The 'Harmonic Timing Effect' in speech

Robert F. Port, Fred Cummins and Keiichi Tajima*
Department of Linguistics,
Department of Computer Science
Program in Cognitive Science
Indiana University, Bloomington, Indiana 47405

August 9, 1996

Abstract

We have discovered that if English speakers are asked to repeat a phrase or short text containing two or more stressed syllables over and over, they tend to adjust the timing such that important syllable onsets (eg, stressed syllables) occur at times that are harmonic fractions of each other – that is, the inter-event time intervals have a strong tendency to exhibit simple ratios close to 1:1, 1:2, 1:3, etc. This behavior is robust and easily obtained although there appear to be major differences between languages in how this timing is arranged. We demonstrate the basic effect in English and in Japanese and explore a little about the conditions under which the effect occurs in English. The harmonic timing effect is so strong, it may prove to be manipulable in ways that will be useful for research on issues like phonetic time measurement, the study of stress, typological differences between languages and possibly other matters.

*TENTATIVE ORDERING OF AUTHORS: The authors are grateful to whoever. This research was supported in part by the Office of Naval Research, N0003-1267.

1 INTRODUCTION

1.1 Repetition induces periodicity

Asking subjects to repeat a phrase over and over has been employed many times in the phonetics literature, at least since Stetson's work beginning in the 1930s (eg, Stetson, Vitikiotis-etal, Marcus, Dauer, Kelso-etal,). In addition, human communities create speaking genres that involve repetitions of short text, for prayers (eg, *'Hail Mary, full of grace...*) or punishment (eg, *I will not chew gum in school, ...*). As far as we are aware, one robust property about these phrase repetition tasks has not been explicitly commented upon or, in any case, its importance has not been pointed out. This property is that some syllables within a phrase (and possibly all the syllables) tend to be distributed in time so as to exhibit harmonic ratios of durations.

We have found that if speakers (of English at least) are asked to simply REPEAT a phrase longer that a couple syllables, they apparently cannot help but adjust the temporal details of their speech in such a way that the most prominent articulatory (and acoustic) events – for English speakers, the onsets of stressed syllables – occur at harmonic fractions of each other.

To observe the effect in simple form, say a short phrase containing at least two stressed syllables repeatedly, eg, *"to beat about the BUSH, to beat about the BUSH, ..."* etc. Note first that the initial word *to* is treated as a kind of throwaway upbeat to the first stressed word, *beat*. Note especially where the onsets of *bush* occur relative to the onsets of *beat*: The word *bush* is likely to occur *half way* between the onsets of the word *beat* (and vice versam, of course). You can verify this (if you are a good speaker of English) by tapping your left hand quietly for each production of the word *beat* and your right hand on *bush*. You will probably find the two hands tapping at the same rate but exactly out of phase with each other. Alternatively, you could tap your right hand twice as fast as the left – tapping, that is, on both *beat* and *bush*. Because *bush* is located halfway between the productions of *beat*, we will say that the second stressed word occurs at a harmonic fraction of the phrase repetition (which begins with the first stressed word *beat*. (An alternative style of repeated production of this phrase, puts the primary stress on *beat* and then locates *bush* about two-thirds of the way between onsets of *beat* – a 'waltz-time reading'.)

None of us is surprised, of course, that people sometimes speak rhythmically. What we seek to point out by the Harmonic Timing Effect is that mere repetitions of phrases – with or without a stabilizing metronome – leads to periodic cycling with very strong phrase-internal timing contraints – just as if the text were treated as a line of poetry or song merely by virtue of being repeated.

After all, the task of merely repeating a phrase would not necessarily imply any constraints on the realtime structure of the speech. Certainly the traditional linguistic descriptions (Chomsky-Halle, 68, Anderson, 82) emphasize that language exhibits merely serially ordered units. As noted by Mark Liberman (1978), predictions regarding equal or harmonic spacing of events could not be derived from, for example, Chomsky and Halle's theory of phonology. Yet these constraints seem intuitively very strong indeed.

Of course, Liberman (1978) attempted to remedy the lack of any notion of phonological time other than serial order. He sought to account for apparent similarities between musical rhythm and linguistic rhythm by propounding the notion of a 'Metrical Grid', a hierarchical, "intuited structuring of time" that specifies time points related by harmonic time intervals, presumably in both duple and triple hierarchical patterns. Lerdahl and Jackendoff (198n) employed the same notion to account for the universality of musical metrical patterns. The Metrical Grid was proposed as something apparently innate and universal that provides a kind of apriori template for scheduling events – a structure containing possible points in real time for phonetic events to be located or for musical events to begin or end. For Liberman, the Metrical Grid served as a complement to his proposed Metrical Patterns, the abstract tree structures of Strong and Weak units. There should be one of these binary, S or W units for each element of linguistic material. However Metrical Patterns require only serial order for their specification. Thus every linguistic utterance has a particular Metrical Pattern that must be aligned with an appropriate Metrical Grid during speech production. In this way Liberman supplements the serial order model of linguistics with a discrete structure in continuous time. This will account for why speech often exhibits music-like temporal constraints. Of course, such an apriori intuitive structure could be invoked quite naturally to account for the phenomena of harmonic timing in repeated phrases that we address here. However, while admiring the boldness of this postulatory stroke, one must continue to wonder what kind of mechanical system could actually generate such a grid as a stable object when needed to account for the rhythmic behavior in music, speech, quasi-musical chants and even dance (see McAuley-95, Large-Kolen-96, Gasser, Deck and Port-inpress).

So, this paper will focus just on experimental evidence that shows just how powerful the harmonic timing effect is, and suggest a couple possible uses that might be made of this effect for research on speech and language. Only later work will address the question of why these constraints exist in the first place. First we will demonstrate the effect, then discuss some possible implications for phonetic and linguistic research.

1.2 Phrase Repetition Task: Entraining speech to itself

In developing a task that will demonstrate harmonic timing in human speakers, the first step is to see the logical necessity for 'beats' – for particular points in time that (a) recur at predictable times, and (b) have some definable property that differentiates them (such that predictions can be tested and either verified or not). Such thinking leads, in the simplest case, to a system of 'repeating beats'.

Our tentative proposal for the nature of beats during speech is for something very similar to the so-called 'P-center' (Marcus-75, Dauer-83). The most important, or beatlike, parts of words are the onsets of vowels – especially stressed or accented vowels. Such events, the onset of strong acoustic energy in the F1 and F2 region, are auditorily important because many fibers in the auditory nerve are known to respond strongly as such onsets (whoever-MIT, Sachs-Young). Thus, we developed a simple 'beat detector that integrates acoustic energy over the region between 200 Hz and 1800Hz, smoothes it and looks for positive-to-negative zero crossings. A pulse is generated at this point whose amplitude is proportional to the peak amplitude of the following vowel. In this way, speech can be automatically reduced to a series of soft or loud pulses whose temporal location can be recorded. This may seem, at first glance, an extreme reduction to make to the speech signal. Our view is that something like this must underlie any 'meter', since periodicity depends on something recurring with regularity. For a signal as complex and variable as speech, only something very abstract and, yet, frequently occurring could support periodicity detection. These are the time points that you would tap your pencil to. We certainly do not claim that our particular method for finding beats is somehow the correct one, nor that there must be some unique method for this purpose. We note merely that this method seems to give reasonable results. We find that an auditory metronome helps stabilize subject performance.

Thus, given a method for extracting beats or pulses, that is, 'special time points,' from continuous speech, we can now just define a task in which subjects repeat a short phrase. [etc]

2 Experiment 1: English and Japanese speakers

In the first experiment here, both English and Japanese speakers repeated short phrases with varying numbers of syllables. The results suggest that the Japanese speakers tend to divide the period of the phrases into roughly equal sized segments, whereas the English speakers, use only the stressed syllables for this purpose.

2.1 Method of beat extraction.

As explained, we need a general method to provide a definition of the beat of a syllable. This should be in approximate correspondence with the time point in the syllable which a subject would use for aligning one syllable with respect to another stimulus, and thus be close to the syllable's "P-center". Based in part on Scott's model for P-center location (Scott, 1993), and Marcus' original model (Marcus, 1981), we used the following algorithmic definition of a syllable's acoustic 'beat:'

After recording the speech at 8000 samples per second directly onto disk, it was passed through a simple auditory filtering model using the Lutear software package. After initial preemphasis, the signal is passed through a bank of six gammatone filters with center frequencies in the range 300 to 2000 Hz. This range should ensure that most of the energy from the first two formants is preserved in the filter output, while F0 and high frequency frication are largely filtered out. The six resulting filter outputs are smoothed and summed to yield an estimate of total signal energy in this range. The signal is rectified, and again smoothed using a simple lowpass filter. This gives a smooth amplitude contour which might be interpreted as a continuous measure of 'sonority.' Any rise in this amplitude is associated with an acoustic beat. The moment of occurrence of the beat is defined as the time point halfway between the local maximum and the preceding local minimum. This beat is based purely on a frequency-dependent amplitude increase in the acoustic signal. With our algorithm, beats are usually located very near the vowel onset of a syllable.

2.2 [The Expt Keiichi is working on.]

Conclusions for Experiment 1. This experiment has shown that for English speakers, when repeating a phrase of two feet, the stressed syllables tend to be located equidistant from each other. In metrical terms, one might say that the two feet tend to be equal in duration. It was not clear that the there any strong constraints on most other syllables. For Japanese speakers of English, on the other hand, a rather different pattern of timing was observed in which each syllable tended to be the same as each other. Of course, precisely these metrical observations have been claimed many times in the past for English and Japanese, but experimental data has always been tenuous. However with the Phrase Repetition Task, these observations appear very distinctly and the language-specific differences appear to be shown especially vividly.

If certain phase angles are preferred by English speakers for the onset of stressed syllables, how can we show most clearly that these these locations are 'attractive'? Is one half the only attractive phase angle? We conceived of an experiment in which asked speakers to place the second stressed syllable of a phrase at a sampling of phase angles from 0.3 to 0.65. Using the phrase '*Take a pack of cards*, we used synthetic speech versions of *take* and *cards* and adjusted their timing.

3 Experiment 2: Aiming for phase targets

3.1 Methods.

Stimuli were constructed consisting of eight repetitions of the words *take* and *cards*. The stimuli were generated with a flat intonation contour using commercial *Eloquence* speech synthesis software. The interval between the beats of successive tokens of *take* was fixed at 1.5 seconds. The phase angle at which the beat of *cards* occurred was varied across trials. If a cycle is the interval between two successive tokens of *take*, the phase of the beat of *cards* was fixed for each trial at a value we will refer to as Φ_k (we use a phase convention specifying phase in the range 0–1). Eight values of Φ_k were used, ranging from 0.3 to 0.65 in increments of 0.05.

On each trial, subjects were presented with the stimulus over headphones. Their task was to try to repeat the phrase *take a pack of cards* in time with the stimulus, such that their productions of *take* and *cards* were temporally aligned with their respective stimulus. No instructions about the relative timing of the other words of the phrase were given. They were given some practice to familiarize themselves with the task before data were collected.

In each trial, they were asked to begin talking on the second repetition of the stimulus (i.e. the second time they heard the word *take*), and to continue speaking after the stimulus stopped. They then continued for seven repetitions of the phrase without the accompanying stimulus. Then they were signaled to pause for about three seconds, and then to produce seven more repetitions, again trying to maintain the relative timing of *take* and *cards* given previously by the stimulus. Subjects were not required to count the number of repetitions they produced, as they received visual signals from the experimenter indicating when they should pause, and when they should stop.

The per-trial data thus fell into three groups—tokens produced together with the stimulus timing pattern, immediately after stimulus cessation, and again after a short pause. There were seven tokens in each group, giving 21 tokens per trial. The eight target phases were presented in three conditions: one in which the target value of Φ_k increased from 0.3 to 0.65 across trials, one in which it decreased, and one in which the target values of Φ_k were randomized. Subjects took a short break between conditions to avoid fatigue. A total of six subjects were run, allowing the order of the conditions to be counterbalanced across subjects. All subjects were graduate students outside of the field of linguistics, and were naive to the purpose of the experiment.

The target values of phase were presented in three orders: ascending phase targets, descending phase targets and randomly ordered phase targets.

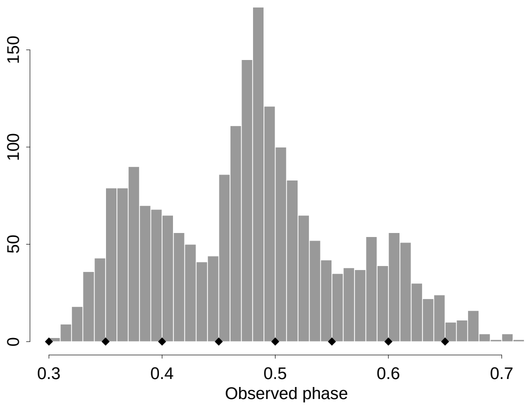

Figure 1: Histogram of all observed phases of the medial stressed syllable *cards* in repetitions of the phrase *Take a pack of cards*. Target phases are marked in black on the abscissa. Although there were 8 target phases, the overall distribution is clearly trimodal.

Measurement procedure. From each group of seven productions, the first was discarded because all subjects showed some uncertainty, both in starting the repetitions, and immediately after the cessation of the stimulus. The value of Φ_k for the next five productions was measured, using the above beat extraction process. The phase at which the beat of *cards* occurred was measured relative to the preceding and following beats associated with the subject's productions of *take*. No measurement was possible for the last production of the latter two within-trial groups as they lacked a following *take* as a referent. For consistency, then, we obtained five phase measurements per group, with three groups per trial. As there were eight target phases per condition and three conditions within six subjects, a total of $3*8*3*5*6 = 2160$ data points were obtained. [[How many per phase angle target? Thats more useful.]]

3.2 Results

The main results pooled across speakers, repetition sets and trial blocks are shown in Figure 1 as a frequency histogram of the measured phase angle of *cards*. The main finding is that subjects could not produce all target phases equally well. Although the target phase angles for the medial stress were equally probable at eight different phase angles, the produced phase angles show a well-defined cluster around 0.5, and somewhat weaker clusters centered around 0.36 and 0.6. These values are close to $\frac{1}{3}$ and $\frac{2}{3}$, predicted by a simple harmonic model for stress location (although the consistent deviation away from actual harmonic predictions merits further attention).

This result seems a rather clear demonstration of the harmonic timing effect. Despite instructions to place the onset of the word *card* at each of these phase angles, subjects instead put many of them at 0.5 and others at certain additional preferred phase angles.

A second important finding, one that is of interest for certain explanations for the effect, is that the produced phase was influenced by the order of presentation of the target phases. That is, apparently the phase produced in the immediately preceding trial influences what subjects do on the current trial, as shown in Figure 2. When phase was *increased* from 0.3 to 0.65 across trials, smaller average phase values were produced than when phase *decreased* across trials. Random ordering of the targets yielded intermediate phase values. Note that this figure shows mean values for what are essentially multimodal distributions (see Figure 1), so that all means are biased towards the strongest mode at 0.5. The effect of recent targets on the distribution of observed phases is potentially important information for inferring the symmetry properties of the underlying dynamic.

Although we expected subjects to lose accuracy when the stimulus demonstrating the target phase was withdrawn, in fact subjects' attempts to reproduce the target phase were no less accurate immediately after its cessation nor after the three second delay than they were when the target phase was demonstrated. A within-subject Anova was performed comparing the effects of target (8 levels), target presentation order (3 levels: ascending, descending and random) and repetition group (3 levels: with stimulus, immediately after stimulus and after the pause) on the mean difference between observed phase and target phase. The only main effect which was significant was target $[F(7,35) = 23.126, p < 0.001]$. The main effect of repetition group, which compares accuracy with and without stimulus, was not significant $[F(2,10) = 0.410, p > 0.4]$. The only other significant effect was the order by target interaction $[F(14,70) = 1.894, p < 0.05]$, which reflects the dependency of observed phase on previous trials.

The data shown in Figures 1 and 2 are averaged across subjects, which ob-

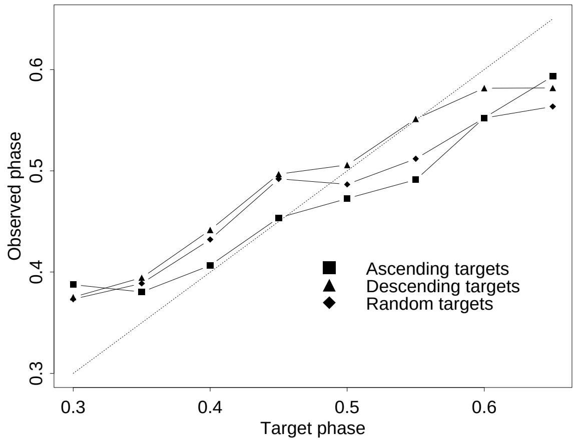

Figure 2: Observed phase of the vowel onset for *cards* as a function of target phase when target phases are increasing, decreasing and randomized across trials. The line $y = x$ is also included for reference. The sequence of trials in which phase increases from trial to trial produces smaller mean phase values that with decreasing phase.

scures considerable intersubject variability. The bias for harmonic fractions can sometimes be seen more clearly when individual speakers are examined. For example, Figure 3 shows some results for a single subject. Each data point shows the mean observed phase and standard deviation for repetitions within a single trial. As in the previous figure, trials are grouped by block (targets are either ascending or descending across trials). The random target presentation condition is not shown. For most target values, the ascending function again lies to the right of the descending function, i.e. smaller mean phases are produced. A clear example of bias for targets at harmonic fractions can be seen in the ascending curve for targets at 0.3, 0.35 and 0.4. All were imitated with the same output phase at about 0.35. Then targets of 0.45, 0.5 and 0.55 were all produced very close to 0.5. As the target changed from 0.4 to 0.45, the subject's productions jumped rather sharply from $\frac{1}{3}$ to $\frac{1}{2}$. It appears that the actual phases produced by subjects are not only dependent on the combination of target phase and nearby harmonic attractors. They are also influenced by context (i.e. productions and/or targets on previous trials).

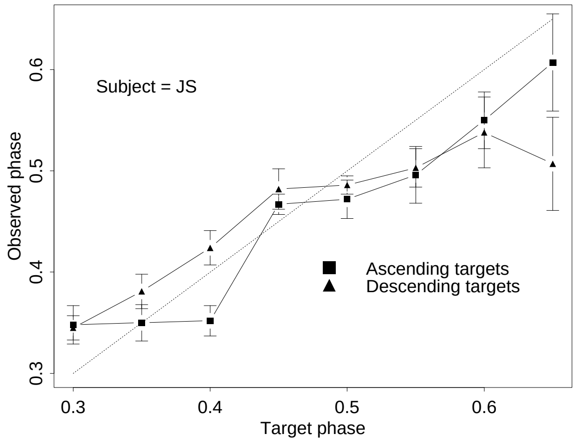

Figure 3: Sample data for one subject showing the mean and standard deviation for each trial. The two conditions plotted are the phase increasing block and the phase decreasing block.

Conclusions for Experiment 2. This experiment shows that even when subjects are specifically asked to begin the second foot a particular phase angles, they cannot. Instead they show a strong preference for a phase of 1/2 and perhaps also for 1/3 and 2/3. The hysteresis effect, the differences in behavior due to the ordering of gestures, implicates a dynamical system.

4 Discussion

4.1 Typology: stress-timed vs. syllable-timed vs. what?

Our little bit of crosslanguage data may be suggestive. By appropriate choice of text materials and task refinement, it might be possible to explore a great many subtle differences in timing between languages, foreign accents and even dialects.

The classification of languages as stress- or syllable-timed has an anachronistic air today, as little or no empirical evidence of isochrony has been found (Classé, 1939; Pike, 1945; Lehiste, 1977; Dauer, 1983). In a 'stress-timed' language such as English, interstress intervals are found to vary greatly and, indeed, to exhibit as much or more variance as similar intervals in a 'syllable-timed' language such as French (Roach, 1982; Williams and Hiller, 1994). Numerous researchers have been led to the conclusion that isochrony (or rhythmicity) in speech should be considered merely a perceptual distortion due to gestalt regularization principles (Lehiste, 1977; Scott et al., 1985; Benguerel and D'Arcy, 1986).

Nonetheless, there is evidence for the claim that interstress intervals in English are governed in part by global constraints. For example, Williams and Hiller (Williams and Hiller, 1994) found that the "foot compression effect" and a foot-dependent syllable shortening effect were nonrandom property of metrical feet, as opposed to random groups of syllables. They conclude that "it is indeed worth attempting to measure rhythmicality, as this has been shown not to be random" (p.438).

This experiment has clearly demonstrated that, given an appropriately constrained task, global timing constraints on stress placement are directly observable from acoustic measurements. Thus Benguerel and d'Arcy's claim that regularity of stress timing is not observable in the acoustic signal may be rejected. Of more interest, however, is the relationship of this regularity to rhythm. Are these global constraints on stress timing a direct manifestation of *rhythm* in speech production?

The observed constraints share many features with rhythmic patterns measured during cyclic activity of the limbs in walking, tapping and waving (Cummins and Port, 1996). Among the characteristics of rhythmic organization are entrainment among constituents, as evidenced by stable, simple phase relations (Kelso, 1995). In our speech data, we found strong preferences for the medial syllable to be placed at roughly $\frac{1}{3}$, $\frac{1}{2}$ or $\frac{2}{3}$ of the overall repetition cycle. This is consistent with the expected behavior of two coupled oscillatory systems, one coextensive with the metrical foot and one with the repetition cycle.

One of the most striking features of these data is the evidence of stability over time. Once subjects pick a phase angle for timing, the can hold it quite reliably well after the stimulus has disappeared. It has often been suggested that one way to simplify the problem of coordinating a complex, high degree of freedom system such as the motor system is to functionally link the action components, so that their mutual coordination is assured and the number of degrees of freedom of the control problem is reduced (Bernstein, 1967; Kugler et al., 1980; Kugler and Turvey, 1987). We interpret the observed linkage between metrical foot and phrase repetition cycles as just such a strategy.

[How exactly? Not clear yet.] In this light, it is not surprising that the linkage is manifested regardless of the externally imposed constraint of the stimulus.

The tendency of produced phases to closely resemble those of the preceding trials is further evidence of an underlying system of coupled oscillatory processes. Both empirical observation of coupled processes (Buchanan and Kelso, 1993; Diedrich and Warren, 1995) and mathematical models of coupled oscillators (Thompson and Stewart, 1986) show that systems which have found a stable coupling relationship will persevere in that coupling mode in the face of external perturbation (here, the changing target phase). This hysteresis is a common property of multistable dynamic systems and is largely independent of the details of the underlying system components.

The experimental evidence presented here constitutes evidence that stress timing in English speech can be easily constrained by the rhythmic coordination of cyclic processes. Given suitable elicitation procedures such as the phrase repetition task this constraint is directly observable in the acoustic signal. An understanding of the global constraints which influence speech timing is an important first step towards a characterization of 'natural' timing patterns of speech.

4.2 Conclusions

First, it appears the the Harmonic Timing Effect is prominent and easily observed in speech under simple experimental conditions.

Second, there is reason to believe that languages may differ in the details of how this effect is executed thereby opening the possibility that diagnostic or training materials might be developed that exploit this effect.

Third, this sort behavior – what we might call 'metrical grid behavior' – seems to underlie not just speech, but also very likely hand motions, dance etc. Thus the effect offers the possibility of clarifying how speech resembles other complex motor skills.

Finally, these results implicate dynamical models of motor control and general cognition (eg, Port-van Gelder, Kelso-95).

References

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

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

Buchanan, J. and Kelso, J. (1993). Posturally induced transitions in rhythmic multijoint limb movements. 94:131–142.

Classé, A. (1939). *The Rhythm of English Prose*. Basil Blackwell, Oxford.

Cummins, F. and Port, R. F. (1996). Rhythmic commonalities between hand gestures and speech. In *Proceedings of the Eighteenth Meeting of the Cognitive Science Society*. To appear.

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

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

Kelso, S. (1995). *Dynamic Patterns: The self-organization of brain and behavior*. MIT Press, Cambridge, MA.

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

Kugler, P. N. and Turvey, M. (1987). *Information, Natural Law, and Self-Assembly of Rhythmic Movement*. Erlbaum, Hillsdale, N.J.

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

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

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

Roach, P. (1982). On the distinction between "stress-timed" and "syllable-timed" languages. In Crystal, D., editor, *Linguistic Controversies*. Edward Arnold, London.

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

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

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

Williams, B. and Hiller, S. M. (1994). The question of randomness in English foot timing: a control experiment. *Journal of Phonetics*, 22:423–439.