The dynamical perspective on speech production: data and theory

J. A. S. Kelso*,†, E. L. Saltzman†‡ and B. Tuller*†

*Center for Complex Systems and Department of Psychology, Florida Atlantic University, Boca Raton, Florida 33431, U.S.A.
†Haskins Laboratories, New Haven, Connecticut, 06511-6695 U.S.A.
‡Department of Psychology, University of Connecticut, Storrs, Connecticut 06268, U.S.A.

Presented here, in preliminary form, is a general theoretical framework that seeks to characterize the lawful regularities in articulatory patterns that occur when people speak. A fundamental construct of the framework is the co-ordinative structure, an ensemble of articulators that functions cooperatively as a single task-specific unit. Direct evidence for co-ordinative structures in speech is presented and a control scheme that realizes both the contextually varying and invariant character of their operation is outlined. Importantly, the space–time behavior of a given articulatory gesture is viewed as the outcome of the system's dynamic parameterization, and the orchestration among gestures is captured in terms of inter-gestural phase information. Thus, both time and timing are deemed to be intrinsic consequences of the system's dynamical organization. The implications of this analysis for certain theoretical issues in coarticulation raised by Fowler (Journal of Phonetics, 1980, 8, 113–133) receive a speculative, but empirically testable, treatment. Building on the existence of phase stabilities in speech and other biologically significant activities, we also offer an account of change in articulatory patterns that is based on the nonequilibrium phase transitions treated by the field of synergetics. Rate scaling studies in speech and bimanual activities are shown to be consistent with a synergetic interpretation and suggest a principled decomposition of languages. The CV syllable, for example, is observed to represent a stable articulatory configuration in space–time, thus rationalizing the presence of the CV as a phonological form in all languages. The uniqueness of the present scheme is that stability and change of speech action patterns are seen as different manifestations of the same underlying dynamical principles—the phenomenon observed depends on which region of the parameter space the system occupies. Though probably wrong, ambitious, and the outcome of much idle speculation, the simplicity of the present scheme is attractive and may offer certain unifying themes for the traditionally disparate disciplines of linguistics, phonetics and speech motor control.
Prologue

The present paper represents, in part, a program of research that seeks to understand the lawful regularities that occur in articulatory patterns when people speak. The term dynamical in the title should not be interpreted as pure biomechanics. Rather, dynamics is used here in the fashion of Maxwell (1877), a forerunner of modern treatments of dynamical systems; namely, as the simplest and most abstract description of the forms of motion produced by a system. In a complex system like that of speech production, it is clearly impossible to investigate the behaviour of each microscopic degree of freedom, however one defines them. The challenge of a dynamical approach is to identify and then lawfully relate macroscopic parameters (that operate on slow time scales) to the behavioral interactions among more microscopic articulatory components (that operate on faster time scales). A putative but important advantage of a dynamical approach that, in principle, may allow for a unification of linguistics, phonetics, and speech motor control is the level independent nature of dynamical description. Thus, dynamics can be specified at a global abstract level for a system of articulators as well as at the local, concrete level of muscle-joint behavior. The issue then becomes less one of translating a “timeless” symbolic representation into space–time articulatory behavior, as it is one of relating dynamics that operate on different intrinsic time scales. Obviously, this is only a way of posing the problem, but we believe—in the absence of evidence to persuade us otherwise—that it offers a principled solution.

1. Introduction

When a speaker produces a word, it is well-known that the physical description of the word (whether acoustic, as displayed in a spectrogram or waveform, or articulatory, as in a cineradiographic sequence) varies widely with many factors. Among these are the rate at which the speaker talks, the word’s pattern of syllabic stress or emphasis within an utterance and the phonetic structure of surrounding words. The variations that arise as a consequence of such factors have long resisted unified systematic descriptions. Despite intensive research efforts, no one has sufficiently described either the acoustic or articulatory information that serves to specify a word in all its various contexts. Nor has anyone, to our knowledge, identified a canonical shape for a word and then transformed it, in a principled fashion, into the many other shapes that it may take.

Along with colleagues at Haskins Laboratories (e.g. Browman, Goldstein, Kelso, Rubin & Saltzman, 1984) we are developing a solution to this problem by treating the units of language—conventionally described by linguists and phoneticians as, for example, phonemes, syllables and words—as the product of time-invariant control structures for a system of vocal tract articulators. We assume that it is from the properties of these dynamically specified control structures, to be described presently, that the observed physical variations naturally arise.

A central hypothesis derived from the present theoretical framework is that articulators seldom move in an isolated, independent fashion (cf. Bernstein, 1967). In speech production they are coordinated with one another in such a way that changes over time in vocal tract shape are produced. Such changes in vocal tract shape structure the

1Note that this claim, to be supported here, seems to run counter to the notion that normal phoneme rates “can be achieved only if separate parts of the articulatory machinery—muscles of the lips, tongue, velum, etc.—can be separately controlled” (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967).
sounds of speech for a listener. A central problem for the theory, then, which we shall address in the present paper, becomes one of characterizing interarticulator cooperation in a multi-degree of freedom system, and identifying the “significant informational units of action” (Greene, 1971, p. xviii) for speech. We, and others, have provided theoretical and empirical support for the hypothesis that, for skilled movements of the limbs or speech articulators, such action units (or coordinative structures) do not entail rigid or hardwired control of joint and/or muscle variables (e.g. Fowler, 1977; Turvey, 1977; Kugler, Kelso & Turvey, 1980; Kelso & Tuller, 1984a; Kelso, in press). Rather, they are defined more abstractly in a task-specific manner, and serve to marshal the articulators temporarily and flexibly into functional groupings or ensembles of joints and muscles that can accomplish particular goals. But what principles govern the assembly of coordinative structures for speech and how can such structures be explicitly modeled?

In Section 2 of the present paper we present evidence for coordinative structures in speech and discuss how they might be used in the production of single syllable utterances. Our focus is primarily on the task-specific stability of coordinative structures in the face of either experimentally induced mechanical perturbations, or “natural” perturbations that might occur during ongoing speech as a result of contextual variations. A key feature of the model we are developing, termed task dynamics (e.g. Saltzman & Kelso, 1983) is that it allows one to define invariant control structures for specific vocal tract gestural goals, from which contextually varying patterns of articulatory trajectories arise. These structures are invariant in two ways, both qualitatively, in terms of the set of relations among dynamic parameters (analogous to mass, stiffness, damping, etc.), and quantitatively, in terms of the parameter values themselves.

Speaking a word entails laryngeal and supralaryngeal gestures involving coordinated activity of many articulators. But words are not simply strings of individual gestures, produced one after the other; rather, each is a particular pattern of gestures, orchestrated appropriately in space and time. A way to probe the nature of the underlying ordering process is to induce naturally occurring scaling transformations, such as changes in speaking rate and degree of prosodic stress, and search for those aspects of the articulatory pattern that remain stable across these transformations. In Section 3 of the present paper we reconceptualize and perform an extensive re-analysis of earlier work that showed that the relative timing among articulators was a crucial feature of intergestural coordination. These steps point to the importance of phase as a key dependent variable, a finding that has empirical and theoretical implications for understanding both the stability of the spatiotemporal orchestration among gestures and the dynamical control structures that underlie such patterns (Kelso & Tuller, 1985a,b). We outline a dynamical account of speech production\(^2\) that differs radically from views that characterize speech as a planned sequence of static linguistic/symbolic units that are different in kind from the physical processes involved in the execution of such a plan. Rather, we hypothesize that the coordinative structures for speech are dynamically defined in a unitary way across both abstract “planning” and concrete articulatory “production” levels. These units are not timeless, but rather incorporate time in an intrinsic manner (cf. Fowler, 1980; Bell-Berti & Harris, 1981).

In the final section we discuss new directions—experimental and theoretical—for enlarging our understanding of the subtleties of dynamical structure that underlie

\(^2\)With implications, no doubt for speech perception which is, however, beyond the mission of this paper (but see, for example, Liberman & Mattingly, 1985).
changes in critically scaled articulatory patterns. We speculate that the form of such changes may in fact offer a window into, and perhaps even rationalize, the basic units of phonological analysis.

2. On coordinative structures in speech

2.1. Theory and data

The production of a single syllable requires the cooperation among a large number of neuromuscular components at respiratory, laryngeal, and supralaryngeal levels, operating on different time-scales. Yet somehow from this huge dimensionality the sound emerges as a distinctive and well-formed pattern. How this “compression” occurs—from a microscopic basis of huge dimensionality to a low-dimensional macroscopic description—is central to many realms of science, not only to understanding the coordination among speech articulators (see, for example, Kelso & Scholz, 1985). For example, there are many neurons, neuronal connections, metabolic components, muscles, motor units, etc. involved in pointing a finger at a target, yet the action itself is nicely modeled by a mass-spring system, a point attractor dynamic in which all system trajectories converge asymptotically at the desired target (e.g. Kelso, 1977; Polit & Bizzi, 1978; Cooke, 1980; Schmidt & McGown, 1980).

Is it, in fact, the case in speech that the higher dimensionality available actually reduces to a lower-dimensional, controllable system? If so, on what principles does such compression or reduction of the many degrees of freedom rest? These questions amount to a basic problem in the control of complex systems, i.e. determining the circumstances under which a small set of control parameters (K) can effectively manipulate a much larger number of degrees of freedom (N). As Rosen (1980) notes, it is usually the case that K \ll N, so that unless further constraints are imposed, it is not possible to impose arbitrary controls on N degrees of freedom.

But what form do such constraints take? Is there any evidence that the many degrees of freedom are actually constrained in a systematic fashion when a person talks? In earlier work, Fowler (1977, 1980) has described some mostly indirect evidence that the many neuromuscular components involved in speech do, in fact, cooperate to form functionally specific action units, or as we prefer to call them, coordinative structures (e.g. Turvey, 1977). Here we supply more direct experimental support.

Support for the hypothesis that a group of relatively independent muscles and joints forms a single functional unit would be obtained if it were shown that a challenge or perturbation to one or more members of the group was, during the course of activity responded to by other remote (non-mechanically linked) members of the group. We have recently found that speech articulators (lips, tongue, jaw) produce functionally specific, near-immediate compensation to unexpected perturbation, on the first occurrence, at sites remote from the locus of perturbation (Kelso, Tuller & Fowler, 1982; Kelso, Tuller, Vatikiotis-Bateson & Fowler, 1984b). The responses observed were specific to the actual speech act being performed: for example, when the jaw was suddenly perturbed while moving toward the final /b/ closure in /bae/, the lips compensated so as to produce the /b/, but no compensation was seen in the tongue. Conversely, the same perturbation applied during the utterance /baez/ evoked rapid and increased tongue muscle activity (appropriate for achieving a tongue–palate configuration for the final fricative sound) but no active lip compensation.
In order to explore the microscopic workings of a coordinative structure, recent work has also varied the phase of the jaw perturbation during bilabial consonant production. Remote reactions in the upper lip were observed only when the jaw was perturbed during the closing phase of the motion, that is, when the reactions were necessary to preserve the identity of the spoken utterance (see also Munhall & Kelso, 1985). Thus the form of cooperation observed is not rigid or "hard wired": the unitary process is flexibly assembled to perform specific functions (for additional evidence in speech and other activities, see Abbs, Gracco & Cole, 1984; Kelso et al., 1984b; Berkenblit, Fel'dman & Fukson, in press). Elsewhere we have drawn parallels between these findings and brain function in general (Kelso & Tuller, 1984a). Just as groups of cells, not single cells appear to be the main units of selection in higher brain function (Edelman & Mountcastle, 1978), so too task-specific ensembles of neuromuscular elements appear to be significant units of control and coordination of action, including speech.

To propose the coordinative structure as a fundamental unit of action does not just involve a change in terminology. It is to take us away from the hard-wired language of reflexes and central pattern generators (CPG) or the had-algorithmic language of computers (formal machines) which is the source of the motor program/CPG idea. reflexes and central pattern generators (CPG) or the had-algorithmic language of in terms of affording an understanding of coherent action. The fact that we observe functionally specific forms of cooperative behavior in many different creatures (e.g. the wiping behavior of the spinal frog: Fukson, Berkenblit & Fel'dman, 1980; Berkenblit et al., in press) with vastly different neuroanatomies suggests that there may be nothing special, a priori, about neural structures and their “wiring” that mandates the existence of coordinative structures. Rather, it suggests that the functional cooperativity—not the neural mechanism per se—is fundamental. Although neural processes serve to instantiate such functions and support such cooperative behaviors, it is the lawful dynamical (rather than neural) basis of these cooperative phenomena that is our primary theoretical and experimental concern. This is where we part company with certain current views of motor control. Contrary to the motor programming formulation that relies on symbol-string manipulation familiar to computer technology [and, we would add, the whole “information processing” perspective of cognitive science (Carello, Turvey, Kugler & Shaw, 1984)], the construct of coordinative structures highlights both the analytic tools of qualitative (nonlinear) dynamics (e.g. Kelso, Holt, Rubin & Kugler, 1981; Saltzman & Kelso; 1983; Kelso, Vatikiotis-Bateson, Saltzman & Kay, 1985) which provides low-dimensional descriptions of forms of motion produced by high-dimensional systems, and the physical principles of cooperative phenomena (e.g. Haken, 1975; Kugler et al., 1980; Kelso, 1981; Kugler, Kelso & Turvey, 1982; Kelso & Tuller, 1984a, b; Haken, Kelso & Bunz, 1985) which accounts for the emergence of order and regularity in nonequilibrium, open systems. Though preliminary, both approaches will be apparent below and in following sections.

2.2. Task dynamic modeling

One way of trying to understand the operation of a coordinative structure is to model it. What type of model could generate, in a task-specific manner, the trajectories characteristic of normal unperturbed speech gestures and the spontaneous, compensatory behaviors discussed above? Here we discuss briefly how these issues of multi-articulator coordination within single speech gestures are treated in a task-dynamic
model (Saltzman & Kelso, 1983; Saltzman, in press) recently developed for effector systems having many articulatory degrees of freedom. Finally, we describe some preliminary attempts to model multiarticulator coordination within two temporally overlapping speech gestures, with reference to "naturally" induced compensatory behaviors (i.e. coarticulation).

Task dynamics is able to model the phenomenon of immediate compensation without requiring explicit trajectory planning or replanning (see Saltzman & Kelso, 1983, for further details). Note that defining invariant patterns of dynamic parameters at the level of articulatory degrees of freedom (e.g. stiffness and damping parameters for the jaw and lips) will not suffice to generate these behaviors. The immediate compensation data for speech described above (Kelso et al., 1984b) could not be generated by a system with a constant rest configuration parameter (i.e. a vector whose components are constant rest positions for the lips and jaw, cf. Lindblom, 1967). As shown in these data, when sustained perturbations were introduced during articulatory closing gestures, the system "automatically" achieved the same constriction as for an unperturbed gesture but with a different final or rest configuration. Thus, immediate compensation appears to result from the way that dynamic parameters at the articulatory level are constrained to change during a gesture in a context-dependent manner. In the task-dynamic model, such patterns of constraint originate in corresponding invariant patterns of dynamic parameters at an abstract, functionally defined level of task description.

There are three main steps involved in simulating coordinate movements of the speech articulators using the task-dynamic model. Since simulations to date have focused mainly on bilabial gestures, we will describe these three steps in some detail with reference to the specific example of a discrete bilabial closure task.

2.2.1. Task space
The first step is to specify the functional aspects of the given speech gesture with reference to the constriction forming movements of an idealized vocal tract. This is done in a two-dimensional task space whose axes represent constriction location and constriction degree, and the topological form of the control regime for each task-space variable is specified according to the functional characteristics of the given speech task. For example, discrete and repetitive speech gestures will have damped (e.g. point attractor) and cyclic (e.g. limit cycle) second-order system dynamics, respectively, along each axis. At the task-space level, then, the dynamical system or control regime is abstract in that the constriction being controlled is independent of any particular set of articulators, and can refer, for example, to either a bilabial constriction produced by the lips and jaw or to a tongue–palate constriction produced by the tongue and jaw. Since we have chosen a discrete closure task to illustrate the steps involved in our task-dynamic simulations, we specify invariant, damped, second-order dynamics for the articulator-independent constriction along each task axis [see Fig. 1(a)].

2.2.2. Body space; tract variables
The second step in modeling bilabial closure is to transform the task-space system kinematically into a two-dimensional body-space system defined in the midsagittal plane of the vocal tract. In contrast to the task-space regime, the body-space dynamics are specific to a given set of articulators whose movements govern the bilabial constriction along the tract variable dimensions of lip aperture (LA) and lip protrusion (LP). These tract variables represent the body-space counterparts of the task-space variables of
constriction degree and location, respectively [see Fig. 1(b)]. Lip aperture is defined by the vertical distance between the upper and lower lips, and lip protrusion by the horizontal distances in the anterior–posterior direction of the upper and lower lips from the upper and lower teeth, respectively. Upper and lower lip protrusion movements are not independent in our preliminary formulation, but have been constrained to be equal in the model for purposes of simplicity. Consequently, like constriction location in task space, lip protrusion in body space currently constitutes only a single degree of freedom. This constraint may be abandoned in future work, as we attempt to model gestures in which the upper and lower lips show very different horizontal positions (e.g. labiodental fricatives). Finally, it should be noted that the result of transforming from task space to body space coordinates is to define a two-dimensional set of motion equations with a constant (although transformed) set of dynamic parameters. The tract variable control regimes are independent, since their corresponding equations of motion are uncoupled.

2.2.3. Model articulator space
The third step in modeling the closure task is to transform kinematically the two-dimensional tract variable regime into the coordinates of a four-dimensional model articulator space. The model articulators are moving segments that have lengths but are massless [see Figure 1(c)], and are defined with reference to the simplified articulatory degrees of freedom adopted in the Haskins Laboratories software articulatory speech synthesizer (Rubin, Baer & Mermelstein, 1981). For bilabial gestures, the set of articulator
movements associated with lip aperture includes rotation of the jaw and vertical displacements of the upper lip and lower lip relative to the upper and lower front teeth, respectively; for lip protrusion, the set of articulator movements includes (currently) yoked horizontal displacements in the anterior–posterior direction of the upper and lower lips relative to the upper and lower front teeth, respectively.

Since there are more model articulator variables than tract variables for the bilabial closure task, the model articulator system is redundant and the inverse kinematic transform from tract variables to model articulator coordinates is indeterminate (e.g. Saltzman, 1979). In order to deal with the indeterminacy or one-to-many property of this transformation, a weighted, least-squares optimality constraint is introduced in the form of a weighted Jacobian pseudoinverse transformation. This pseudoinverse has also been used in control schemes for robot arms that have a surplus number of degrees of freedom (i.e. the number of joints in the arm is greater than the number of task-relevant, spatial degrees of freedom for the hand, e.g. Whitney, 1972; Benati, Gaglio, Morasso, Tagliasco & Zaccaria, 1980; Klein & Huang, 1983). Specifically, the pseudoinverse is a function of two matrix components—the Jacobian and articulator weighting matrices. The Jacobian matrix defines the transformation that relates motions of the articulators at their current configuration or posture to corresponding tract-variable motions of the bilabial constriction. The elements of the Jacobian matrix are nonlinear functions of the current articulatory posture. The elements of the articulator weighting matrix, however, are constant during a given gesture. In current modeling, a given set of articulator weightings constrains the motion of the articulators in direct proportion to the relative magnitude of the corresponding weighting element. Hence, different articulator weighting patterns are associated with different amounts of relative motion on the part of the four articulators responsible for controlling the tract variables of the bilabial constriction. In this sense, elements of the articulator weighting matrix used in the associated pseudoinverse define a further set of constant parameters for the bilabial constriction’s equation of motion.

To summarize, in the task-dynamic model one may interpret the task-specific, coherent movements of the model articulatory system as resulting from the way that instantaneous tract-variable “forces” acting on a particular vocal tract constriction are distributed across the model articulators during the course of the tract variable gesture. At any given instant during this gesture, the partitioning is based on two factors:

(a) the task-specific, constant set of articulator weightings and tract-variable dynamic parameters (e.g. lip aperture stiffness and damping); and
(b) the current values of elements in the posturally dependent Jacobian matrix: because these elements are functions of the current posture of the model articulators, the dynamic parameters defined at the level of the model articulator variables (e.g. stiffness and damping of the jaw, upper lip, and lower lip) are also functions of the evolving articulatory configuration.

2.2.4. Example 1: discrete bilabial closures: unperturbed gestures
Given a fixed set of dynamic parameter values for the tract variables of lip aperture and lip protrusion, and a set of initial positions and velocities for the jaw, upper lip and lower lip, the equations of motion for the model articulators will generate a pattern of coordinated articulatory movements that will achieve the task goal (e.g. bilabial closure) specified for the tract variables. For an initial configuration corresponding to open
and relatively unprotruded lips, and with initial articulator velocities of zero, these coordinated movements will reflect the evolving task-specific motions of the tract variables en route to their specified targets, with motion characteristics (e.g. speed, degree of overshoot, etc.) specified by the pattern of tract-variable dynamic parameter values. Assuming the system is not perturbed during its motion trajectory, the relative extents of movements for the jaw and lips will be specified by the relative values of the associated articulator weightings. Thus, one weighting pattern might correspond to predominant jaw motion, while a second weighting pattern might correspond to predominant vertical motion of the lips for a given lip aperture trajectory.

Figure 2 (configurations a and b) illustrates an unperturbed movement from an initially open and relatively unprotruded configuration (Fig. 2a) to a closed and relatively protruded final configuration (Fig. 2b). Since the articulators associated with lip aperture were weighted equally in the corresponding weighting matrix, the extents of motion for these articulators were equal over the course of the gesture.

2.2.5. Example 2: bilabial closure, immediate compensation, perturbed gestures
As discussed in Section 2.1, Kelso et al. (1984b) demonstrated that if the jaw was retarded en route to a bilabial closure for /b/, the closure was still attained and the final articulatory configuration for the perturbed movement was different from the final configurations for unperturbed movements. Significantly, upper lip compensation was absent if the jaw was perturbed en route to an alveolar closure for /z/. These results show that an invariant dynamic description of a movement does not apply at the articulator level, since the articulatory-dynamic parameters (e.g. rest-configuration) must be able to change according to a movement’s context in an utterance-specific (i.e. /b/ vs. /z/) manner. Furthermore, the speed of these compensatory behaviors suggests that they must occur “automatically” without reference to traditional stimulus–response reaction time correction procedures.

The task-dynamic model handles such immediate compensation as follows. Bilabial closure gestures are simulated as discrete movements toward target constrictions, using point attractor dynamics for the local tract variables of lip aperture and protrusion. When the simulated jaw is “frozen” in place during the closure gesture at the level of the model effector system, the main qualitative features of the perturbation data are captured, in that (a) compensation to the jaw perturbation is immediate in the upper and lower lips, i.e. the system does not require reparameterization in order to compensate; and (b) the target bilabial closure is reached (although with different final articulator
configurations and, hence, different jaw-space locations for the closure) for both perturbed (Fig. 2c) and unperturbed (Fig. 2b) "trials".

2.2.6. Example 3: coarticulation, gestural coproduction, bilabial and tongue dorsum gestures

In the task dynamic model, coarticulatory effects may originate in two ways. Passive carryover effects that are due to inherent system "sluggishness" (i.e. the time constants of the different tract variables; see also Henke, 1966; Coker, 1976) are implicit in the functioning of the model. Additionally, and more interestingly, other coarticulatory effects (both anticipatory and carryover) result from the temporally overlapping demands (conflicting or synergistic) made by the same or different tract variables on a common articulator subset (e.g. bilabial and tongue dorsum gestures with reference to the shared jaw articulator). We have begun to model these latter "active" coarticulatory effects by using the articulatory synthesizer to define articulator subsets for two new tract variables associated with the vocal tract constriction formed by movements of the tongue body dorsum. Thus, constriction location for tongue body dorsum is associated with the articulatory degrees of freedom of jaw rotation, and radial and angular displacements of the tongue body relative to the jaw; constriction degree for tongue body dorsum is associated with the same articulatory subset.

In preliminary simulations, we modeled the hypothetical case in which bilabial and tongue dorsum gestures either do not overlap in time or were totally synchronous. Both gestures were identical in durational and damping factors, and all articulators had equal weightings. For both gesture types in the non-overlapping case, the model articulators started at the same initial "neutral" configuration (corresponding to slightly open lips and a schwa-like position for the tongue dorsum), and attained their respective bilabial closure and tongue dorsum constriction targets. The final articulatory configurations were different for both gesture types and, in particular, the final jaw position for the single bilabial gesture was higher than that for the single tongue dorsum gesture. Recalling previous discussions in this section, these final configurational (and jaw positional) differences resulted from the different ways that the instantaneous, evolving task space forces were distributed across each gesture type's articulatory subset during the course of the movement. Roughly speaking, if we focus on the net "force" distributed to the jaw during the movement, we can say that more net force was delivered to the jaw during the simulated bilabial than during the tongue dorsum gesture, resulting in greater and lesser jaw displacements, respectively. Starting from the same initial configuration but with synchronous gestures, both the bilabial and tongue dorsum targets were again reached. However, the final articulatory configuration was different from those observed when either of the gestures occurred in isolation. The final jaw height for the gesturally synchronous case was halfway between the final jaw positions attained for the nonoverlapping gestures. This compromise jaw position resulted from the fact that, in the model, the net force delivered to the jaw over the gesturally synchronous movement was (roughly) the weighted average of the net jaw forces delivered during each of the nonoverlapping gestures.

We are extending our simulations currently to include cases in which different gestures overlap only partially in time (a more realistic assumption with reference to speech coarticulatory phenomena). In these cases, the net force distributed to the jaw (and hence total jaw displacement) during periods of gestural overlap will reflect the weighted averages of the jaw forces associated with each gesture over these periods. The predicted behavior of the model is consistent, in fact, with coarticulation data for V1CV2
utterances presented by Sussman, MacNeilage & Hanson (1973). As a first approximation toward modeling such utterances, we will treat bilabial consonants as closing gestures of a lip-jaw system associated with the tract variables for bilabial constrictions. Similarly, we will treat vowels as opening gestures of the jaw–tongue system associated with the tract variables for tongue dorsum constrictions. We realize, of course, that this description represents only a preliminary, simplified account of the data which will be modified as experiments and simulations progress. For example, at least the early portions of consonantal release gestures appear to depend on the manner class (e.g. stops vs. fricatives) of the consonants themselves. However, given these assumptions, we may represent the V1C/V2 productions as temporally overlapping sequences of opening (vocalic) and closing (consonantal) tract variable gestures. Since the vowel and consonant gestures share the jaw as a common articulator, the net movement of the jaw during periods of gestural overlap (i.e. the period of jaw motion during which the V1C closing gesture overlaps the CV2 opening gesture) will be determined by the weighted average of the respective “demands” made on the jaw by each gesture during these periods. Hence, for example, the vertical upward displacement of the jaw for a V1C gesture (and hence, the jaw height at closure) will be influenced by the height of V2. Specifically, the net upward demand or “force” delivered to the jaw for low V2 (/ɛ/) will be less during the period of gestural overlap than it would be for high V2 (/i/), and should generate the anticipatory coarticulatory effect of greater V1C displacement for high V2 than for low V2 observed by Sussman et al. (1973).

3. On gestural orchestration: from relative timing to phase stability

In the previous section we focused on the intrinsic properties of functional units of action, but have not discussed the sequencing or orchestration of these units over time. One way to explore the processes underlying such orchestration is to transform a given action pattern as a whole (e.g. by scaling on movement rate, amplitude, etc.) and search for what remains stable across the transformation.

Much evidence now exists that the relative timing of movement events is stable across certain scaling changes and hence provides a more appropriate metric than their absolute durations. Although early demonstrations of relative temporal stability were provided from activities that are qualitatively repetitive and potentially pre-wired (e.g. locomotion, respiration, and mastication; see Grillner, 1977, for review), more recent work has revealed that less repetitive activities show similar organizational features (e.g. two-handed movements, typing, handwriting, postural control and speech–manual coordination; Nashner, 1977; Kelso, Southard & Goodman, 1979; Lestienne, 1979; Viviani & Terzuolo, 1980; Hollerbach, 1981; Shapiro, Zernicke, Gregor & Diestel, 1981; Schmidt, 1982; Kelso, Tuller & Harris, 1985). Importantly, there is some limited evidence that the production of speech can be described by a similar style of organization, and we will now describe this work in some detail.

In a set of previous experiments (Tuller, Kelso & Harris, 1982, 1983, Tuller & Kelso, 1984; Harris, Tuller & Kelso, 1986) Tuller and colleagues have shown that, across variations in speaking rate and stress, the timing of articulatory events associated with consonant production remains stable relative to the interval between events associated with flanking vowels. Consider a very simple, but paradigmatic case in which the latency (in ms) of onset of upper lip motion for a medial consonant is measured relative to the interval (in ms) between onsets of jaw motion for flanking vowels. In Fig. 3(a), we see
the particular intervals measured for one token of the utterance /baPAB/, spoken at a conversational rate with primary stress on the second syllable. The movement data were obtained by recording from infrared LEDs attached to the subject’s lips and jaw. Here, the interval from V1 to V2 represents the time between onsets of jaw motion for successive vowels. The interval V1–UL represents the latency of onset of medial consonant-related movement in the upper lip. These points were obtained from zero crossings of velocity traces. The main empirical question was: Do the intervals V1–V2 and V1–UL change in a systematically related way as syllable stress and speaking rate vary?

Figure 3(b), taken from Tuller & Kelso (1984), plots the latency of upper lip movement relative to the vowel period for one of the four speakers. The data were similar for all subjects. The utterance shown, /baPAB/, spoken at two rates and with two stress patterns illustrates the main result, viz. that over changes in speaking rate and stress, the measured temporal intervals and articulatory displacements change considerably but the relative timing is preserved. The overall relationship can be described by a linear function defined by two parameters—a positive slope and a non-zero intercept. This high correlation of two event durations across rate and stress in different speakers has since been
replicated by other investigators (Bladon & Al-Bamerini, personal communication; Linville, 1982; Lubker, 1983; Gentil, Harris, Horiguchi & Honda, 1984; Munhall, 1985).

How is this stability of relative timing to be rationalized? A popular view in the motor control literature is that time is metered out by a central program that instructs or commands the articulators when to move, how far to move and for how long (e.g. Schmidt, 1982). However, a reconceptualization by Kelso & Tuller (1985a/in press) and subsequent real analysis of the original data (Tuller & Kelso, 1984) strongly suggest that their findings can be understood without recourse to an extrinsic timer or timing metric. In fact, a very different view of articulatory “timing” emerges when the articulatory movements are re-analysed as trajectories on the phase plane. These phase plane trajectories provide a geometric or kinematic description that usefully captures the forms of patterned motion produced by the articulators. A brief tutorial follows.

3.1. The phase portrait: a tutorial (cf. Kelso, Tuller & Harris, 1984a, 1986)

All possible system states can be represented in the phase plane, whose axes are the articulator’s position (x) and its velocity (ẋ). As time varies, the point P (x, ẋ) describing the motion of the articulator moves along a certain path on the phase plane. Figure 4 illustrates the mapping from time domain to phase plane trajectories. Hypothetical jaw and upper lip trajectories (position as a function of time) are shown for an unstressed /bab/ [Figure 4(a) left] and a stressed /bab/ [Fig. 4(b) left]. On the right are shown the

3A preliminary report of these data was given by Kelso & Tuller (1985b).
corresponding phase plane trajectories. In this figure, and those following, we have reversed the typical orientation of the phase plane so that position is shown on the vertical axis and velocity on the horizontal axis. Thus, downward movements of the jaw are displayed as downward movements of the phase path. The vertical crosshair indicates zero velocity and the horizontal crosshair indicates zero position (midway between minimum and maximum displacement). As the jaw moves from its highest to its lowest point (from A to C in Figure 4), velocity increases (negatively) to a local maximum (B) then decreases to zero when the jaw changes direction of movement (C). Similarly, as the jaw is raised from the local vowel /a/ into the following consonant constriction, velocity peaks approximately midway through the gesture (D) then returns to zero (A).

Phase plane trajectories preserve some important differences between stressed and unstressed syllables. For example, maximum lowering of the jaw for the stressed vowel is greater than lowering for the unstressed vowel and maximum articulator velocity differs noticeably between these two orbits (e.g. Kelso et al., 1985; MacNeilage et al., 1970; Stone, 1981; Tuller, Harris & Kelso, 1982). In contrast, the different durations taken to traverse the orbit as a function of stress are not represented explicitly in this description. That is, although time is implicit and usually recoverable from phase plane trajectories, it does not appear explicitly.

It is possible to transform the Cartesian $x, \dot{x}$ coordinates into equivalent polar coordinates, namely a phase angle, $\phi = \tan^{-1}([y/x])$, and a radial amplitude, $R = [x^2 + \dot{x}^2]^{1/2}$. These polar coordinates are indicated on the phase planes shown in Fig. 4. The phase angle has been a key (computed) dependent variable in our re-analysis of interarticulator timing. It allows us to rephrase the traditional question of how the lip knows when to begin its movement for the medial consonant by asking where on the cycle of jaw states that the lip motion for medial consonant production begins. One possibility is that lip motion begins at the same phase angle of the jaw across different jaw motion orbits (i.e. across rate and stress). This outcome is not necessarily entailed, or predicted by, the relative timing results. For example, Fig. 4(a)–(c) shows three utterances whose vowel-to-vowel periods and consonant latencies do not change in a linearly related fashion. Nevertheless, the phase angle at which upper lip motion begins relative to the cycle of jaw states is identical in the three cases. Thus, the information for “timing” of a remote articulator (e.g. the upper lip) may not be time itself, nor absolute position of another articulator (e.g. the jaw), but rather a relationship defined over the position-velocity state (or, in polar coordinates, the phase angle) of the other articulator. Although this conceptualization is intriguing, we want to re-emphasize that it constitutes an alternative description of the relative timing data set. For example, Fig. 5 illustrates the converse of Fig. 4, namely, that two (hypothetical) utterances with identical vowel-to-vowel periods (P) and consonant latencies (L) can nonetheless show very different phase angles for upper lip movement onset. To be specific, the phase angle analysis incorporates the full trajectory of motion; the relative timing analysis is independent of trajectory once movement has begun and is based on the onsets and offsets of movement events.

*Note that there is an important caveat for the phase notion. Though phase has been illustrated here at a very “simple” interarticulator level, we do not want to suggest that this is necessarily the appropriate frame of reference for speech production and perception. However, the point is that regardless of the particular frame of reference (for example, events defined at muscle, articulator, tract variable levels, etc.) a concept such as phase will be crucial to specifying the sequence of events.*
Dynamical perspective on speech production

Figure 5. Two hypothetical utterances having identical vowel-to-vowel periods (P) and consonant (upper lip) latencies (L) but different phase angles of upper lip onset. (See caption Fig. 4.) (From Kelso & Tuller, 1985a.)

Figure 6. Left: jaw cycle on the phase plane for the first token produced of stressed /'baaab/ (top) and unstressed /ba#b/ (bottom), spoken at a fast rate. Each token shown is the first instance produced of the utterance type. On the left is the entire jaw cycle for each stress pattern; on the right, the jaw cycle is reproduced only until the point of onset of upper lip movement downward for production of the medial labial consonant, as measured from the first deviation from zero velocity. The calculated phase angle\(^5\) at which upper lip

\(^5\)The computation of the phase angle of a point on the phase plane is problematical. Because the units of position and velocity are incommensurate (mm vs. mm/s, for example), applying inverse trigonometric functions directly to the data yields meaningless results. To avoid this problem, we normalize both position and velocity to the same numerical interval, \(-1\) to \(+1\), (not necessarily the unit circle for periodic data) and then apply the inverse tangent function to the normalized data. The normalization of position over a cycle of data proceeds via the following linear transform:
motion begins is indicated for each token. Notice that the jaw displacement and velocity are both greater for the stressed than the unstressed syllable. Nevertheless, upper lip motion begins at essentially the same phase angle for both tokens. If upper lip motion began at a phase angle of 180°, it would be synchronous with the jaw “turnaround” point.

3.2. New results

Table I shows the mean data and the standard error of the mean for all four speakers from the Tuller & Kelso (1984) study. Analyses of variance (2×2) for each utterance type showed no significant main effects of rate or stress or their interaction on the phase angle of upper lip onset for medial consonant production. For /babab/, $F(1,27)$ ranged from 0.02 to 2.97; for /babab/, $F(1,30)$ ranged from 0.01 to 2.39; for /bawab/, $F(1,29)$ ranged from 0.01 to 2.80, giving $p$ values greater than 0.1. Although phase angle was invariant across speaking rate and stress, Table I also shows some differences in phase angle as a function of the medial consonant. There is some tendency for upper lip phase for /p/ to be smaller than /b/. This result may be consistent with acoustic findings that vowels are longer before voiced than voiceless consonants. There is also a strong tendency for the upper lip phase of /w/ to be greater than the stops. However, this result could be artifactual: the movement measures did not include a horizontal component (potentially larger for /w/ than /b/ or /p/). In addition, our subjects produced /w/ with much smaller and slower upper lip movements, making measurement of movement onset more difficult.

3.3. Empirical and theoretical implications

There are at least two empirical advantages of these phase angle analyses over our previous relative timing descriptions. First, in the relative timing analysis, the overall correlations across rate and stress conditions were very high, but the within-condition slopes tended to vary somewhat. In the phase analysis, on the other hand, the mean phase angle is the same across conditions. Second, recall that the relative timing data were fitted by linear functions described by two parameters. The phase description requires only a single parameter and, if nothing else, is the more parsimonious description.

The phase angle conceptualization also offers a number of theoretical advantages over the original relative timing analyses. First, once articulatory motions are represented

$$P_{\text{norm}} = 2P/(P_{\text{max}} - P_{\text{min}}) = (P_{\text{max}} + P_{\text{min}})/(P_{\text{max}} - P_{\text{min}}),$$

where $P_{\text{norm}}$ is the normalized position, $P$ is the actual position, and $P_{\text{max}}$ and $P_{\text{min}}$ are the maximum and minimum position values over a cycle.

This has the effect of (i) rescaling the data to a range of two units and (ii) shifting the equilibrium position to zero, i.e., the interval −1 to +1 is achieved. Velocity is normalized according to which half-cycle the point of interest is in, to put the half-cycle of articulator raising on the interval 0 to 1 and the half-cycle of articulator lowering on the interval 0 to −1. That is,

$$V_{\text{norm}} = V/\text{abs}(V_{\text{max}}),$$

where $V_{\text{norm}}$ is the normalized velocity, $V$ is the actual velocity, and $V_{\text{max}}$ is the maximum velocity during the corresponding half-cycle.

The arctangent was then computed using the normalized position and velocity values,

$$\text{phase angle} = \arctan(V_{\text{norm}}/P_{\text{norm}})$$

The final value obtained is a number from 0 to 360°, which increases in value in the direction opposite to the unwinding of trajectories on the phase plane, as per mathematical convention.
Table I. Mean upper lip phase (± SE) relative to vowel-to-vowel jaw trajectory for subject JE, NM, BT, and CH.

<table>
<thead>
<tr>
<th></th>
<th>/baba/</th>
<th></th>
<th>/bapa/</th>
<th></th>
<th>/bawa/</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lip phase</td>
<td>SE</td>
<td>Lip phase</td>
<td>SE</td>
<td>Lip phase</td>
</tr>
<tr>
<td><strong>JE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS*</td>
<td>212</td>
<td>7.05</td>
<td>184</td>
<td>6.41</td>
<td>207</td>
</tr>
<tr>
<td>SU</td>
<td>205</td>
<td>2.83</td>
<td>183</td>
<td>2.07</td>
<td>205</td>
</tr>
<tr>
<td>FS</td>
<td>197</td>
<td>2.83</td>
<td>177</td>
<td>3.82</td>
<td>212</td>
</tr>
<tr>
<td>FU</td>
<td>203</td>
<td>4.26</td>
<td>179</td>
<td>3.81</td>
<td>203</td>
</tr>
<tr>
<td><strong>NM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>182</td>
<td>2.0</td>
<td>178</td>
<td>2.35</td>
<td>193</td>
</tr>
<tr>
<td>SU</td>
<td>178</td>
<td>2.74</td>
<td>175</td>
<td>3.21</td>
<td>193</td>
</tr>
<tr>
<td>FS</td>
<td>184</td>
<td>3.14</td>
<td>176</td>
<td>1.73</td>
<td>197</td>
</tr>
<tr>
<td>FU</td>
<td>183</td>
<td>3.49</td>
<td>172</td>
<td>2.50</td>
<td>189</td>
</tr>
<tr>
<td><strong>BT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>168</td>
<td>2.76</td>
<td>163</td>
<td>2.93</td>
<td>196</td>
</tr>
<tr>
<td>SU</td>
<td>168</td>
<td>5.83</td>
<td>174</td>
<td>3.50</td>
<td>198</td>
</tr>
<tr>
<td>FS</td>
<td>166</td>
<td>4.58</td>
<td>166</td>
<td>2.68</td>
<td>192</td>
</tr>
<tr>
<td>FU</td>
<td>164</td>
<td>3.36</td>
<td>167</td>
<td>3.96</td>
<td>191</td>
</tr>
<tr>
<td><strong>CH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>184</td>
<td>4.38</td>
<td>188</td>
<td>6.45</td>
<td>203</td>
</tr>
<tr>
<td>SU</td>
<td>186</td>
<td>3.93</td>
<td>184</td>
<td>3.01</td>
<td>208</td>
</tr>
<tr>
<td>FS</td>
<td>181</td>
<td>6.38</td>
<td>183</td>
<td>3.94</td>
<td>207</td>
</tr>
<tr>
<td>FU</td>
<td>182</td>
<td>3.16</td>
<td>177</td>
<td>4.08</td>
<td>196</td>
</tr>
</tbody>
</table>

*SS = Slow (normal) speaking rate, first syllable stressed; SU = Slow (normal) speaking rate, first syllable unstressed; FS = Fast speaking rate, first syllable stressed; FU = Fast speaking rate, first syllable unstressed.

geometrically on the phase plane, the phase angle serves to normalize duration across speaker, stress, speaking rate, etc. Second, these analyses potentially provide a grounding for so-called intrinsic timing theories of speech production (e.g. Fowler, 1980; Fowler et al., 1980), since neither absolute nor relative durations need be monitored or controlled extrinsically, and no time-keeping mechanisms or time controllers are required in this formulation. As with a candle (which provides a metric for time by a change in its length) or a water clock (where the metric is number of drops), the units of time are defined entirely in terms of the dynamical processes involved. Time itself is not a fundamental variable, and is not likely to be a possessed, programmed or represented property of the speech production system (Kelso et al., 1984a; Kelso & Tuller, 1985a). As an aside, it has never been clear how the speech system could keep track of time, at least peripherally, because there is no known afferent basis (such as time receptors) for time-keeping in the articulatory structures themselves (Kelso, 1978). On the other hand, an informational basis (e.g. in position and velocity sensitivities of muscle spindle and joint structures) is a physiological given in the phase angle characterization. It might well be the case that certain critical phase angles provide information for coordination between articulators (beyond those considered here) and/or vocal tract configurations, just as phase angles of the leg joints provide coupling information for locomotory coordination (Shik & Orlovskii, 1965). Third, as Fowler (1980) notes, a dominant assumption of what she calls extrinsic timing theories of coarticulation is that
phonological segments are considered to be discrete in the sense that their boundaries are straight lines perpendicular to the time axis. Yet, as is well known, discrete segments are not seen by perpendicular cuts of the physical records of speech (acoustic, kinematic, physiological measurements) along the time axis. In the phase angle analysis, however, no a priori assumptions are made regarding the issue of segmentation per se, and the overlap (or coproduction) among gestures is captured in a natural way while still preserving a separation between consonantal and vocalic events. A final implication of the view presented here is that “segments” or phonological units as typically defined by linguists may not be relevant to the speech production system. Rather, phonological units might be profitably reconceptualized in terms of characteristic interarticulator phase structures (see also Browman & Goldstein, in press, for related notions). Note that the phase structure description minimizes the mind/body problem for speech production by avoiding the translation step between psychological planning units and the physical execution of those units. On the other hand, different issues are immediately raised—such as whether there are a restricted number of stable phase structures (which one might expect if they are to be tagged with linguistic descriptors), and if so, why some configurations appear and not others. Experimental inroads into these issues can be made in the present perspective with a minimum of ad hoc assumptions, and with little resort to a priori linguistic categories.

4. Instabilities: nonequilibrium phase transitions and phonetic change

The phase analysis of simple speech utterances indicated that certain phase relations among the articulators remain unaltered across manifold speaker characteristics. Such critical phase angles are revealed by the flow of the dynamics of the system; they are not externally defined. As Sleigh & Barlow (1980) note in their comparative analysis of creatures that use a wide variety of propulsive structures for their activities, phase appears to provide essential information for stable coupling among the components of the system. How, then, do we conceive of the processes underlying change in articulatory pattern? What factors mediate the emergence of new (or different) spatiotemporal patterns? Such questions are at the heart of a theory of pattern generation. Below we offer an interpretation of certain kinds of articulatory (and phonetic) change in terms of the non-equilibrium phase transitions treated by synergetics (Haken, 1975, 1977; Haken et al., 1985). The central aspects of the theoretical model will be introduced briefly using an example from hand movements, and an application to speech will follow that focuses primarily on the effects of scaling changes in speaking rate, one of Stetson's (1928/1951) “great causes of phonetic modification” (p. 67). Importantly, the present analysis attests to the further significance of phase information—both within the stable and transition regions of the speech system’s parameter space—in guaranteeing phonetic stability on the one hand and promoting phonetic change, on the other. Moreover, along with other kinds of data (primarily on phonological development) the form of articulatory change may help rationalize a particular phonological unit, as a “natural” or intrinsically stable unit in the production of speech.

4.1. A synergetic outline: pattern formation and change

For some years now, we have advocated an approach in which the control and coordination of multidegree of freedom speech and limb activities are treated in a manner...
continuous with cooperative phenomena in other physical, chemical, and biological systems, i.e. as synergetic or dissipative structures (e.g. Kelso et al., 1981, 1983; Kugler et al., 1980; Kelso & Saltzman, 1982). These are systems—like that for speech production—that are composed of very many subsystems. In synergetics, when a certain parameter or combination of parameters (generally referred to as “controls”) are scaled in sometimes quite non-specific ways (i.e. the control prescription is not highly detailed), well-defined spatiotemporal patterns can form. The latter are maintained by a continuous flux of energy (or matter) through the system (e.g. Yates, Marsh & Iberall, 1973; Haken, 1975). Although there is pattern formation in the nonequilibrium phenomena treated by synergetics (e.g. the hexagonal forms produced in the Bénard convection instability, the transition from incoherent to coherent light waves in the laser, the oscillating waves and macroscopic patterns of various kinds of chemical reaction, etc.) there are, strictly speaking, no special mechanisms—that contain or represent the pattern before it appears (for further examples see Kelso & Tuller, 1984a).

How pattern formation occurs in these systems can be visualized roughly as follows. Imagine an open, dissipative system, one into which energy is continuously fed and from which it is continually dissipated. Certain configurations, called modes, are more capable of absorbing the energy flow than others. At a critical point, a linearized stability analysis reveals that the amplitude of these so-called unstable modes grows exponentially whereas the other modes (the so-called “damped” modes) decay. In many non-equilibrium systems, close to critical (or bifurcation) points, the number of unstable modes can be shown to be much smaller than the number of stable, damped modes. In fact, the latter can be completely eliminated mathematically, according to Haken’s so-called slaving principle, thereby allowing a tremendous reduction of the degrees of freedom. For example, in the laser (see Haken, 1975), a reduction from $10^{14}$ degrees of freedom to a single degree of freedom has been obtained.

More formally, the slaving principle states that the amplitudes of the damped modes can be expressed by means of a small set of “unstable” mode amplitudes (the so-called order parameters). The consequence is that all the damped modes follow the order parameters adiabatically, so that the behavior of the whole system is then governed by the order parameters alone (see Haken, 1977, Chapter 7). Watching Bénard convection, for example, one is impressed how the total behavior—at a critical point—is completely captured by a macroscopic, modal action. The motions of the many microscopic, molecular components are completely irrelevant at this point: a low dimensional, macroscopic observable (the order parameter) specifies the system’s evolving pattern.

However, identifying order parameters, even for many physical and chemical systems is not always an easy matter. Certain guidelines do exist, however, which can be used to select viable candidates. A main one is that the order parameter changes much more slowly than the subsystems it is said to govern. Relative phase fits this criterion quite well, since it is the phasing structure of many different activities that is preserved across scalar transformations (Section 3). Thus the individual articulatory components change.

---

4A key feature in the development of science has been to define limits or constraints on natural phenomena. Once such constraints are known, much new understanding results (Prigogine & Stengers, 1984). Phase has this constraint-like property. Our experiments described in Section 4 reveal the limits over which one organization (a given phase relation) can remain stable. Also, because in those experiments it is phase (and phase alone, as far as we know) that changes dramatically, we have some reason to suppose that phase is a key parameter even in the stable range of performance, i.e. that phase represents a fundamental constraint (see Section 3).
quite a bit (kinematically and electromyographically) but the phase does not—at least in a given region of the parameter space.

4.2. Phase transitions in movement: an explicit example

Using relative phase as an order parameter, Haken et al. (1985) have offered an explicit theoretical model of phase transitions that occur in bimanual activity (see Kelso, 1981, 1984). The basic phenomenon is as follows: A human subject is asked to cycle his/her fingers at a preferred frequency using an out-of-phase, antisymmetrical motion. That is flexion [extension] of one hand is accompanied by extension [flexion] of the other. Under an instruction to increase cycling frequency, i.e. a systematic rate increase, the movements shift abruptly to an in-phase, symmetrical mode involving activation of homologous muscle groups. When the transition frequency was expressed in units of preferred frequency, the resulting dimensionless ratio or critical value was constant for all subjects but one (who was not naive and who purposely resisted the transition—although with certain energetic consequences, see Kelso, 1984). A frictional resistance to movement lowered both preferred and transition frequencies, but did not change the critical ratio (~1.33), suggesting the presence of an intrinsic invariant metric.

For present purposes, the main features of the bimanual experiments are (1) the presence of only two stable phase (or "attractor") states between the hands (see also Yamanishi, Kawato & Suzuki, 1980 for further evidence); (2) an abrupt transition from one attractor state to the other at a critical, intrinsically defined frequency; (3) beyond the transition, only one mode (the symmetrical one) is observed; and (4) when the driving frequency is reduced, the system does not return to its initially prepared state, i.e. it remains in the basin of attraction for the symmetrical mode.

The theoretical strategy employed by Haken et al. (1985) to account for the foregoing findings may be worth noting. First, they specified a potential function corresponding to the layout of modal attractor states (i.e. the stable in-phase and out-of-phase patterns), and showed how that layout was altered as a control parameter (driving frequency) was scaled. From the behavior of the potential function, they then derived the equations of motion for each hand, and a nonlinear coupling structure between the hands. Analytic derivations and consequent numerical simulation revealed that if the model system was started, or "prepared" in the out-of-phase mode, and driving frequency was increased slowly, the system remained in that mode until the solution of the coupled equations of motion became unstable. At this point, a jump occurred and the only stable stationary solution produced by the system corresponded to the in-phase mode (see Haken et al., 1985, for more details). Ongoing theoretical (Schöner, Haken & Kelso, 1976) and empirical (Kelso & Scholz, 1985) work has revealed that the non-linear coupling strength as well as fluctuations (both intrinsically generated due to noise in system parameters and extrinsically generated due to an added random forcing function) play an important role in effecting the modal transitions between the hands. Thus, Kelso & Scholz (1985), in new experiments, have found both "critical slowing down" and enhanced fluctuations in order parameter behavior as the transition is approached. These predictions follow directly from the synergetic treatment of nonequilibrium phase transitions (see, for example, Haken, 1984; Haken et al., 1985; Schöner et al., 1986) and are simply not part of more conventional accounts of "switching" behavior based on motor programs (cf. Schmidt, 1982, p. 316) or central pattern generators (cf. Grillner, 1982, p. 224).
4.3. Stetson's (1951) experiments

Let us now see how this view of spatiotemporal pattern formation and change may apply to speech production. To do this we draw initially on Stetson's (1951) work and offer a theoretical interpretation of his experiments that is consistent with synergetics. Then we mention some new (as yet preliminary) data of our own (Kelso, Munhall, Tuller & Saltzman, in preparation) suggesting that certain kinds of phonetic change correspond directly to phase transitions among articulatory gestures.

Stetson (1951) recognized that "... the modifications of the articulations is one of the most important aspects for study in experimental phonetics" (p. 67), and that scaling changes in speaking rate offered a window into the "various types of modification of the factors of the syllable, or the changing conditions that throw them together or force them apart" (p. 67). We also hypothesize—by analogy to our discussion above—that rate changes may (a) reveal the most stable modes of coordination of the articulatory system and, in turn, (b) that these stable modes may rationalize why one phonological form, the CV syllable, tends to be a universal feature of all languages (cf. Abercrombie, 1967; Bell, 1971; Clements & Keyser, 1983).

Consider first an example, discussed in some detail by Stetson (1951). A subject produces the CVC syllable "pup" repetitively. As speaking rate is gradually increased, Stetson (1951) describes the following changes: The syllables, "pup, pup, ... " at first distinct, come closer together. As rate increases the arresting consonant of each syllable "doubles" with the releasing consonant of the next syllable. Thus the first change can be annotated as: "pup, pup" ... → "pup-pup, ...". At still higher rates, according to Stetson, it becomes impossible to execute the prescribed number of consonants per second, and the arresting consonant of each syllable drops out. This second change can be referred to as "singling"; "pup-pup, ... " goes to "pu' pu'. ...".

Such changes induced by increasing speaking rate are brought about, in Stetson's words, by the tendency of movements "either to get into step or to drop out in order to simplify the coordination" (p. 71) and, relatedly, because of a "universal tendency to simplify by eliminating the arresting consonant" (p. 81). But why this particular tendency should prevail is unclear. On the one hand, elimination of the arresting consonant "simplifies" coordination; on the other, the process is dictated by maximum articulatory rates: "singling" must occur at rates of around 4.5 syllables/s, because such rates in turn entail 8 consonant movements per second (Stetson, 1951).

"Simplification" as a function of maximum articulatory rate cannot be the whole story, of course. For example, often "singling" occurs at a rate as low as 2.5 syllables/s. Also, the arresting consonant does not always drop out; often it is said to "fuse" with the releasing consonant (Stetson, 1951). Therein lies a potential clue. That is, it may be the phasing among component gestures—one with another—that is a central aspect of phonetic stability and change (cf. Section 3). For example, in his studies of the combination of abutting consonants, Stetson notes that the arresting consonant of one syllable (e.g. "p" in "sap") and the releasing consonant of the next (i.e. "s" in "sap"), "quickly overlap and soon become simultaneous ... a striking illustration of the movements of speech to get in phase" (p. 78). Moreover, as rate decreases, the movements of arresting and releasing consonants "merely slide apart" (p. 78). Thus, there is a strong hint in Stetson's experiments and writings that under scaling influences of speaking rate, certain phase relations among gestures are more stable than others. For example, in all "phonetic coordinations" to use Stetson's phrase, there is a preferred relation between
the releasing consonant and the beat stroke of the syllable. Namely, the releasing consonant never drops out. According to Stetson (1951), it retains its position because it coincides with the syllable's beat stroke. In addition, compound consonants are said to be produced by the "sliding" of the two movements, e.g. the continuant labial "m" and the continuant lingual "s" in the syllable "mass" slide to form "sma". Stetson's descriptive language in this respect is almost prophetic of current formalisms: abutting consonants are "attracted" (p. 80) one to the other. As part of the tendency for movements to coincide, one consonant movement is delayed and the other advanced (cf. Stetson, 1951, p. 80).

Though phase was never explicitly measured in Stetson's work, and his account of phonetic change is largely posed within his "chest pulse" framework, there appears to be a strong linkage between his results and our previous discussion of hand movements. In particular, it seems possible that both may fall under the theoretical rubric of nonequilibrium phase transitions. Such a view is supported on at least two grounds. First, in an as yet quite limited data set, we have examined interarticulator phasing when subjects produced the vowel-consonant combination /ip/ at progressively faster rates (Kelso et al., in preparation). A shift to the CV form /pi/, occurred at a given rate and was characterized by an abrupt change in the phase relation between glottal aperture and lip aperture. And second, it seems possible to reinterpret some of Stetson's own data on syllable duration when speaking rate is increased, as consistent with our previous theoretical discussion of phase transitions.

4.4. Phase transitions in speech: some direct evidence

The design of the following experiment was extremely simple. Infra-red light emitting diodes were placed on the subjects' lips and jaw, thus allowing us to obtain the trajectories of these articulators. Similarly, the opening and closing of the glottis was monitored by transillumination (e.g. Baer, Löfquist & McGarr, 1983). In a similar way to some of Stetson's (1951) work, the subject was invited to produce the syllable /ip/ at a slow speaking rate and then instructed to simply speed up in a step-like manner. A complete trial consisted of a series of repetitive syllables produced in a single breath. Typically, a trial lasted about 10 to 12 s. An identical procedure was employed for the syllable /pi/. Subjects performed at least five trials per syllable. Although data collection and analysis are not yet completed (presently three subjects have been run), the data are quite clear thus far. Trajectories over time of lip aperture (i.e. a single variable representing vertical distance between upper and lower lips) and glottal aperture are shown in Fig. 7(1b) and 7(2b) for part of a representative trial for each utterance. In these subfigures the vertical ticks denote the onset of lip opening (consonant release) and the

1The beat stroke as defined by Stetson (1951) is "always ballistic ... and can hardly be longer than 40–100 ms" (p. 29). He continues: "The unit movement of speech is the pulse which produces the syllable, a pulse of air through the glottis made audible by the vocal folds in speaking aloud and stopped and started by the chest muscles or by the auxiliary movements of the consonants" (p. 30). And later on in a discussion of consonant release, Stetson indicates that "... the stroke of the expiratory chest muscles and the beat stroke of the consonant occur at the same time" (p. 46). We include this definition and clarification of "beat stroke" for mostly historical reasons. Some of Stetson's claims about syllable pulses have been seriously questioned (Ladefoged, Draper & Whitteridge, 1958). The present analysis, of course, does not rely on such notions.

2Lip aperture was estimated simply by subtracting the position of the lower lip from that of the upper lip. Note, however, that the same data pattern was obtained when the movement of a single labial articulator (e.g. the lower lip) was compared to glottal aperture.
occurrence of peak glottal opening (i.e. maximum vocal fold abduction). The corresponding relative phase between the lip and glottal aperture motions is shown in Fig. 7(1a) and 7(2a). The movements shown were sampled at 200 Hz (for details of signal processing techniques, see Kay, Munhall, Vatikiotis Bateson & Kelso, 1985).

In the case of both /ip/ and /pi/ it is quite obvious that the phase relation between lip opening onset and peak glottal opening is practically invariant, but different for the two syllables, over the range of speaking rates examined (approximately 1 to 5 syllables/s). For /pi/, peak glottal openings lag the onset of oral opening by a constant amount, roughly 40–50°. For /ip/ the two events are almost coincident, up to a speaking rate of approximately 4 syllables/s. Then, a clear jump in phase occurs, practically within a single cycle, to the phasing pattern for /pi/. Note that like the hand movement data, the phase transition occurs as well below maximum syllable rates, at least for CV syllables. Again, both forms of coordination are quite stable below the critical region: only the coordinative mode characteristic of the CV, however, exists beyond the transition. Except for the quantitatively different phase relations observed, these speech data mimic the pattern of results observed in the bimanual data.

Several issues remain to be addressed, however. First, we need to know much more about what goes on in the region of the transition itself. Second, a continuous estimate of relative phase should be obtained (see Kelso & Scholz, 1985). The point estimate presented here requires the rather arbitrary selection of peaks or valleys in the time-series.
data as reference and/or target events. Given previous work on laryngeal-oral coordination (e.g. Löfqvist & Yoshioka, 1981) the selection of the present events (i.e. lip opening onset and peak glottal opening) seems reasonable. Obviously different events, for example, the onset of movement toward oral closure relative to peak glottal opening, would yield the same pattern but different phase values. A continuous estimate, based on a sample-by-sample phase difference, would not require one to make such a choice a priori. Third, Stetson (1951) reports that the original form restores with a decrease in rate, i.e. the VC form returns and that "... this tendency to restoration ... is the great conservative factor in pronunciation" (p. 74). We suspect otherwise, though we have yet to formally check our suspicions. That is, once a transition to the CV form occurs, the system exhibits hysteresis—it tends to remain in the currently displayed form. If this is so, then we have a model of phase transitions in speech that is formally equivalent to that of Haken et al. (1985) developed for phase transitions in hand movements.

4.5. Theory and theoretical implications

A further hint that the discontinuities created by rate scaling are at least consistent with a nonequilibrium phase transition interpretation can be gleaned from Stetson (1951, figure 51). In this figure, whose main features are reproduced here [Fig. 8(a)], distribution curves are presented showing the rates at which "doubling" and "singling" occur when a single syllable is repeated at varying rates. Although "doubling" occurs at rates between 1 and 3.5 syllables/s, a peak for doubling is present at 2.5 syllables/s. Another way to envisage these data is to invert the curves: Two minima are then apparent, one for each of the two articulatory patterns, separated by a local hill or maximum [Fig. 8(b)]. As speaking rate is increased, it becomes increasingly difficult (as indicated by progressively fewer and fewer observations) for a subject to maintain "doubling". Then, at a critical point, a shift to the next "equilibrium" configuration

*Stetson (1951) says little about the instructions given to subjects when speaking rate is reduced. Obviously, with slowing they will be able to say /ip/ below a certain rate. The question is when they do so spontaneously if, for example, they could not hear the consequences of their production.
occurs, corresponding to the “singing” pattern. Analogous to our discussion above, it seems plausible to suggest that (a) the doubling and singling patterns correspond to distinct system modes, each characterized by specific phasing relations among articulatory gestures; and (b) the transition from doubling to singling beyond a critical production rate reflects a system bifurcation.

These speech data of Stetson, therefore, bear a striking resemblance to our speech and hand movement data as well as recent work on locomotor gait transitions (see Fig. 9, reproduced from Hoyt & Taylor, 1981),10 an interpretation of which is given in Kelso & Tuller (1984a) and Kelso & Scholz (1985). In the case of quadruped gait, the modes correspond to particular phasing relations among the limbs, which, when the animal is allowed to locomote freely, correspond to regions of minimum oxygen consumption. Hoyt & Taylor (1981), however, forced ponies to locomote away from these stability regions by increasing the speed of a treadmill on which the ponies walked. That is, according to our interpretation, they experimentally displaced the ponies away from equilibrium. As locomotor velocity is scaled, it becomes metabolically costly for the animal to maintain a given interlimb configuration; a switch into the next stable region, i.e. the next local minimum, occurs (e.g. walking shifts to trotting). Like the hand movement data, when a critical value is reached (a point at which the “forces” driving

10This resemblance is not only qualitative but perhaps quantitative as well. It may be purely serendipity that the ratio of the “doubling” mode frequency (~ 2.5 syllables/s) to the critical frequency (~ 3.1 syllables/s) shown in Fig. 8, bears a close correspondence (~ 1.24) to the dimensionless ratios computed for Kelso’s bimanual (~ 1.31) and Hoyt & Taylor’s gait (~ 1.33) data (see Kelso, 1984). These dimensionless numbers, analogous to Reynolds’ numbers in fluid dynamics, may be a reflection of the system’s intrinsic “distance from equilibrium”. That is, they may index how far beyond a “preferred” steady state a given pattern can persist before it fractures into a new configuration.
through the system—roughly equated with increases in neural activation of muscle groups induced by rate scaling—compete with, and overcome the "forces" holding the system together, i.e. characteristic phase relations) the system bifurcates and a new (or different) spatiotemporal ordering emerges. We want to emphasize that such ordering changes are not strictly fixed for any of these situations. Horses, for example, can trot at speeds at which they normally gallop (as a visit to Yonker's race track to observe the trotters will quickly reveal), but it is metabolically expensive to do so. Similarly "doubling" is possible beyond the bifurcation point in Fig. 8, as illustrated by the dashed line, but there are so few observations there (at rates between 3 and 4 syllables/s), as to suggest that coordination in that region is highly unstable.

In summary, although there are obvious differences between the various critical phenomena discussed here, there is reason to suppose that all of them—hand movement, speech, and gait—correspond to instabilities that arise as the particular system is driven experimentally away from equilibrium.Obviously, much more work needs to be done to ground this conjecture (see Kelso & Scholz, 1985; Schöner et al., 1986, for possible experimental directions). In each case, new stabilities arise—indexed by particular phase relations between the components—as a result of competition between energy flowing into the operational components (i.e. a scaling influence) and the ability of those components to absorb the energy flow in their new configuration. In the hand movement case, and, by hypothesis, in speech as well, we expect that higher bifurcations are possible because the system has available additional degrees of freedom (see Kelso & Scholz, 1985). That is, more configurations are possible—some of which will be stable and others not—precisely because of the availability of these extra degrees of freedom. In this view, the latter are not a curse (cf. Bellman, 1961) but a tremendous advantage. In addition, fluctuations (the "noise" often removed by engineers) can be shown to permit the discovery of new modes or phasing structures (cf. Schöner et al., 1986).

Not only does the present theoretical perspective afford the potential of a principled analysis of pattern formation (for many more details see Haken et al., 1985) but, as we mentioned earlier, the nature of the pattern change itself may prove rather informative. The theory predicts that the states that emerge under scaling influences are the most favored ones, and empirical evidence supports this interpretation. Thus, within the range of driving frequencies examined in the experiments of Kelso and colleagues, shifts to the symmetrical mode of coordination occur but not vice versa. Similarly, in speech, the articulatory configuration supporting the consonant–vowel form is the more fundamental: utterances never shift to the VC form under rate increases when the system is originally prepared in the "CV state". In each case, one structure can be fractured but not the other. By the arguments and data discussed here, this is because certain phase relations among the articulators—which can be modeled as an order parameter for the total articulatory ensemble—are more stable than others.

Clearly both CV and VC forms (like symmetric and antisymmetric hand movements) can be produced easily in a given region of parameter space. The question of which of the two patterns is more basic, is answered by determining which remains beyond a critical point. The fact that the consonant–vowel form "wins out" when the system is scaled is thus a consequence of the stability of the articulatory configuration for that

---

11One difference, for example, between the hand and speech data and the gait analysis is that, in the former, the various modal patterns can coexist in stable forms at subcritical rates. Galloping, on the other hand, is not observed at slow walking speeds. Though it may be available, it is simply not a stable locomotor mode in that region of the parameter space (see Fig. 9).
form. That is, certain configurations can absorb the energy input more efficiently than others. The universal tendency (Stetson, 1951) to simplify coordination by eliminating the arresting consonant (i.e. the one tied to the previous vowel), suggests that it is in some sense “easier” for the system to produce movements “in-phase”, than otherwise. However, this does not have to be the case according to the present thesis. It remains very much an open question—to be pursued empirically—as to which phase relations are more stable than others. In the case of speech, unlike the hand movement case (at least in the most primitive, paradigmatic case studied in our experiments) we would expect a much larger and more varied (but perhaps nested) set of stable phasings. A similar hypothesis applies to studies of phasing in skilled pianists which we are presently analysing. In each case, the layout of the attractor states should be much more “wrinkled” then the “simple” bistable potential—differentiating in-phase and antiphase modes—that we have studied thus far.

In spite of the foregoing caveats, the present analysis may rationalize, in an elegant fashion, why the consonant–vowel is a core syllable type in all languages (cf. Abercrombie, 1967; Clements & Keyser, 1983). Such a rationale has been missing in phonological theory which starts off with the CV core unit as a basic assumption. Moreover the developmental evidence reviewed by Locke (1983) reveals a strong tendency for syllable initial (CV) forms to predominate in infant babbling. Rate scaling studies may reveal these primitive forms of coordination in the mature organism and therefore offer a window into the building blocks of language, a principled decomposition of which has been lacking. Like the particle accelerator that breaks atoms apart to reveal their secrets, so forcing the articulatory system to perform at unusual rates may reveal the primitive units of language and, more important, their interactions with other units. Lest this image be interpreted as too mechanical or immutable let us allay the readers’ concern; nothing could be further from our intent. Just as the cheetah does not have to proceed through the locomotory gaits when it pursues its prey, so the system that realizes language does not have to traverse through any fixed set of phase relations to reveal its intent. What this experimental program may reveal, and this theoretical framework rationalize, is a design that allows for, rather exploits, the low energy switching among its articulatory configurations. In short, a design appropriate for intentional systems.

5. Summary

Presented here, in preliminary form, is a general theoretical framework that seeks to characterize the lawful regularities in articulatory pattern that occur when people speak. A fundamental construct of the framework is the coordinative structure, an ensemble of articulators that functions cooperatively as a single task-specific unit. Direct evidence for coordinative structures in speech is presented and a control scheme that realizes both the contextually varying and invariant character of their operation is outlined. Importantly, the space–time behavior of a given articulatory gesture is viewed as the outcome of the system’s dynamic parameterization, and the orchestration among gestures is captured in terms of intergestural phase information. Thus, both time and timing are deemed to be intrinsic consequences of the system’s dynamical organization. The implications of this analysis for certain theoretical issues in coarticulation raised by Fowler (1980) receive a speculative, but empirically testable, treatment. Building on the existence of phase stabilities in speech and other biologically significant activities, we also offer an account of change in articulatory patterns that is based on the nonequilibrium phase transitions
treated by the field of synergetics. Rate scaling studies in speech and bimanual activities are shown to be consistent with a synergetic interpretation and suggest a principled decomposition of languages. The CV syllable, for example, is observed to represent a stable articulatory configuration in space–time, a possible rationalization for the presence of the CV as a phonological form in all languages. The uniqueness of the present scheme is that stability and change of speech action patterns are seen as different manifestations of the same underlying dynamical principles—the phenomenon observed depends on which region of the parameter space the system occupies. Though probably wrong, ambitious, and the outcome of much idle speculation, the simplicity of the present scheme is attractive and may offer certain unifying themes for the traditionally disparate disciplines of linguistics, phonetics, and speech motor control.

This paper is dedicated to the daughter of the first and third authors, Kathleen Sasha Scott Kelso, born on 7 July 1985. She cannot, however, be held responsible for any of the errors and speculations contained herein. We also thank Kevin Munhall for stimulating discussions on Section 4 and for performing the data analysis of the experiment described in Section 4.4. The authors’ research is supported by NINCDS Program Project NS13617 and BRS Grant RR-05596, by U.S. Office of Naval Research (Psychological Sciences Division) Contract N00014-83-C-0083, and by a grant from The Stuttering Center, Baylor College of Medicine.

References
Dynamical perspective on speech production


