Synergetic Organization in Speech Rhythm

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

February 12, 1997

Abstract

The Speech Cycling Task (SCT) is a novel experimental paradigm developed together with Robert Port and Keiichi Tajima at Indiana University. In a SCT, subjects repeat a phrase containing multiple prominent, or stressed, syllables in time with an external timer, which can be simple or complex. A phase-based collective variable is defined in the acoustic signal. This paper reports on two experiments using the SCT which together reveal many of the hallmarks of hierarchically coupled oscillatory processes.

The first experiment requires subjects to place the final stressed syllable of a small phrase at specified phases within the overall Phrase Repetition Cycle (PRC). It is clearly demonstrated that only three patterns, characterized by phases around $\frac{1}{3}$, $\frac{1}{2}$ or $\frac{2}{3}$ are reliably produced, and these points are attractors for other target phases. The system is thus multistable, and the attractors correspond to stable couplings between the metrical foot and the PRC. A second experiment examines the behavior of these attractors at increased rates. Faster rates lead to mode jumps between attractors. Previous experiments have also illustrated hysteresis as the system moves from one mode to the next.

The dynamical organization is particularly interesting from a modeling point of view, as there is no single part of the speech production system which cycles at the level of either the metrical foot or the phrase repetition cycle. That is, there is no continuous kinematic observable in the system. Nonetheless, there is strong evidence that the macroscopic behavior of the entire production system is correctly described as hierarchically coupled oscillators. There are many parallels between this organization and the forms of inter-limb coupling observed in locomotion and rhythmic manual tasks.

1 Introduction

The model system of bimanual coordination as studied by Kelso and co-workers and modeled originally in Haken, Kelso and Bunz (1985; hereafter, HKB) is well known. Typically, subjects are asked to oscillate two effectors (hands, fingers, etc) simultaneously. A collective variable ϕ is defined as the difference between the phases of the two hands. At moderate rates, two patterns are stable, in-phase ($\phi = 0$) and anti-phase ($\phi = \frac{1}{2}$), while at fast rates, only the in-phase pattern is stable. The intrinsic dynamics of the system features bistability at most rates, with a bifurcation to a monostable regime at a critical frequency. Supporting evidence for the bifurcation comes from critical fluctuations, critical slowing down, characteristic distribution of switching time and more. All these features have been accommodated within the HKB model (Schöner and Kelso, 1988a; Schöner and Kelso, 1988b; Scholz et al., 1987; Scholz and Kelso, 1990). The original observation of bistability was made informally, prior to experimental verification, and is easy to replicate. The system is simple enough that one can be reasonably certain that no hidden modes remain untried.

An important experiment, first done by Yamanishi, Kawato and Suzuki (1980), and later in modified form by Tuller and Kelso (1989) involves subjects trying to maintain a fixed phase relationship between taps performed with one finger of each hand. A target phase, ψ, is given by a pair of visual metronomes which flash with constant periods and a fixed offset. Data can be obtained either while the target is visually present (Tuller and Kelso, 1989), or, following training with feedback, after cessation of the target (Yamanishi et al., 1980). The principal results are the same. Targets at the attractors of the intrinsic system ($\psi \in \{0.5, 0\}$) are reproduced accurately and with low variance. Intermediate targets are produced with values biased towards the attractors (so, e.g. $\psi = 0.4$ results in $\phi \approx 0.45$) and with higher variance.

Apart from adding richness to the model, by allowing coupling between the (known) intrinsic dynamics and an environmental source of information, this procedure is important because the intrinsic dynamics shows up so clearly in the resulting data. That is, if one did not know what the attractor set of the intrinsic system was, this procedure would reveal that set. This is the reasoning behind application of the procedure to the study of the formation of novel attractors at learned phases (Schöner et al., 1992; Zanone and Kelso, 1994).

Speech is a much more complex form of action. Various parts of the speech production system operate at timescales ranging from about 200 Hz to about 0.3 Hz. Neither the motion of the articulators, nor the acoustic product, are easily divisible into discrete components. Studies of reiterant speech, where the syllables of a phrase are all replaced by a single syllable such as /ma/ have revealed that articulator motion shares a common dynamical organization with motion of the limbs, and may be appropriately modeled by a second-order limit cycle dynamics (Ostry et al., 1983; Vatikiotis-Bateson and Kelso, 1993). These studies are based on kinematic data obtained from the motion of the articulators, comparable to the kinematic data obtained in the above studies of bimanual coordination.

Many linguistic and phonetic studies point to the metrical foot as being the rhythmical unit of English speech (Classé, 1939; Lehiste, 1977; Couper-Kuhlen, 1993). The metrical foot, as traditionally defined, comprises a stressed syllable and any and all following unstressed syllables. In the present context, that definition will be slightly extended (see below). Unfortunately, no kinematic variable which cycles once per foot is available, and so a synergetic approach to studying speech rhythm must first arrive at a suitable collective variable. This is described in the next section.

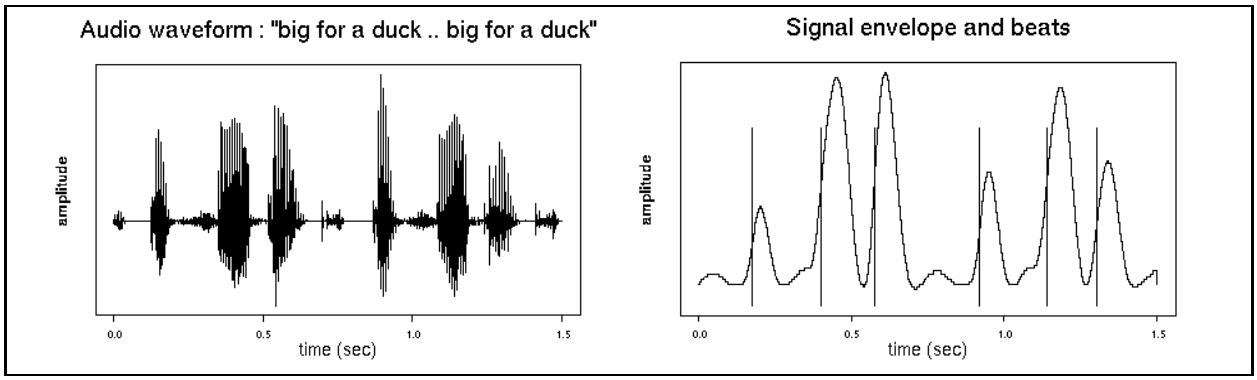

Figure 1: Left panel: waveform of the phrase *big for a duck* repeated twice. Right panel: The signal has been bandpass filtered around 1000 Hz, rectified and smoothed. Beats (vertical lines) are located halfway through a local rise in amplitude.

2 Speech cycling

An event-based conception of rhythm requires the identification of punctate events, or beats. Figure 1 illustrates a procedure for locating beats in the acoustic signal. This procedure, described in detail in Cummins and Port (1997), is derived from models of P-center location (Morton et al., 1976; Scott, 1993), and it places the beat close to the onset of the vowel. A metrical foot can now be given an operational definition as the interval between two beats associated with stressed syllables.

Now consider the data obtained by a subject who repeats a short phrase, say *big for a duck*, over and over at a self-selected comfortable rate. If we ignore breath pauses, a reasonably regular sequence of beats associated with the syllables *big* and *duck* emerges. We can define the Phrase Repetition Cycle (PRC) to be the interval between the beats of successive phrase onsets, that is, the beats of successive tokens of the word *big*. What we find, in general, is that unconstrained repetition will result in the beat of *duck* occurring half way through the PRC ($\phi_{duck} = 0.5$). For a simple two-stress phrase, we have now defined a collective variable, ϕ, which is the phase of the PRC at which a medial beat occurs. The initial impression is that there is a stable mode at $\phi = 0.5$, corresponding to an isochronous series of stresses. Unlike the bimanual system studied by Kelso et al., it is not immediately obvious what the complete set of stable modes might be. For this reason, a procedure similar to that of Yamanishi et al. (1980) and of Tuller and Kelso (1989) is defined. This procedure can be called Targeted Speech Cycling.

In a targeted speech cycling experiment, subjects hear a succession of alternating high and low beeps over headphones. The low beep is located at a target phase ψ of the cycle defined by the succession of high beeps. Their task is to repeat a short phrase, as above, trying to line up the beats of two given syllables with the two tones. That is, they try to produce a specified phase $\phi = \psi$. In order to simplify matters (for the experimenter), subjects are trained to stop repeating the phrase when they need to breathe. In this way, data from the cycle which contains the breath can easily be identified and omitted.

This outlines the basic speech cycling procedure. Many variations on this basic theme are, of course, possible, some of which will be described below. What is important to note is that the collective variable ϕ has been defined over the *acoustic signal*, not over the motion of the articulators. This strategy allows the empirical investigation of organization at the level of the metrical foot and

above. One drawback of this method is that ϕ is a pointwise measure, providing a single value per cycle. No continuous measurement is available, making the observation of fine detail such as the presence of critical fluctuations difficult.

3 Experiment 1: Mapping attractors in a speech cycling task

A baseline experiment, reported in full in Cummins and Port (1997), was designed to look for attractive phases over a range of $0.3 < \phi < 0.7$. Note that not all phases between 0 and 1 can be examined, as some time must elapse between the phrase onset and the medial beat if speech is to be produced at all, and similarly, some time must be provided between the medial beat and the onset of the following phrase if the final word is to be completely produced.

Stimuli The stimuli consisted of 14 pairs of alternating high and low sinusoidal beeps. The interval from high to low tone was fixed throughout this experiment at 700 ms. The target phase for each trial was drawn from a uniform random distribution with end points at 0.3 and 0.7. The interval from low to high tone was then computed to provide the desired target phase, $\psi = (\text{high}_i - \text{low}_i)/(\text{high}_{i+1} - \text{high}_i)$. This gave a high–high cycle length within the range 1.0 sec (for a target phase of 0.7) to 2.333 sec (for a target of 0.3). Stimuli were played at a self-selected comfortable listening level over headphones. The intensity of the last two pairs of tones were scaled down by factors of 0.66 and 0.33, respectively, so that the tones faded out rather than stopping abruptly.

Task Subjects were required to listen to the first two pairs of tones, and then to join in, repeating a given phrase in time with the stimulus. They were to attempt to line up the first syllable with the high tone and the final syllable with the low tone. When the stimulus stopped, they were to continue repeating the phrase, attempting to maintain the same pattern, until signaled to stop by the experimenter. Approximately equal amounts of data were collected with and following the stimulus. The phrases came from a set of 30, all of the form *X for a Y*, where X and Y were chosen subject to certain phonetic constraints which helped to ensure consistency in beat measurement. Although all words were English words, most of the phrases were meaningless. In a first pass of the experiment, 4 subjects took part, each completing 90 trials distributed over 3 sessions. As post hoc analysis showed that the one subject who was a non-musician differed markedly from the others, a control group of 3 non-musician and one musician was added, each of which completed 30 trials. Finally, 4 of these subjects took part in another condition where the interval from high to low tone was reduced to 450 ms, which required a substantially faster speaking rate. In this condition 30 trials with target phases evenly spaced over the range [0.3–0.7] were used.

Results As full results have been presented elsewhere (Cummins and Port, 1997), representative data from one subject only will be shown here. The data are from 90 trials. The left hand panel of Figure 2 is a histogram showing the distribution of trial medians of ϕ, the produced phase. Recall that targets were drawn from a uniform distribution between 0.3 and 0.7. The distribution of phases produced is clearly trimodal, with phases at about $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$ much more likely to be produced than other, intermediate, phases. In the middle panel the difference between the produced phase and the target phase is plotted as a function of the target. Accurate reproduction of the target would result in data on the line $y = 0$. The data cluster into three well defined groups, corresponding

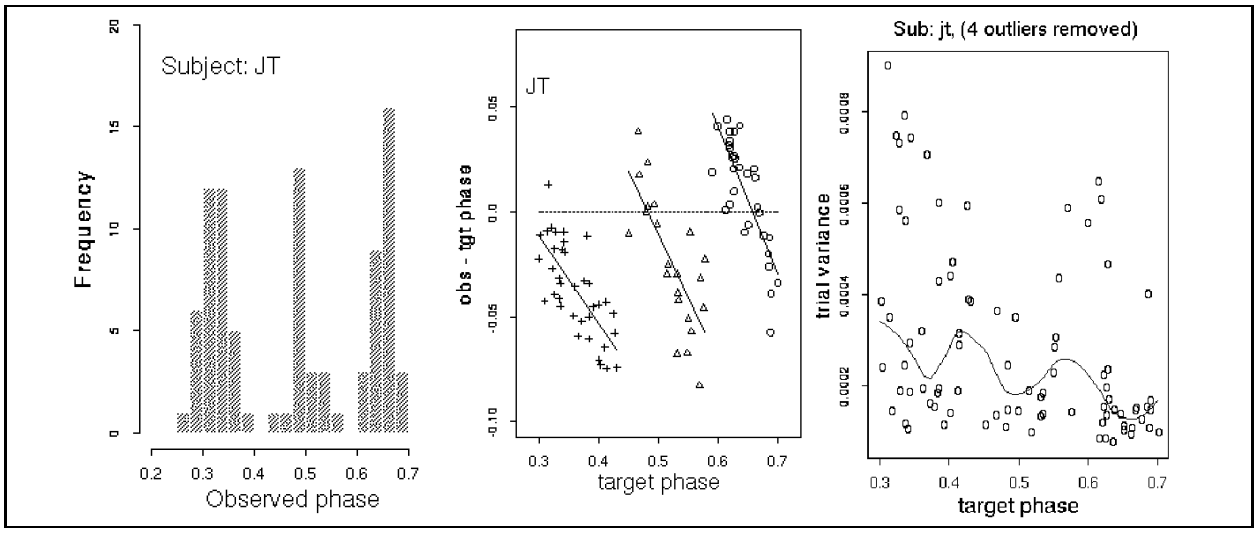

Figure 2: Left: Histogram showing the distribution of observed phases for one subject over 90 trials. Each datum is a trial median. Center: $\psi - \phi$ as a function of ψ. Data on the line $y = 0$ reflect accurate reproduction of the target phase. The three plotting symbols serve only to distinguish three clusters in the data. A local regression line has been fitted to each cluster. Right: Variance as a function of ψ. A local curve fitting routine has been used to identify approximate minima.

directly to the three modes of the histogram in the left hand panel. These clusters are marked by distinct plotting symbols, and a local regression line has been fitted to each cluster. The strength of the linear relationship and the negative slope of each line demonstrates that the three preferred patterns are, indeed, attractive patterns, with intermediate targets eliciting phases biased towards the attractors. Finally, the right hand panel plots the trial variance as a function of target phase. A locally weighted quadratic regression procedure is used to fit a smooth curve. Minima are observed close to the attractor phases.

Some other features of the results deserve mention. Firstly, no difference was observed between data elicited together with the stimulus and after stimulus cessation. Two of eight subjects showed evidence of two patterns only, at about $\frac{1}{3}$ and $\frac{1}{2}$. These subjects experienced difficulty with the task as set. It appears that they simply did not discover the missing pattern at $\frac{2}{3}$. No feedback was provided, which might have helped them in the task. The fact that the only such subject among the initial 4 subjects was also the only non-musician of the group led to a control group being tested which included more non-musicians. The non-musicians of the control group produced patterns similar to the musicians of the initial group, while the one musician of the control group (a tuba player) also produced only $\frac{1}{3}$ and $\frac{1}{2}$. Although this and other experiments have suggested that musicians may find speech cycling tasks easier than non-musicians, the finding of three discrete and stable patterns is not restricted to musicians.

When speaking rate was increased by shortening the interval between high and low tones, two of the four subjects produced trimodal data, exactly as in the slower condition. The other two, however, produced data which showed attractors at $\frac{1}{3}$ and $\frac{1}{2}$ only. This is shown in Figure 3 by the plots of observed phase minus target phase as a function of target phase for all four subjects. This also serves to graphically illustrate the remarkable discreteness of the behaviors observed.

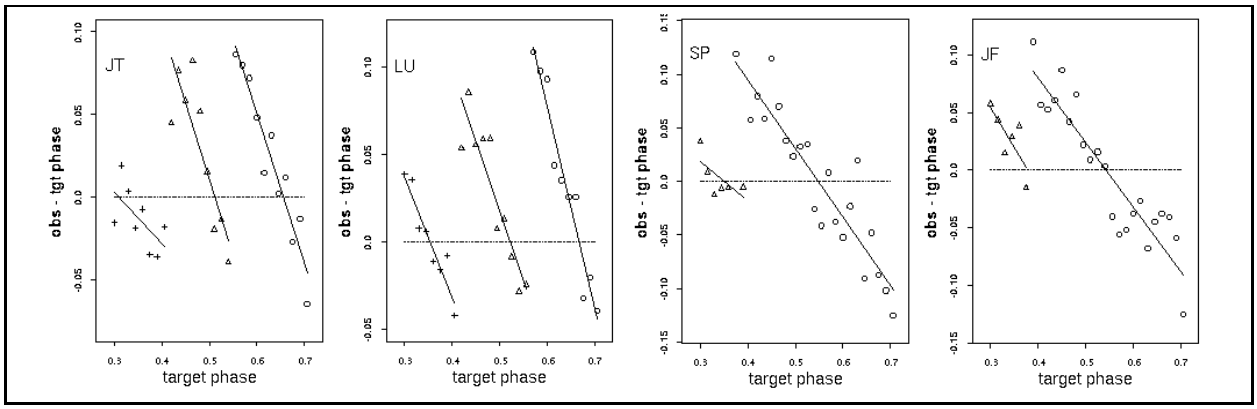

Figure 3: Plot of $\phi - \psi$ as a function of ψ for the four subjects of the fast condition. Two subjects show three attractors; the other two exhibit only two each, losing the attractor at $\frac{2}{3}$. Again the data have been clustered and local regression lines fitted. There are 30 trials per subject.

To summarize, the data lay bare the gross features of an intrinsic dynamics characterized by attractors at $\phi = \frac{1}{3}$, $\frac{1}{2}$ or $\frac{2}{3}$. There may be some initial evidence of differential stability among the attractors, as the attractor at $\frac{2}{3}$ is not found for two of eight subjects at moderate speaking rate and for two of four at a fast rate. The observed behaviors are remarkably consistent across subjects, and are seen whether the target phase is physically present or is taken from recent memory.

4 Mode jumping

The previous experiment demonstrated that three patterns can be reliably produced at moderate speaking rates. The next experiment had two goals: to see if mode jumps between these discrete patterns could be elicited, and to see if differential stability among the attractors could be observed as rate is increased. Speech rate is given an operational definition here by simply shortening the PRC. Clearly, for a given, short, PRC of, say, 0.6 sec, the pattern which best fills the interval between phrase onsets is the one with an attractor at $\frac{2}{3}$; that with an attractor at $\frac{1}{3}$ will be harder, as the bulk of the speech material must be fitted into the first 200 ms.

Method Subjects were presented with a stimulus which set a target phase which was one of $\psi = \frac{1}{3}$, $\frac{1}{2}$ or $\frac{2}{3}$. The high–low interval (that is, the interval from phrase onset to final stress) was fixed at 700 ms, so the overall PRC was 2.1, 1.4 or 1.05 seconds respectively. They listened to the target, and began repeating a simple phrase, as in the previous experiment. After 6 repetitions, the low tone, which provided the target phase, was switched off, leaving only the high tone which cued the start of the PRC. Thereafter, the length of the PRC was decreased by 2% on each cycle, until it became so short that the subject could no longer repeat the phrase, at which point they simply stopped. Subjects were instructed to try and maintain the original pattern, but in the event of difficulty, they should try to repeat the phrase once per beep. Six subjects completed six trials at each of the three starting conditions. As before, breath pauses were excluded from the data.

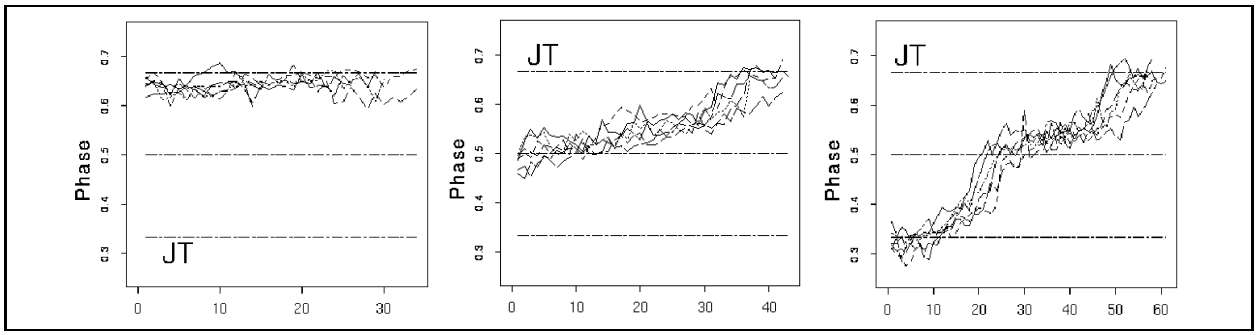

Figure 4: Time series data from a single subject. The abscissa indicates repetition number, the ordinate is phase. The PRC shortens by a constant proportion from one repetition to the next. Initial target phase is (left) $\frac{2}{3}$, (mid) $\frac{1}{2}$ or (right) $\frac{1}{3}$. These values are marked by horizontal lines. There are six trials for each condition.

Results The analysis of the data from this experiment is still underway. Sample data from a single subject (subject JT, as in the first experiment) are presented in Figure 4. In all three conditions, the initial pattern is produced with reasonable accuracy at the start of the trial. When the initial target is $\frac{2}{3}$, the PRC is already quite short, and after about 30 repetitions (plus breath pauses, not shown), the subject stops. When the initial target is $\frac{1}{2}$, more repetitions are obtained. It can be seen that there is a gradual drift to larger phase values. Given the continuous decrement in the interval between beeps, some shift to a higher phase value is understandable. It is possible that some trials show a jump to the attractor at $\frac{2}{3}$ towards the end of the trial, but the data stop so shortly thereafter that this interpretation is not unequivocal. The more interesting case is the final condition, where the initial target is $\frac{1}{3}$, and the PRC starts as a relatively long interval. As the PRC is gradually shortened, the subject finds herself trying to cram all the speech material into the first third of the cycle, while the remaining two-thirds is largely pausal. Eventually this leads to discrete jumps, first to the attractor at $\frac{1}{2}$, and then, later, on to $\frac{2}{3}$. All six subjects show the first jump, and three of the six reliably show the second. One subject (not shown) even jumps directly from $\frac{1}{3}$ to $\frac{2}{3}$, bypassing $\frac{1}{2}$, on three of the six trials.

5 Discussion

A technique was applied in Tuller and Kelso (1989), and in Yamanishi et al. (1980) which used a range of environmentally specified phases to map out the intrinsic dynamics of a system of bimanual coordination. This found later application in studies of attractor emergence during learning (Kelso, 1990; Schöner et al., 1992; Zanone and Kelso, 1994). The present work seeks to expand the application of this procedure to the more complex domain of speech. This is only possible by defining a collective variable in the acoustic, rather than the articulatory, domain. The expedient of assigning beats to the beginnings of stressed syllables allows the definition of a variable which is easy to measure, although it delivers only a single scalar per phrase repetition.

The first experiment demonstrated multistability, with three discrete patterns being produced. The reader can reproduce these three patterns by following the musical transcription shown in Figure 5. Note that this experiment explored only the range [0.3–0.7]. The possibility of attractors at, say, 0.25 and 0.75 cannot be excluded; in fact these seem quite probable. In the second experi-

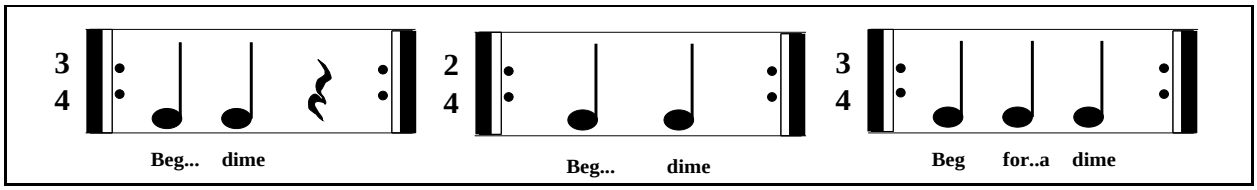

Figure 5: Musical representation of the three discrete patterns produced in the first experiment. The subject repeats the phrase *beg for a dime*. Left: $\phi = \frac{1}{3}$. Middle: $\phi = \frac{1}{2}$. Right: $\phi = \frac{2}{3}$.

ment, it was shown that discrete jumps among the attractors may be observed. Differential stability among the attractors was not established unequivocally. A previous experiment (Cummins and Port, 1996a,b) demonstrated that the position of the jump between attractors depends on the direction in which the control variable (rate) is manipulated. That is, the transition is hysteretic.

So what do these experiments tell us about speech production? The experimental task used here imposes an artificial periodicity on the speech, the Phrase Repetition Cycle. Under these conditions, another period becomes manifest, cycling twice (for $\phi = \frac{1}{2}$) or three times ($\phi \in \{\frac{1}{3}, \frac{2}{3}\}$) for each period of the PRC. This period corresponds to the metrical foot or inter-stress interval, well known to metrical phonologists (Hayes, 1985; Kiparsky and Youmans, 1989). The collective variable, ϕ, captures in succinct form the nesting relationship which obtains between these two cycles. The coupling which is evident in these experiments constitutes strong evidence for the reality of the metrical foot as a well-defined unit in the production of speech. Furthermore, the observation of mandatory coupling with an imposed cycle suggests the description of the temporal evolution of the foot as a second order limit cycle dynamics. This serves to establish a commonality in the geometry underlying the production of the foot with production of gestures (Ostry et al., 1983; Vatikiotis-Bateson and Kelso, 1993) and with movements of the limbs (Kugler et al., 1980; Kelso et al., 1980; Cummins and Port, 1996a).

6 Acknowledgments

This work grew from many brainstorming sessions and some pilot experiments together with Robert Port and Keiichi Tajima. Thanks also to Betty Tuller for crucial pointers at an early stage, and to Mauri Kaipainen for some helpful comments on this manuscript. The work was funded by an Indiana University Research Incentive Dissertation Year Fellowship and by a grant from the Cognitive Science Program at Indiana University.

References

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

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

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

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

Cummins, F. and Port, R. F. (1997). Rhythmic constraints on stress timing in English. *Journal of Phonetics*. Submitted.

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

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

Kelso, J. A. S. (1990). Phase transitions: foundations of behavior. In Haken, H. and Stadler, M., editors, *Synergetics of Cognition*, volume 45 of *Springer Series in Synergetics*, pages 249–268. Springer Verlag, Berlin.

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

Kiparsky, P. and Youmans, G., editors (1989). *Rhythm and Meter*, volume 1 of *Phonetics and Phonology*. Academic Press, San Diego.

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

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

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

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

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

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

Schöner, G., Zanone, P. G., and Kelso, J. A. S. (1992). Learning as change of coordination dynamics: Theory and experiment. *Journal of Motor Behavior*, 24(1):29–48.

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

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

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

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

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

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

Zanone, P.-G. and Kelso, J. A. S. (1994). The coordination dynamics of learning. In *Interlimb Coordination: Neural, Dynamical, and Cognitive Constraints*, chapter 22, pages 461–490. Academic Press.