Rhythmic Constraints on English Stress Timing

Fred Cummins and Robert F. Port

Department of Linguistics and Cognitive Science Program
Indiana University, Bloomington, IN 47405

ABSTRACT

The evidence for isochrony of stress timing is weak at best for ordinary prose, but this does not mean that the timing of stresses is always unaffected by global constraints. We asked subjects to continually repeat the phrase *Take a pack of cards* and to temporally align the words *take* and *cards* with an auditorily presented stimulus consisting of just the words *take* and *cards* repeated several times. The phase of the *cards* stimulus relative to a reference cycle defined by the *take–take* interval was varied over the range 0.3-0.65 in eight equal-sized phase steps. The distribution of actually produced phases for the vowel onset of the syllable *cards*, however, was strongly trimodal. Subjects showed a powerful preference for phases close to 0.5, and somewhat weaker preferences for phases near 0.36 and 0.6. These values are close to (although systematically different from) 0.33 and 0.66 predicted by a simple harmonic model for stress timing. The observed distribution had this form whether the subjects were speaking along with the stimulus, or trying to maintain the prescribed timing after cessation of the stimulus. Furthermore, the observed phase was influenced by the phase produced on previous trials, suggesting dynamic control with hysteresis between competing stable patterns of timing. These results demonstrate strong rhythmic constraints on the timing of stresses within a phrase, where the domain of 'phrase' in this artificial speaking task is simply the repeated text. The rhythmic constraints are similar to those observed for limb movements. Modeling these constraints should provide insight into the form of a general dynamic control regime for global speech timing, and may allow improved characterization of 'natural' timing patterns in English speech.

1. INTRODUCTION

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 [4, 13, 11, 6]. 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 [14, 18]. Numerous researchers have been led to the conclusion that isochrony (or rhythmicity) in speech should be considered a perceptual phenomenon, derived from the known bias of perceptual systems to regularize in the face of noise [11, 15, 1]. This viewpoint is forcefully expressed by Benguerel and d'Arcy [1], who claim:

> It seems surprising that anyone interested in finding evidence for *perceptual* regularity would expect measurements (no matter how careful) of the *acoustic* signal to provide such evidence . . . In our view, if there exists any regularity or rhythmicity in speech, it is at the perceptual level, and possibly at the pre-production level, but not at the acoustic level. (p.244)

Nonetheless, there is evidence for the claim that interstress intervals in English are governed in part by global constraints. For example, Williams and Hiller [18] 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:

> . . . it can now be stated that it is indeed worth attempting to measure rhythmicality, as this has been shown not to be random. (p.438)

One way to investigate the nature of these global constraints is to devise a speech task which requires subjects to attempt to produce a range of interstress intervals. In order to reduce as far as possible the many uncontrolled variables associated with continuous speech, we devised a phrase repetition task where subjects repeated a simple phrase many times. The simplification of suprasegmental features which results is somewhat akin to reiterant speech, in which segmental variation within the syllable is controlled. By thus defining a phrase repetition cycle, we were able to specify precisely the point in time relative to this cycle at which subjects would attempt to place a medial stress.

2. METHODS

For both stimuli and human data in the following experiment, we needed to provide a definition of the 'moment of occurrence,' or 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. The optimal definition of a beat is an empirical issue. In the context of this work, the beat's function is to provide a localized event associated with a syllable, which is treated as the 'time of occurrence' of the syllable. Based in part on Scott's

model for P-center location [16], and Marcus' original model [12], 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, and, strictly speaking, should be distinguished from the term P-center. With our algorithm, beats are usually located very near the vowel onset of a syllable.

Stimuli were constructed consisting of eight repetitions of the words *take* and *cards*, produced consecutively. The stimuli were generated with a flat intonational contour using commercial speech synthesis software (Eloquence). The interval between the beats of successive tokens of *take* was fixed at 1.5 seconds. The point in time at which the beat of *cards* occurred was varied across trials. Defining a cycle as 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 the stimulus. No instructions about the relative timing of the other words of the phrase were given. 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. When they had completed seven repetitions of the phrase without accompanying stimulus, 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, immediately after the stimulus, and 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 naïve to the purpose of the experiment.

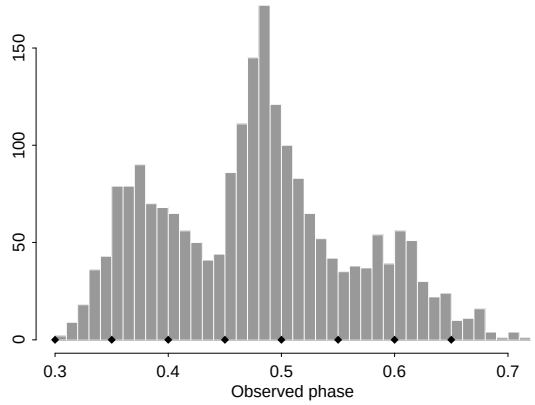

Figure 1: Histogram of all observed phases of the medial stressed syllable *cards* in repetitions of the phrase *Take a pack of cards*. Target phases are marked in black on the abscissa. Although there were 8 target phases, the overall distribution is clearly trimodal.

2.1. 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.

3. 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 first main finding was 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).

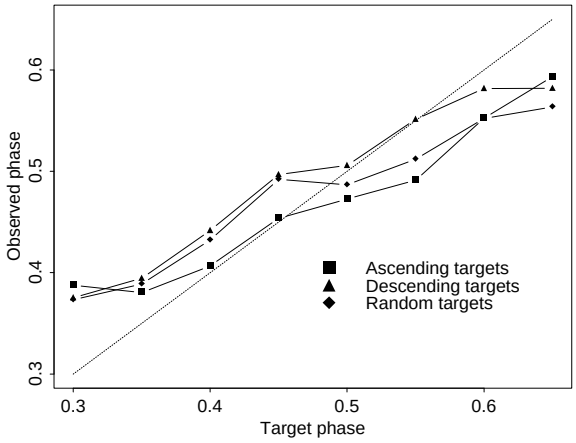

Figure 2: Observed phase of the vowel onset for *cards* as a function of target phase when target phases are increasing, decreasing and randomized across trials. The line $y = x$ is also included for reference. The sequence of trials in which phase increases from trial to trial produces smaller mean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

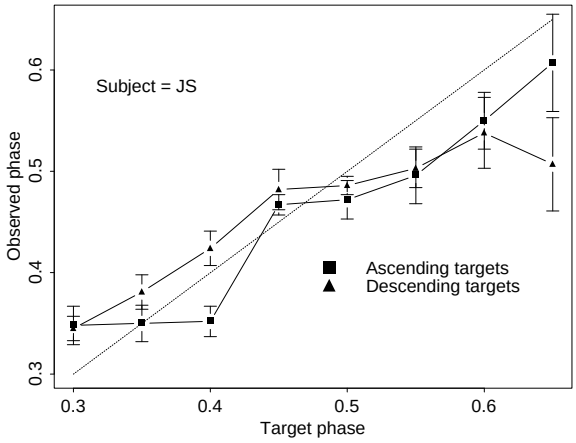

Figure 3: Sample data for one subject showing the mean and standard deviation for each trial. The two conditions plotted are the phase increasing block and the phase decreasing block.

A second important finding is that the produced phase was influenced by that produced on immediately preceding trials, as shown in Figure 2. When phase increased from 0.3 to 0.65 across trials, smaller average phase values were produced than when phase decreased across trials. Random target ordering 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.

Contrary to our expectations, subjects' attempts to reproduce the target phase were no more accurate when speaking simultaneously with the prompt than they were immediately after its cessation or after the three second delay.

These findings were substantiated by quantitative statistical analysis. 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 obscures considerable intersubject variability. The bias for harmonic fractions can be seen more clearly when individual speakers are examined. By way of 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 discretely 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).

4. Discussion

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?

We have argued elsewhere that the observed constraints share many features with rhythmic patterns measured during cyclic activity of the limbs in walking, tapping and waving [5]. Among the characteristics of rhythmic organization are entrainment among constituents, as evidenced by stable, simple phase relations [8]. 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.

Further evidence of a stable organization, characterized by coupling between its components, comes from the rather surprising finding that subjects were no more accurate in producing target phases when speaking with the stimulus than they were after it was turned off, or even after a further three second pause. 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 action components, so that their mutual coordination is assured and the number of degrees of freedom of the control problem is reduced [2, 9, 10]. We interpret the observed linkage between metrical foot and phrase repetition cycles as just such a strategy. 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 [3, 7] and mathematical models of coupled oscillators [17] 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). 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 strong evidence that stress timing in English speech is 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.

5. REFERENCES

1. André-Pierre Benguerel and Janet D'Arcy. Time-warping and the perception of rhythm in speech. *Journal of Phonetics*, 14:231–246, 1986.
2. N. Bernstein. *The Coordination and Regulation of Movements*. Pergamon Press, London, 1967.
3. J.J. Buchanan and J.A.S. Kelso. Posturally induced transitions in rhythmic multijoint limb movements. *Experimental Brain Research*, 94:131–142, 1993.
4. André Classé. *The Rhythm of English Prose*. Basil Blackwell, Oxford, England, 1939.
5. Fred Cummins and Robert F. Port. Rhythmic commonalities between hand gestures and speech. In *Proceedings of the Eighteenth Meeting of the Cognitive Science Society*. To appear, 1996.
6. R.M. Dauer. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11:51–62, 1983.
7. Frederick J. Diedrich and William H. Warren. Why change gaits? Dynamics of the walk-run transition. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1):183–202, 1995.
8. J. A. Scott Kelso. *Dynamic Patterns*. MIT Press, Cambridge, MA, 1995.
9. Peter N. Kugler, J.A. Scott Kelso, and M.T. Turvey. On the concept of coordinative structures as dissipative structures: I. Theoretical lines of convergence. In G.E. Stelmach and J. Requin, editors, *Tutorials in Motor Behavior*. North-Holland, 1980.
10. Peter N. Kugler and M.T. Turvey. *Information, Natural Law, and Self-Assembly of Rhythmic Movement*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.
11. Ilse Lehiste. Isochrony reconsidered. *Journal of Phonetics*, 5:253–263, 1977.
12. S.M. Marcus. Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics*, 30:247–256, 1981.
13. K.L. Pike. *The Intonation of American English*. University of Michigan Press, Ann Arbor, MI, 1945.
14. P. Roach. On the distinction between "stress-timed" and "syllable-timed" languages. In David Crystal, editor, *Linguistic Controversies*. Edward Arnold, London, 1982.
15. Donia R. Scott, S. D. Isard, and Bénédicte de Boysson-Bardies. Perceptual isochrony in English and French. *Journal of Phonetics*, 13:155–162, 1985.
16. Sophie K. Scott. *P-centers in Speech: An Acoustic Analysis*. PhD thesis, University College London, 1993.
17. J.M.T. Thompson and H.B. Stewart. *Nonlinear Dynamics and Chaos*. John Wiley and Sons, New York, NY, 1986.
18. Briony Williams and Steven M. Hiller. The question of randomness in English foot timing: a control experiment. *Journal of Phonetics*, 22:423–439, 1994.