Processing View of Hesitation Phenomena*. Language and Speech, Vol. 24, Part 2, 1981.

Howell P. (1988). *Prediction of P-center location from the distribution of energy in the amplitude envelope*. Perceptual Psychophysics. Jan; 43(1):90-3.

Howell, P., & Au-Yeung, J. (2002). The EXPLAN theory of fluency control and the diagnosis of stuttering. In E. Fava (Ed.), *Pathology and therapy of speech disorders.* Amsterdam: John Benjamins.

Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T. 1991. C*ooperative and critical phenomena in the human brain revealed by multiple SQuIDs*. In: Duke D, Pritchard W, editors. Measuring chaos in the human brain. Teaneck, NJ: World Scientific. p 97–12.

Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T. 1992. *A phase transition in human brain and behavior*. Phys Lett A 169:134 –144.

Large, E.W. and Jones, M.R. (1999). *The Dynamics of Attending: How People Track Time-Varying Events*. Psychological Review.

Levelt, W. (1983). *Monitoring and self-repair in speech*. Cognition, 14, 41-104.

Liberman, M. (1975). The intonational system in English. MIT dissertation.

Liberman, M. and Prince, A. (1977). *On stress and Linguistic rhythm*. Linguistic Inquiry. 15:33-74.

Maclay, H., and Osgood, C. E. (1959). *Hesitation phenomena in spontaneous English speech*. Word 15: 19–44.

Morton, J., Marcus, S., and Frankish, C. (1976). *Perceptual Centers (P-centers)*. Psychological Review. Vol. 83, No. 5, 405-408.

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

Patel, A.D., Löfqvist, A. and Naito, W. (1999). *The Acoustics and Kinematics of Regularly Timed Speech: A Database and Method for the Study of the P-Center Problem.* Proceedings of the 14th International Congress of Phonetic Sciences, August 1999, San Francisco.

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

Plauché, M. and Shriberg, E. (1999). *Data driven subclassification of disfluent repetitions based on prosodic features*. In Proceedings of the International Congress of Phonetic Sciences, San Francisco.

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

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

Port, R.F., Dalby, J. and O'Dell, M. (1987). Evidence for Mora timing in Japanese. Journal of the Acoustic Society of America. May;81(5):1574-85.

Postma, A., and Kolk, H. (1993). *The Covert Repair Hypothesis: Prearticulatory Repair Processes in Normal and Stuttered Disfluencies*, Journal of Speech and Hearing Research, Vol. 36, 472-487.

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

Saltzman, E. L. & Kelso, J.A.S. (1987). *Skilled Actions: A Task-Dynamic Approach*. Psychological Review, 94. 84-106.

Schnabel, R.B., Koonatz, J.E., and Weiss, B.E. (1985). *A modular system of algorithms for unconstrained minimization*. ACM Transactions on Mathematical Software (TOMS).

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

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

Semjen, A., and Ivry, R. (2001). *The Coupled Oscillator Model of Between-Hand Coordination in Alternate-Hand Tapping: A Reappraisal*. Journal of Experimental Psychology: Human Perception and Performance. 27:2. 251-265.

Shriberg, E. (1999). *Phonetic Consequences of Speech Disfluency*. In *Proceedings of the XIVth International Congress on Phonetic Sciences* (pp. 619–622). San Francisco.

Shriberg, E. (2001). *To 'errrr' is human: ecology and acoustics of speech disfluencies.* Journal of the International Phonetic Association. 31/1.