Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 38-68

A Dynamical Approach to Gestural Patterning in Speech Production*

Elliot L. Saltzman and Kevin G. Munhall†

In this article, we attempt to reconcile the linguistic hypothesis that speech involves an underlying sequencing of abstract, discrete, context-independent units, with the empirical observation of continuous, context-dependent interleaving of articulatory movements. To this end, we first review a previously proposed task-dynamic model for the coordination and control of the speech articulators. We then describe an extension of this model in which invariant speech units (gestural primitives) are identified with context-independent sets of parameters in a dynamical system having two functionally distinct but interacting levels. The *intergestural* level is defined according to a set of *activation* coordinates; the *interarticulator* level is defined according to both *model articulator* and *tract-variable* coordinates. In the framework of this extended model, coproduction effects in speech are described in terms of the blending dynamics defined among a set of temporally overlapping active units; the relative timing of speech gestures is formulated in terms of the serial dynamics that shape the temporal patterning of onsets and offsets in unit activations. Implications of this approach for certain phonological issues are discussed, and a range of relevant experimental data on speech and limb motor control is reviewed.

INTRODUCTION

The production of speech is portrayed traditionally as a combinatorial process that uses a limited set of units to produce a very large number of linguistically "well-formed" utterances (e.g., Chomsky & Halle, 1968). For example, /mæd/ and /dæm/ are characterized by different underlying sequences of the hypothesized segmental units /m/, /d/, and /æ/. These types of speech units are usually seen as discrete, static, and invariant across a variety of contexts. Putatively, such characteristics allow speech production to be generative, because units of this kind can be concatenated easily in any order to form new strings. The reality of articulation, however, bears little resemblance to this depiction. During speech production, the shape of the vocal tract changes constantly over time. These changes in shape are produced by the movements of a number of relatively independent articulators (e.g., velum, tongue, lips, jaw, etc.). For example, Figure 1 (from Krakow, 1987) displays the vertical movements of the lower lip, jaw and velum for the utterance "it's a /bamib/ sid." The lower lip and jaw cooperate to alternately close and open the mouth during /bamib/ while, simultaneously, the velum alternates between a closed and open posture. It is clear from this figure that the articulatory patterns do not take the form of discrete, abutting units that are concatenated like beads on a string. Rather, the movements of different articulators are interleaved into a continuous gestural flow. Note, for example, that velic lowering for the /m/ begins even before the lip and jaw complete the bilabial opening from the /b/ to the /a/.

We acknowledge grant support from the following sources: NIH Grant NS-13617 (Dynamics of Speech Articulation) and NSF Grant BNS-8520709 (Phonetic Structure Using Articulatory Dynamics) to Haskins Laboratories, and grants from the Natural Science and Engineering Research Council of Canada and the Ontario Ministry of Health to Kevin G. Munhall. We also thank Philip Rubin for assistance with the preparation of the Figures in this article, and Nancy O'Brien and Cathy Alfandre for their help in compiling this article's reference section. Finally, we are grateful for the critical and helpful reviews provided by Michael Jordan, Bruce Kay, Edward Reed, and Philip Rubin.

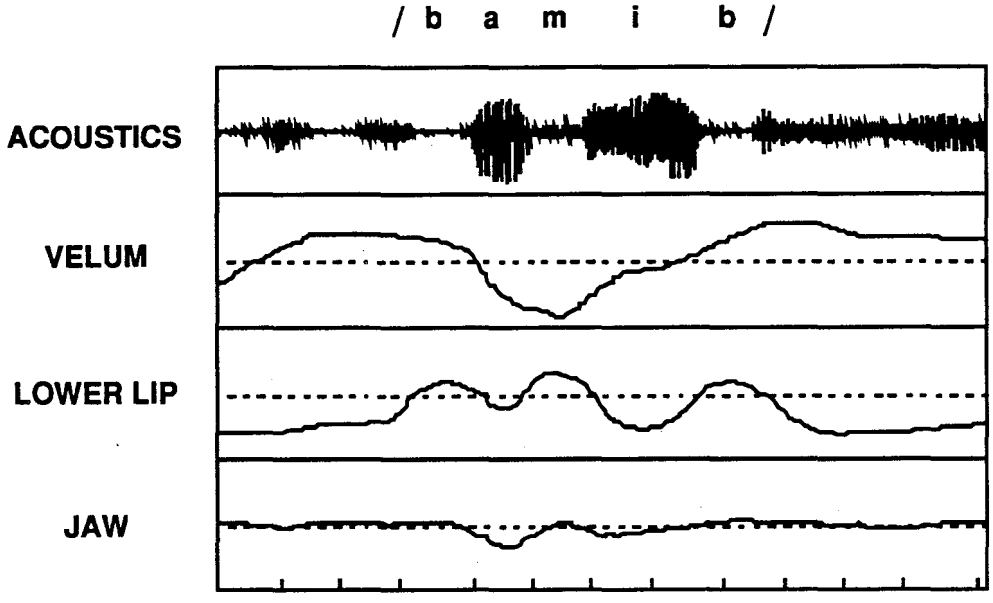

Figure 1. **Acoustic waveform and optoelectronically monitored vertical components of the articulatory trajectories accompanying the utterance "It's a /bamib/ sid". (From Krakow, 1987; used with author's permission).**

In this article, we focus on the patterning of speech gestures,[1] drawing on recent developments in experimental phonology/phonetics and in the study of coordinated behavior patterns in multi-degree-of-freedom dynamical systems. Our key questions are the following: How can one best reconcile traditional linguistic analyses (discrete, context-independent units) with experimental observations of speech articulation and acoustics (continuous, context-dependent flows)? How can one reconcile the hypothesis of underlying invariance with the reality of surface variability? We try to answer these questions by detailing a specific dynamical model of articulation. Our focus on dynamical systems derives from the fact that such systems offer a theoretically unified account of: a) the kinematic forms or patterns displayed by the articulators during speech; b) the stability of these forms to external perturbations; and c) the lawful warping of these forms due to changing system constraints such as speaking rate, casualness, segmental composition, or suprasegmental stress. For us the primary importance of the work lies not so much in the details of this model, but in the problems that can be delineated within its framework.[2] It has become clear that a complete answer to these questions will have to address (at least) the following: 1) the nature of the *gestural units* or *primitives* themselves; 2) the articulatory consequences of partial or total temporal overlap (*coproduction*) in the activities of these units that results from gestural interleaving; and 3) the *serial coupling* among gestural primitives, i.e., the processes that govern intergestural relative timing and that provide intergestural cohesion for higher-order, multigesture units.

Our central thesis is that the spatiotemporal patterns of speech emerge as behaviors implicit in a dynamical system with two functionally distinct but interacting levels. The *intergestural* level is defined according to a set of *activation* coordinates; the *interarticulator* level is defined according to both *model articulator and tract-variable* coordinates (see Figure 2). Invariant gestural units are posited in the form of relations between particular subsets of these coordinates and sets of context-independent dynamical parameters (e.g., target position and stiffness). Contextually-conditioned variability across different utterances results from the manner in which the influences of gestural units associated with these utterances are gated and blended into ongoing processes of articulatory control and coordination. The activation coordinate of each unit can be interpreted as the strength with which the associated gesture "attempts" to shape vocal tract movements at any given point in time. The tract-variable and model articulator coordinates of each unit specify the particular vocal-tract constriction (e.g., bilabial) and set of articulators (e.g., lips and jaw) whose behaviors are directly affected by the associated unit's activation. The intergestural level accounts for patterns of

relative timing and cohesion among the activation intervals of gestural units that participate in a given utterance, e.g., the activation intervals for tongue-dorsum and bilabial gestures in a vowel-bilabial-vowel sequence. The interarticulator level accounts for the coordination among articulators evident at a given point in time due to the currently active set of gestures, e.g., the coordination among lips, jaw, and tongue during periods of vocalic and bilabial gestural coproduction.[3]

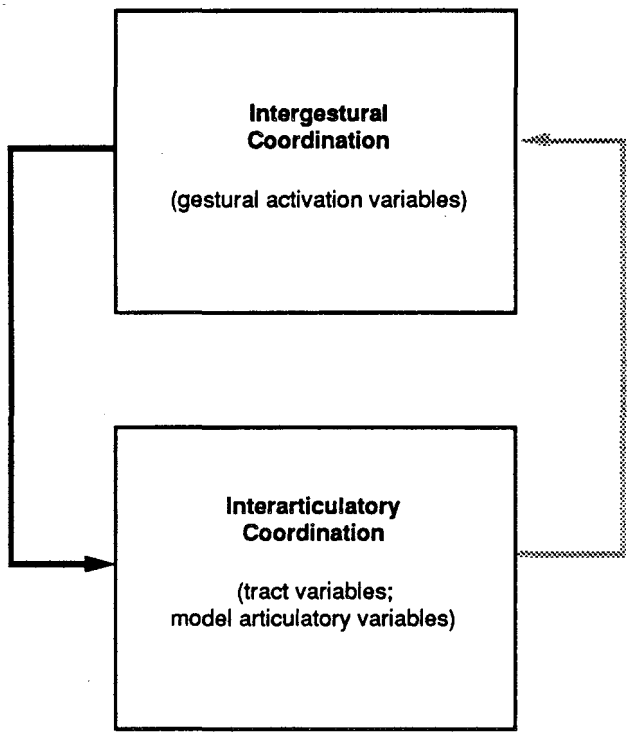

Figure 2. **Schematic illustration of the proposed two-level dynamical model for speech production, with associated coordinate systems indicated. The darker arrow from the intergestural to the interarticulator level denotes the feedforward flow of gestural activation. The lighter arrow indicates feedback of ongoing tract-variable and model articulator state information to the intergestural level.**

In the following pages we take a stepwise approach to elaborating upon these ideas. First, we examine the hypothesis that the formation and release of local constrictions in vocal tract shape are governed by active gestural units that serve to organize the articulators temporarily and flexibly into functional groups or ensembles of joints and muscles (i.e., synergies) that can accomplish particular gestural goals. Second, we review a recent, promising extension of this approach to the related phenomena of coarticulation and coproduction (Saltzman, Rubin, Goldstein, & Browman, 1987). Third, we describe some recent work in a connectionist, computational framework (e.g., Grossberg, 1986; Jordan, 1986, in press; Lapedes & Farber, cited in Lapedes & Farber, 1986; Rumelhart, Hinton, & Williams, 1986) that offers a dynamical account of intergestural timing. Fourth, we examine the issue of intergestural cohesion and the relationships that may exist between stable multiunit ensembles and the traditional linguistic concept of phonological segments. In doing so, we review the work of Browman and Goldstein (1986) on their *articulatory phonology*. Fifth, and finally, we review the influences of factors such as speaking rate and segmental composition on gestural patterning, and speculate on the further implications of our approach for understanding the production of speech.

Gestural primitives for speech: A dynamical framework

Much theoretical and empirical evidence from the study of skilled movements of the limbs and speech articulators supports the hypothesis that the "significant informational units of action" (Greene, 1971, p. xviii) do not entail rigid or hard-wired control of joint and/or muscle variables. Rather, these units or *coordinative structures* (e.g., Fowler, 1977; Kugler, Kelso & Turvey, 1980, 1982; Kugler & Turvey, 1987; Saltzman, 1986; Saltzman & Kelso, 1987; Turvey, 1977) must be defined abstractly or functionally in a task-specific, flexible manner. Coordinative structures have been conceptualized within the theoretical and empirical framework provided by the field of (dissipative) nonlinear dynamics (e.g., Abraham & Shaw, 1982, 1986; Guckenheimer & Holmes, 1983; Haken, 1983; Thompson & Stewart, 1986; Winfree, 1980). Specifically, it has been hypothesized (e.g., Kugler et al., 1980; Saltzman & Kelso, 1987; Turvey, 1977) that coordinative structures be defined as task-specific and autonomous (time-invariant) dynamical systems that underlie an action's form as well as its stability properties. These attributes of task-specific flexibility, functional definition, and time-invariant dynamics have been incorporated into a *task-dynamic* model of coordinative structures (Kelso, Saltzman & Tuller, 1986a, 1986b; Saltzman, 1986; Saltzman & Kelso, 1987; Saltzman et al., 1987). In the model, time-invariant dynamical systems for specific skilled actions are defined at an abstract (task space) level of system description. These invariant dynamics underlie and give rise to contextually-

dependent patterns of change in the dynamic parameters at the articulatory level, and hence to contextually-varying patterns of articulator trajectories. Qualitative differences between the stable kinematic forms required by different tasks are captured by corresponding topological distinctions among task-space *attractors* (see also Arbib, 1984, for a related discussion of the relation between task and controller structures). As applied to limb control, for example, gestures involving a hand's discrete motion to a single spatial target and repetitive cyclic motion between two such targets are characterized by time-invariant *point attractors* (e.g., as with a damped pendulum or damped mass-spring, whose motion decays over time to a stable equilibrium point) and *periodic attractors* (*limit cycles*; e.g., as with an escapement-driven pendulum in a grandfather clock, whose motion settles over time to a stable oscillatory cycle), respectively.

Model articulator and tract variable coordinates

In speech, a major task for the articulators is to create and release constrictions locally in different regions of the vocal tract, e.g., at the lips for bilabial consonants, or between the tongue dorsum and palate for some vowels.[4] In task-dynamics, constrictions in the vocal tract are governed by a dynamical system defined at the interarticulator level (Figure 2) according to both tract variable (e.g., bilabial aperture) and model articulator (e.g., lips and jaw) coordinates. Tract variables are the coordinates in which context-*independent* gestural "intents" are framed, and model articulators are the coordinates in which context-*dependent* gestural performances are expressed. The distinction between tract-variables and model articulators reflects a behavioral distinction evident in speech production. For example, in a vowel-bilabial-vowel sequence a given degree of effective bilabial closure may be achieved with a range of different lip-jaw movements that reflects contextual differences in the identities of the flanking vowels (e.g., Sussman, MacNeilage, & Hanson, 1973).

In task-dynamic simulations, each constriction type (e.g., bilabial) is associated with a pair (typically) of tract variables, one that refers to the location of the constriction along the longitudinal axis of the vocal tract, and one that refers to the degree of constriction measured perpendicularly to the longitudinal axis in the sagittal plane. Furthermore, each gestural/constriction type is associated with a particular subset of model articulators. These simulations have been implemented using the Haskins Laboratories software articulatory synthesizer (Rubin, Baer & Mermelstein, 1981). The synthesizer is based on a midsagittal view of the vocal tract and a simplified kinematic description of the vocal tract's articulatory geometry. Modeling work has been performed in cooperation with several of our colleagues at Haskins Laboratories as part of an ongoing project focused on the development of a gesturally-based, computational model of linguistic structures (Browman & Goldstein, 1986, in press; Browman, Goldstein, Kelso, Rubin, & Saltzman, 1984; Browman, Goldstein, Saltzman, & Smith, 1986; Kelso et al., 1986a, 1986b; Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985; Saltzman, 1986; Saltzman et al., 1987).

Figures 3 and 4 illustrate the tract variables and articulatory degrees-of-freedom that are the focus of this article. In the present model, they are associated with the control of bilabial, tongue-dorsum, and "lower-tooth-height" constrictions.[5] Bilabial gestures are specified according to the tract variables of lip protrusion (LP; the horizontal distance of the upper and lower lips to the upper and lower front teeth, respectively) and lip aperture (LA; the vertical distance between the lips). For bilabial gestures the four modeled articulatory components are: yoked horizontal movements of the upper and lower lips (LH), jaw angle (JA), and independent vertical motions of the upper lip (ULV) and lower lip (LLV) relative to the upper and lower front teeth, respectively. Tongue-dorsum gestures are specified according to the tract variables of tongue-dorsum constriction location (TDCL) and constriction degree (TDCD). These tract variables are defined as functions of the current locations in head-centered coordinates of the region of maximum constriction between the tongue-body surface and the upper and back walls of the vocal tract. The articulator set for tongue-dorsum gestures has three components: tongue body radial (TBR) and angular (TBA) positions relative to the jaw's rotation axis, and jaw angle (JA). Lower-tooth-height gestures are specified according to a single tract variable defined by the vertical position of the lower front teeth, or equivalently, the vertical distance between the upper and lower front teeth. Its articulator set is simply jaw angle (JA). This tract variable is not used in most current simulations, but was included in the model to test hypotheses concerning suprasegmental control of the jaw (Macchi, 1985), the role of lower tooth height in tongue blade fricatives, etc.

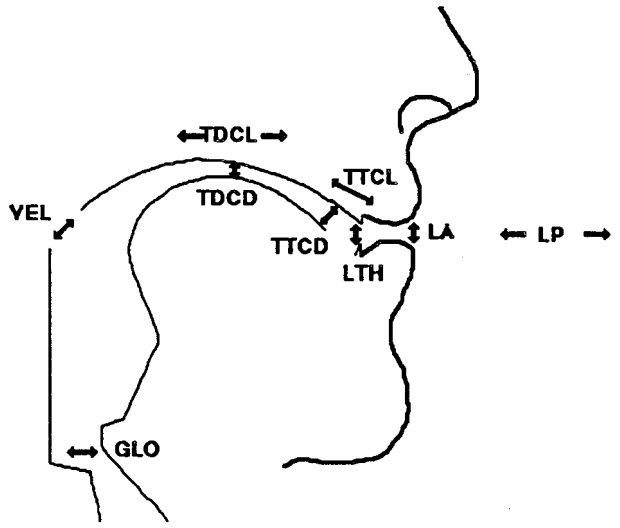

Figure 3. **Schematic midsagittal vocal tract outline, with tract-variable degrees of freedom indicated by arrows. (see text for definitions of tract-variable abbreviations used).**

TRACT VARIABLES (Z_i; i = 1, 2, ...,m; m=9)	MODEL ARTICULATOR VARIABLES ($Ø_j$; j = 1, 2, ...,n; n=10)									
	LH ($Ø_1$)	JA ($Ø_2$)	ULV ($Ø_3$)	LLV ($Ø_4$)	TBR ($Ø_5$)	TBA ($Ø_6$)	TTR ($Ø_7$)	TTA ($Ø_8$)	V ($Ø_9$)	G ($Ø_{10}$)
LP (Z_1)	●									
LA (Z_2)		●	●	●						
TDCL (Z_3)		●			●	●				
TDCD (Z_4)		●			●	●				
LTH (Z_5)		●								
TTCL (Z_6)		●			●	●	●	●		
TTCD (Z_7)		●			●	●	●	●		
VEL (Z_8)									●	
GLO (Z_9)										●

Figure 4. **Matrix representing the relationship between tract-variables (z) and model articulators (ø). The filled cells in a given tract-variable row denote the model articulator components of that tract-variable's articulatory set. The empty cells indicate that the corresponding articulators do not contribute to the tract-variable's motion. (See text for definitions of abbreviations used in the figure.)**

Each gesture in a simulated utterance is associated with a corresponding tract-variable dynamical system. At present, all such dynamical systems are defined as tract-variable point-attractors, i.e., each is modeled currently by a damped, second-order linear differential equation (analogous to a damped mass-spring). The corresponding set of tract-variable motion equations is described in Appendix 1. These equations are used to specify a functionally equivalent dynamical system expressed in the model articulator coordinates of the Haskins articulatory synthesizer. This model articulator dynamical system is used to generate articulatory motion patterns. It is derived by transforming the tract-variable motion equations into an articulatory space whose components have geometric attributes (size, shape) but are massless. In other words, this transformation is a strictly kinematic one, and involves only the substitution of variables defined in one coordinate system for variables defined in another coordinate system (see Appendix 2).

Using the model articulator dynamical system (Equation [A4] in Appendix 2) to simulate simple utterances, the task-dynamic model has been able to generate many important aspects of natural articulation. For example, the model has been used to reproduce experimental data on *compensatory articulation*, whereby the speech system quickly and automatically reorganizes itself when faced with unexpected mechanical perturbations (e.g., Abbs & Gracco, 1983; Folkins & Abbs, 1975; Kelso, Tuller, Vatikiotis-Bateson & Fowler, 1984; Munhall & Kelso, 1985; Munhall, Löfqvist & Kelso, 1986; Shaiman & Abbs, 1987) or with static mechanical alterations of vocal tract shape (e.g., Gay, Lindblom, & Lubker, 1981; MacNeilage, 1970). Such compensation for mechanical disturbances is achieved by readjusting activity over an entire subset of articulators in a gesturally-specific manner. The task-dynamic model has been used to simulate the compensatory articulation observed during bilabial closure gestures (Saltzman, 1986; Kelso et al., 1986a, 1986b). Using point-attractor (e.g., damped mass- spring) dynamics for the control of lip aperture, when the simulated jaw is "frozen" in place during the closing gesture, at least the main qualitative features of the data are captured by the model, in that: 1) the target bilabial closure is reached (although with different final articulator configurations) for both perturbed and unperturbed "trials," and 2) compensation is immediate in the upper and lower lips to the jaw perturbation, i.e., the system does not require reparameterization in order to compensate. Significantly, in task-dynamic modeling the processes governing intra-gestural motions of a given set of articulators (e.g., the bilabial articulatory set defined by the jaw and lips) are *exactly the same* during simulations of both unperturbed and mechanically perturbed active gestures. In all cases, the articulatory movement patterns emerge as implicit consequences of the gesture-specific dynamical parameters (i.e., tract-variable parameters and articulator weights; see Appendices 1 and 2), and the ongoing postural state (perturbed or not) of the articulators. Explicit trajectory planning and/or replanning procedures are not required.

Gestural activation coordinates

Task dynamics identifies several different time spans that are important for conceptualizing the dynamics of speech production. For example, the *settling time* of an unperturbed discrete bilabial gesture is the time required for the system to move from an initial position with zero velocity to within a criterion percentage (e.g., 2%) of the distance between initial and target positions. A gesture's settling time is determined jointly by the inertia, stiffness, and damping parameters intrinsic to the associated tract-variable point attractor. Thus, gestural duration or settling time is implicit in the dynamics of the interarticulator level (Figure 2, bottom), and is not represented explicitly. There is, however, another time span that is defined by the temporal interval during which a gestural unit actively shapes movements of the articulators. In previous sections, the concept of gestural activity was used in an intuitive manner only. We now define it in a more specific fashion.

Intervals of active gestural control are specified at the intergestural level (Figure 2, top) with respect to the system's activation variables. The set of activation variables defines a third coordinate system in the present model, in addition to those defined by the tract variables and model articulators (see Figures 2 & 5). Each distinct tract-variable gesture is associated with its own activation variable, a_{ik}, where the subscript-i denotes numerically the associated tract variable ($i = 1, ..., m$), and the subscript-k denotes symbolically the particular gesture's linguistic affiliation (k = /p/, /i/, etc.). The value of a_{ik} can be interpreted as the strength with which the associated tract-variable dynamical system "attempts" to shape vocal tract movements at any given point in time.

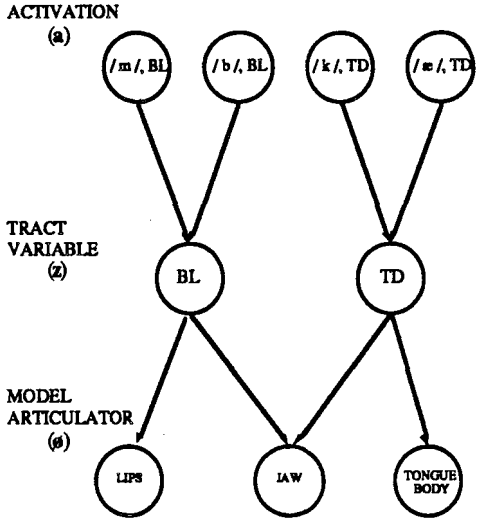

Figure 5. **Example of the "anatomical" connectivity pattern defined among the model's three coordinate systems. BL and TD denote tract-variables associated with bilabial and tongue-dorsum constrictions, respectively.**

In current simulations the temporal patterning of gestural activity is accomplished with reference to a *gestural score* (Figure 6) that represents the activation of gestural primitives over time across parallel tract-variable output channels. Currently, these activation patterns are not derived from an underlying implicit dynamics. Rather, these patterns are specified explicitly "by hand", or are derived according to a rule-based synthesis program called GEST that accepts phonetic string inputs and generates gestural score outputs (Browman et al., 1986). In the gestural score for a given utterance, the corresponding set of activation functions is specified as an explicit matrix function of time, $A(t)$. For purposes of simplicity, the activation interval of a given gesture-ik is specified according to the duration of a step-rectangular pulse in a_{ik}, normalized to unit height ($a_{ik} \in \{0, 1\}$). In future developments of the task-dynamic model (see the *Serial Dynamics* section later in the article), we plan to generalize the shapes of the activation waves and to allow activations to vary continuously over the interval ($0 \leq a_{ik} \leq 1$).[6]

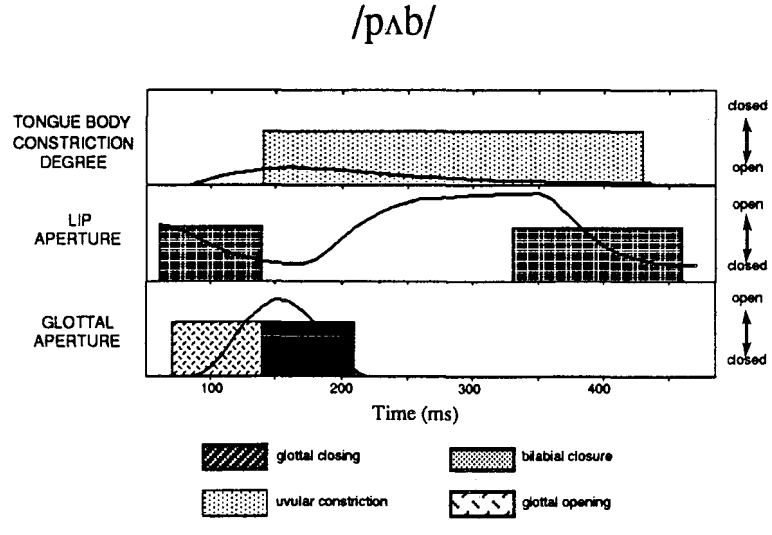

Figure 6. **Gestural score used to synthesize the sequence /pʌb/. Filled boxes denote intervals of gestural activation. Box heights are uniformly either 0 (no activation) or 1 (full activation). The waveform lines denote tract-variable trajectories generated during the simulation.**

COPRODUCTION

Temporally discrete or isolated gestures are at best rare exceptions to the rule of temporal interleaving and overlap (coproduction) among gestures associated with nearby segments (e.g., Bell-Berti & Harris, 1981; Fowler, 1977, 1980; Harris, 1984; Keating, 1985; Kent & Minifie, 1977; Öhman, 1966, 1967; Perkell, 1969; Sussman et al., 1973; for an historical review, see Hardcastle, 1981). This interleaving is the source of the ubiquitous phenomenon of coarticulation in speech production. Coarticulation refers to the fact that at any given point during an utterance, the influences of gestures associated with several adjacent or near-adjacent segments can generally be discerned in acoustic or articulatory measurements. Coarticulatory effects can occur, for example, when lip protrusion for a following rounded vowel begins during the preceding phonologically unrounded consonant, thereby coloring the acoustic correlates of the consonant with those of the following vowel. Similarly, in a vowel-/p/-vowel sequence the formation of the bilabial closure for /p/ (using the jaw and lips) appears to be influenced by temporally overlapping demands associated with the following vowel (using the jaw and tongue) by virtue of the shared jaw component.[7] In the context of the present model (see also Coker, 1976; Henke, 1966), these overlapping demands can be represented as overlapping activation patterns in a corresponding set of gestural scores. The specification of gestural scores (either by hand or by synthesis rules) thereby allows rigorous experimental control over the temporal onsets and offsets of the activations of simulated gestures, and provides a powerful computational framework and research tool for exploring and testing hypotheses derived from current work in experimental phonology/phonetics. In particular, these methods have facilitated the exploration of coarticulatory phenomena that have been ascribed to the effects of partial overlap or coproduction of speech gestures. We now describe in detail how gestural activation is incorporated into ongoing control processes in the model, and the effects of coproduction in shaping articulatory movement patterns.

Active gestural control: Tuning and gating

How might a gesture gain control of the vocal tract? In the present model, when a given gesture's activation is maximal (arbitrarily defined as 1.0), the gesture exerts maximal influence on all the articulatory components associated with the gesture's tract-variable set. During each such activation interval, the evolving configuration of the model articulators results from the gesturally- and posturally-specific way that driving influences generated in the tract-variable space (Equation [A1], Appendix 1) are distributed across the associated sets of articulatory components (Equations [A3] and [A4], Appendix 2) during the course of the movement. Conversely, when the gesture's activation is minimal (arbitrarily defined as 0.0), none of the articulators are subject to active control influences from that gesture. What, then, happens to the model articulators when there is no active control? We begin by considering the former question of active control, and treat the latter issue of nonactive control below in the section *Nonactive Gestural Control.*

The driving influences associated with a given gesture's activation "wave" (see Figure 6) are inserted into the interarticulator dynamical system in two ways in our current simulations. The first way serves to define or *tune* the current set of dynamic parameter values in the model (i.e., K, B, $\mathbf{z}_0$, and W in Equations [A3] and [A4], Appendix 2; see also Saltzman & Kelso, 1983, 1987 for a related discussion of parameter tuning in the context of skilled limb actions). The second way serves to implement or *gate* the current pattern of tract-variable driving influences into the appropriate set of articulatory components. The current use of tuning and gating is similar to Bullock and Grossberg's (1988a, 1988b; see also Cohen, Grossberg & Stork, 1988; Grossberg, 1978) use of target specification and "GO signals," respectively, in their model of sensorimotor control.

The details of tuning and gating processes depend on the ongoing patterns of overlap that exist among the gestures in a given utterance. The gestural score in Figure 6 captures in a descriptive sense both the temporal overlap of speech gestures as well as a related spatial type of overlap. As suggested in the figure, coproduction occurs whenever the activations of two or more gestures overlap partially (or wholly) in time within and/or across tract-variables. Spatial overlap occurs whenever two or more coproduced gestures share some or all of their articulatory components. In these cases, the influences of the spatially and temporally overlapping gestures are said to be *blended*. For example, in a vowel-consonant-vowel (VCV) sequence, if one assumes that the activation intervals of the vowel and

consonant gestures overlap temporally (e.g., Öhman, 1967; Sussman et al., 1973), then one can define a continuum of supralaryngeal overlap that is a function of the gestural identity of the medial consonant. In such sequences, the flanking vowel gestures are defined, by hypothesis, along the tongue-dorsum tract variables and the associated articulatory set of jaw and tongue body. If the consonant is /h/, then there is no supralaryngeal overlap. If the consonant is /b/, then its gesture is defined along the bilabial tract variables and the associated lips-jaw articulator set. Spatial overlap occurs in this case at the shared jaw. If the consonant is the alveolar /d/, its gesture is defined along the tongue-tip tract variables and the associated articulator set of tongue tip, tongue body, and jaw. Spatial overlap occurs then at the shared jaw and tongue body. Note that in both the bilabial and alveolar instances the spatial overlap is not total, and there is at least one articulator free to vary, adaptively and flexibly, in ways specific to its associated consonant. Thus, Öhman (1967) showed, for a medial alveolar in a VCV sequence, that both the location and degree of tongue-tip constriction were unaffected by the identity of the flanking vowels, although the tongue-dorsum's position was altered in a vowel-specific manner. Finally, if the medial consonant in a VCV sequence is the velar /g/, the consonant gesture is defined along exactly the same set of tract variables and articulators as the flanking vowels. In this case, there is total spatial overlap, and the system shows a loss of behavioral flexibility. That is, there is now contextual variation evident even in the attainment of the consonant's tongue-dorsum constriction target; Öhman (1967), for example, showed that in such cases the velar's place of constriction was altered by the flanking vowels, although the constriction degree was unaffected.

Blending due to spatial and temporal overlap occurs in the model as a function of the manner in which the current gestural activation matrix, A, is incorporated into the interarticulator dynamical system. Thus, blending is implemented with respect to both the gestural parameter set (tuning) and the transformation from tract-variable to articulator coordinates (gating) represented in Equations (A3) and (A4). In the following paragraphs, we describe first the computational implementation of these activation and blending processes, and then describe the results of several simulations that demonstrate their utility.

Parameter tuning. Each distinct simulated gesture is linked to a particular subset of tract-variable and articulator coordinates, and has associated with it a set of time-invariant parameters that are likewise linked to these coordinate systems. For example, a tongue-dorsum gesture's stiffness, damping, and target parameters are associated with the tract variables of tongue-dorsum constriction location and degree; its articulator weighting parameters are associated with the jaw angle, tongue-body radial, and tongue-body angular degrees of freedom. Values for these parameters are estimated from kinematic speech data obtained by optoelectronic or X-ray measurements (e.g., Kelso et al., 1985; Smith, Browman, & McGowan, 1988; Vatikiotis-Bateson, 1988). The parameter set for a given gesture is represented as:

$$k_{ik}^{+},\ b_{ik}^{+},\ z_{oik}^{+}, w_{ikj},$$

where the subscripts denote numerically either tract variables (i = 1, ..., m) or articulators (j = 1, ..., n), or denote symbolically the particular gesture's linguistic affiliation (k = /p/, /i/, etc.). These parameter sets are incorporated into the interarticulator dynamical system (see Equations [A3] and [A4], Appendix 2) as functions of the current gestural activation matrix, A, according to explicit algebraic blending rules. These rules define or tune the current values for the corresponding components of the vector $\boldsymbol{z_o}$ and matrices K, B, and W in Equations (A3) and (A4) as follows:

$$b_{ii} = \sum_{k \in Z_i} (p_{Tik} b_{ik}^{+}) \ ; \tag{1a}$$

$$k_{ii} = \sum_{k \in Z_i} (p_{Tik} k_{ik}^{+}) \ ; \tag{1b}$$

$$z_{oi} = \sum_{k \in Z_i} (p_{Tik} z_{oik}^{+}) \ ; \text{ and} \tag{1c}$$

$$w_{jj} = \sum_{i \in \Phi_j} \left(\sum_{k \in Z_i} p_{Wikj} w_{ikj}^{+} \right) + g_{Njj} \ , \tag{1d}$$

where p_{Tik} and p_{Wikj} are variables denoting the *post*-blending strengths of gestures whose activations influence the ongoing values of tract-variable and articulatory-weighting parameters, respectively, in Equations A1 - A4 (Appendices 1 and 2); Z_i is the set of gestures associated with the i^{th} tract-variable; Φ_j is the set of tract-variables associated with the j^{th} model articulator (see Figure 4); and $g_{Njj} = 1.0 - \min \left[1, \sum_{i \in \Phi_j} \left(\sum_{k \in Z_i} a_{ik} \right) \right]$. The subscript N denotes the fact that, for parsimony's sake, g_{Njj} are the same elements used to gate in the j^{th} articulatory component of the *neutral attractor* (see the *Nonactive Gestural Control* section to follow). In Equations (1a-1c), the parameters of the i^{th} tract-variable assume default values of zero at times when there are no active gestures that involve this tract-variable. Similarly, in Equation (1d), the articulatory weighting parameter of the j^{th} articulator assumes a default value of 1.0, due to the contribution of the g_{Njj} term, at times when there are no active gestures involving this articulator.

The p_{Tik} and p_{Wikj} terms in Equation 1 are given by the steady-state solutions of a set of feedforward, competitive-interaction-network dynamical Equations (see Appendix 3 for details). These solutions are expressed as follows:

$$p_{Tik} = a_{ik} \Big/ \left(1 + \beta_{ik} \sum_{\substack{l \in Z_i \\ l \neq k}} [\, \alpha_{il} a_{il}] \right); \tag{2a}$$

$$p_{Wikj} = a_{ik} \Big/ \left(1 + \beta_{ik} \sum_{i \in \Phi_j} \left[\sum_{\substack{l \in Z_i \\ l \neq k}} \alpha_{il} a_{il} \right] \right), \tag{2b}$$

where α_{il} = competitive interaction (lateral inhibition) coefficient from gesture-*il* to gesture-*ik*, for $l \neq k$; and β_{ik} = a "gatekeeper" coefficient that modulates the incoming lateral inhibition influences impinging on gesture-*ik* from gesture-*il*, for $l \neq k$; For parsimony, β_{ik} is constrained to equal $1.0/\alpha_{ik}$, for $\alpha_{ik} \neq 0.0$. If $\alpha = 0.0$, β_{ik} is set to equal 0.0 by convention. Implementing the blended parameters defined by Equations (1a-1c) into the dynamical system defined by Equations (A3) and (A4) creates an attractor layout or field of driving influences in tract-variable space that is specific to the set of currently active gestures. The blended parameters defined by Equation (1d) create a corresponding pattern of relative "receptivities" to these driving influences among the associated synergistic articulators in the coordinative structure.

Using the blending rules provided by Equations (1) and (2), different *forms* of blending can be specified, for example, among a set of temporally overlapping gestures defined within the same tract variables. The form of blending depends on the relative sizes of the context-independent (time-invariant) α and β parameters associated with each gesture. For $a_{ik} \in \{0, 1\}$, three possibilities are averaging, suppressing, and adding. For the set of currently active gestures along the i^{th} tract variable, if all α's are equal and greater than zero (all β's are then equal by constraint), then the $\sum_{k \in Z_i} p_{Tik}$ is normalized to equal 1.0 and the tract-variable parameters blend by simple averaging. If the α's are unequal and greater than zero, then the $\sum_{k \in Z_i} p_{Tik}$ is also normalized to equal 1.0 and the parameters blend by a weighted averaging. For example, if gesture-*ik*'s α_{ik} = 10.0 and gesture-il's α_{il} = 0.1, then gesture-*ik*'s parameter values dominate or "suppress" gesture-il's parameter values in the blending process when both gestures are co-active. Finally, if all α's = 0.0, then all β's = 0.0 by convention, and the parameters in Equation (1) blend by simple addition. Currently, all gestural parameters in Equation (1) are subject to the same form of competitive blending. It would be possible at some point, however, to implement different blending forms for the different parameters, e.g., adding for targets and averaging for stiffnesses, as suggested by recent data on lip protrusion (Boyce, 1988) and laryngeal abduction (Munhall & Löfqvist, 1987).

Transformation gating. The tract-variable driving influences shaped by Equations (1) and (2) remain implicit and "disconnected" from the receptive model articulators until these influences are gated explicitly into the articulatory degrees of freedom. This gating occurs with respect to the weighted Jacobian pseudoinverse (i.e., the transformation that relates tract-variable motions

to articulatory motions) and its associated orthogonal projection operator (see Appendix 2). Specifically, J^* and I_n are replaced in Equation (A4) by gated forms, J^*_G and G_P, respectively. J^*_G can be expressed as follows:

$$J^*_G = W^{-1}J_G^T(C + [I_m - G_A])^{-1}, \quad (3a)$$

where $J_G = G_A J$, and G_A is a diagonal $m \times m$ *gating* matrix for the active tract-variable gestures. Each $g_{Aii} = min\left(\sum_{k \in Z_i} a_{ik}\right)$, where the summation is defined as in Equations (1) and (2). Each g_{Aii} multiplies the i^{th} row of the Jacobian. This row relates motions of the articulators to motions defined along the i^{th} tract variable (see Equation [A2]). When $g_{Aii} = 1$ (or 0), the i^{th} tract variable is gated into (out of) J^*_G and contributes (does not contribute) to $\ddot{\phi}_A$, the vector of active articulatory driving influences (see Equations [A3] and [A4]); $C = J_G W^{-1} J_G^T$. C embodies the kinematic interrelationships that exist among the currently active set of tract variables. Specifically, C is defined by the set of weighted, pairwise inner products of the gated Jacobian rows. A diagonal element of C, c_{ii}, is the weighted inner product (sum of squares) of the i^{th} gated Jacobian row with itself; an off-diagonal element, c_{hi} $(h \neq i)$, is the weighted inner product (sum of products) of the h^{th} and i^{th} gated Jacobian rows. A pair of gated Jacobian rows has a (generally) nonzero weighted inner product when the corresponding tract variables are active and share some or all articulators in common; the weighted inner product of two gated Jacobian rows equals zero when the corresponding tract variables are active and share no articulators; the inner product also equals zero when one or both rows correspond to a nonactive tract variable; and I_m = a $m \times m$ identity matrix.

The gated orthogonal projection operator is expressed as follows:

$$[G_P - J^*_G J], \quad (3b)$$

where G_P = a diagonal $n \times n$ gating matrix.

Each element $g_{Pjj} = min\left[1, \sum_{i \in \Phi_j}\left(\sum_{k \in Z_i} a_{ik}\right)\right]$, where the summations are defined as in Equations (1) and (2).

For example, if there are no active gestures then $G_A = 0$, $G_P = 0$, and $(C + [I_m - G_A]) = I_m$ Consequently, both J^*_G and the gated orthogonal projection operator equal zero, and $\ddot{\phi}_A = 0$ according to Equation (A4). If active gestures occur simultaneously in all tract variables, then $G_A = I_m$, $G_P = I_n$, and $J^*_G = J^*$. That is, both the gated pseudoinverse and orthogonal projection operator are "full blown" when all tract variables are active, and $\ddot{\phi}_A$ is influenced by the attractor layouts and corresponding driving influences defined over all the tract variables. If only a few of the tract variables are active, these terms are not full blown and $\ddot{\phi}_A$ is only subject to driving influences associated with the attractor layout in the subspace of active tract variables.

Nonactive gestural control: The neutral attractor

We return now to the question of what happens to the model articulators when there is no active control in the model. In such cases, articulator movements are shaped by a "default" *neutral attractor*. The neutral attractor is a point attractor in model articulator space, whose target configuration corresponds to schwa /ə / in current modeling. It is possible, however, that this neutral target may be language-specific. The articulatory degrees of freedom in the neutral attractor are uncoupled dynamically, i.e., point attractor dynamics are defined independently for all articulators. At any given point in time, the neutral attractor exerts a set of driving influences on the articulators, $\ddot{\phi}_N$, that can be expressed as follows:

$$\ddot{\phi}_N = G_N (- B_N \dot{\phi} - K_N[\phi - \phi_{N0}]), \quad (4)$$

where ϕ_{N0} is the neutral target configuration; B_N and K_N are $n \times n$ diagonal damping and stiffness matrices, respectively. Because their parameters never change, parameter tuning or blending is not defined for the neutral attractor. The components, k_{Njj}, of K_N are typically defined to be equal, although they may be defined asymmetrically to reflect hypothesized differences in the biomechanical time constants of the articulators (e.g., the jaw is more sluggish [has a larger time constant] than the tongue tip). The components, b_{Njj}, of B_N are defined at present relative to the corresponding K_N components to provide critical

damping for each articulator's neutral point attractor. For parsimony's sake, B_N is also used to define the orthogonal projection vector in Equation (A4); and G_N is a $n \times n$ diagonal *gating* matrix for the neutral attractor. Each element $g_{Njj} = 1.0 - \min\left[1, \sum_{i \in \Phi_j}\left(\sum_{k \in Z_i} a_{ik}\right)\right]$, where the summations are defined as in Equations (1-3). Note that $G_N = I_n - G_P$, where I_n is the n x n identity matrix and G_P is defined in Equation (3b).

The total set of driving influences ($\ddot{\phi}_T$) on the articulators at any given point in time is the sum of an active component ($\ddot{\phi}_A$; see Equations [A3], [A4], and [3]) and a neutral component ($\ddot{\phi}_N$; see Equation [4]), and is defined as follows:

$$\ddot{\phi}_T = \ddot{\phi}_A + \ddot{\phi}_N \quad (5)$$

For example, consider a time when only a tongue-dorsum gesture is active. Then g_{N11} (for LH) = g_{N33} (for ULV) = g_{N44} (for LLV) = 1.0, and g_{N22} (for JA) = g_{N55} (for TBR) = g_{N66} (for TBA) = 0. Active control will exist for the gesturally involved jaw and tongue (JA, TBR, and TBA), but the noninvolved lips (LH, ULV, and LLV) will "relax" independently from their current positions toward their neutral positions according to the specified time constants. If only a bilabial gesture is active, then the complementary situation holds, with $g_{N11} = g_{N33} = g_{N44} = 0.$, and $g_{N22} = g_{N55} = g_{N66} =$ 1.0. The jaw and lips will be actively controlled, and the tongue will relax toward its neutral configuration. When both bilabial and tongue-dorsum gestures are active simultaneously, all g_{Njj} components equal zero, $\ddot{\phi}_N = 0$, and the neutral attractor has no influence on the articulatory movement patterns. When there is no active control, all g_{Njj} components equal one, and all articulators relax toward their neutral targets.

Simulation examples

We now describe results from several simulations that demonstrate how active and neutral control influences are implemented in the model, focusing on instances of gestural coproduction.

Parameter tuning. The form of parameter blending for speech production has been hypothesized to be tract-variable-specific (Saltzman et al., 1987). As already discussed in the *Active Gestural Control* section, Öhman (1967) showed that for VCV sequences, when the medial consonant was a velar (/g/ or /k/), the surrounding vowels appeared to shift the velar's place of constriction but not its degree. These results have been simulated (qualitatively, at least) by superimposing temporally the activation intervals for the medial velar consonant and the flanking vowels. During the resultant period of consonant-vowel coproduction, an averaging blend was implemented for tuning the tract-variable parameters of tongue-dorsum-constriction-location, and a suppressing blend (velar suppresses vowel) was implemented for tongue-dorsum-constriction-degree (see Figure 7; Saltzman et al., 1987; cf., Coker, 1976, for an alternative method of generating similar results).

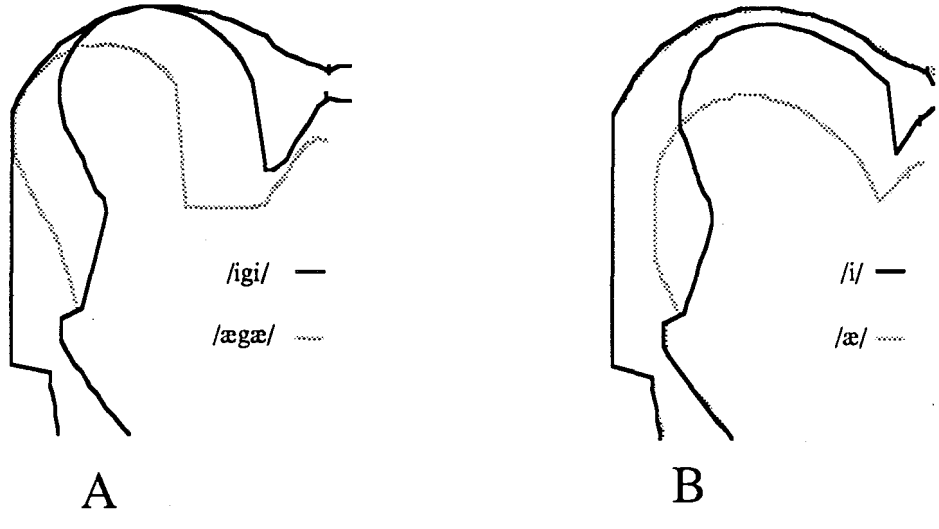

Figure 7. **Simulated vocal tract shapes. A. First contact of tongue-dorsum and upper tract wall during symmetric vowel-velar-vowel sequences. B. Corresponding steady-state vowel productions.**

This blending scheme for constriction degree is consistent with the assumption in current modeling that the amount of suppression during blending is related to differences in the sonority (Jesperson, 1914) or openness of the vocal tract associated with each of the blended gestures. Gestural sonority is reflected in the constriction degree target parameters of each gesture. For tongue-dorsum gestures, vowels have large positive-valued targets for constriction degree that reflect their open tract shapes (high sonority), and stops have small negative target values that reflect contact-plus-compression against the upper tract wall (low sonority).

Transformation gating. Simulations described in this article of blending for gestures defined along different tract variables have been restricted to periods of temporal overlap between pairs of bilabial and tongue-dorsum gestures. Under these circumstances, articulatory trajectories have been generated for sequences involving consonantal gestures superposed onto ongoing vocalic gestures that match (qualitatively, at least) the trajectories observed in X-ray data. In particular, Tiede and Browman (1988) analyzed X-ray data that included the vertical motions of pellets placed on the lower lip, lower incisor (i.e., jaw), and "mid-tongue" surface during /pV_1pV_2p/ sequences. The mid-tongue pellet height corresponds, roughly, to tongue-dorsum height in the current model. Tiede and Browman found that the mid-tongue pellet moved with a relatively smooth trajectory from its position at the onset of the first vowel to its position near the offset of the second vowel. Specifically, when V_1 was the medium height vowel /ɛ/ and V_2 was the low vowel /a/, the mid-tongue pellet showed a smooth lowering trajectory over this gestural time span (see Figure 8). During this same interval, the jaw and lower lip pellets moved through comparably smooth gestural sequences of lowering for V_1, raising for the medial /p/, and lowering for V_2.

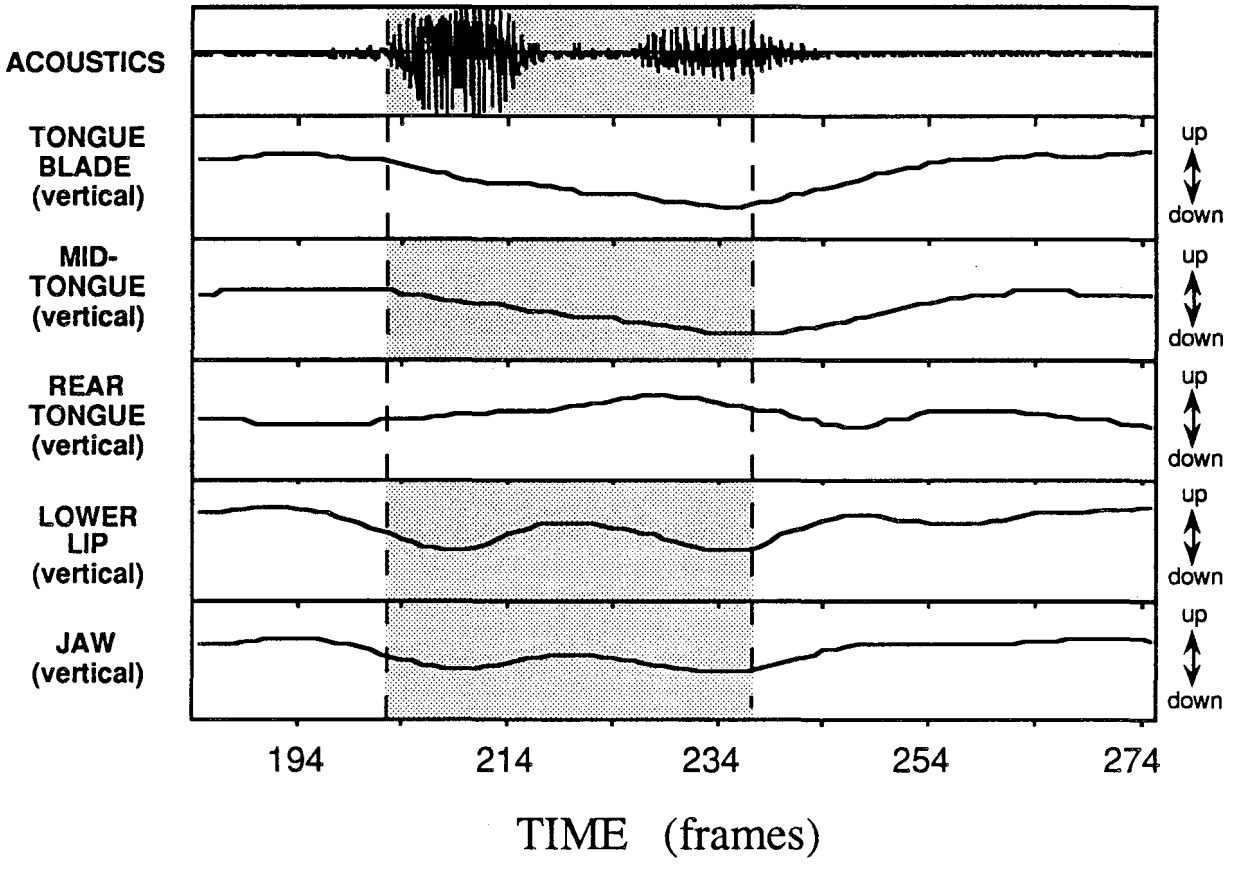

Figure 8. **Acoustic waveform and vertical components of articulatory X-ray pellet data during the utterance /pɛpap/. (From Tiede & Browman, 1988; used with authors' permission).**

Figure 9a shows a simulation with the current model of a similar sequence /əbæbæ/. The main point is that the vowel-to-vowel trajectory for tongue-dorsum-constriction-degree is smooth, going from the initial schwa to the more open /æ/. This tongue-dorsum pattern occurs simultaneously with the comparably smooth closing-opening gestural sequences for jaw height and lip aperture.

Two earlier versions of the present model generated nonacceptable trajectories for this same sequence that are instructive concerning the model's functioning. In one version (the "modular" model), each constriction type operated independently of the other during periods of coproduction. For example, during periods of bilabial and tongue-dorsum overlap, driving influences were generated along the tract-variables associated with each constriction. These influences were then transformed into articulatory driving influences by separate, constriction-specific Jacobian pseudoinverses (e.g., see Equations [A3] and [A4]). The bilabial pseudoinverse involved only the Jacobian rows (see Equation [A2] and Figure 4) for lip aperture and protrusion, and the tongue-dorsum pseudoinverse involved only the Jacobian rows for tongue-dorsum constriction location and degree. The articulatory driving influences associated with each constriction were simply averaged at the articulatory level for the shared jaw. The results are shown in Figure 9b, where it is evident that the tongue-dorsum does not display the relatively smooth vowel-to-vowel trajectory seen in the X-ray data and with the current model. Rather, the trajectory appears to be perturbed in a complex manner by the simultaneous jaw and lip aperture motions. It is hypothesized that these perturbations are due to the fact that the modular model did not incorporate, by definition, the off-diagonal elements of the C-matrix used currently in the gated pseudoinverse (Equation [3]). Recall that these elements reflected the kinematic relationships that exist among different, concurrently active tract-variables by virtue of shared articulators. In the modular model, these terms were absent because the constriction-specific pseudoinverses were defined explicitly to be independent of each other. Thus, if the current model is a reasonable one, it tells us that knowledge of inter-tract-variable kinematic relationships must be embodied in the control and coordinative processes for speech production.

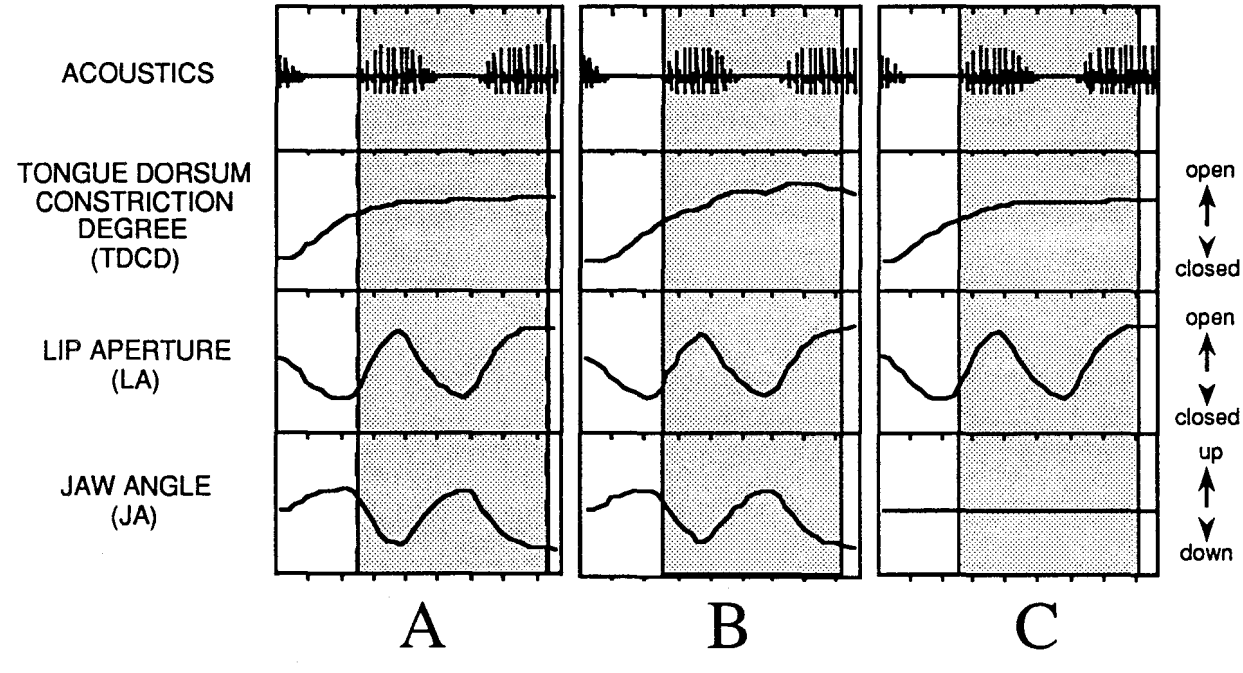

Figure 9. **Simulations of the sequence /əbæbæ/. A. Current model. B. Older "modular" version. C. Older "flat jaw" version.**

A different type of failure by a second earlier version of the model provides additional constraints on the form that must be taken by such inter-tract-variable knowledge. En route to developing the current model, a mistake was made that generated a perfectly flat jaw trajectory (the "flat jaw" model) for the same sequence (/əbæbæ/; see Figure 9c). Interestingly, however, the tongue-dorsum trajectory was virtually identical to that generated with the current model. The reason for this anomalous jaw behavior was that the gated pseudoinverse (Equation [3]) had been forced accidentally to be "full blown" regardless of the ongoing state of gestural activation. This meant that all tract variables were gated on in this transformation, even when the associated gestures were not activated. The specification of the attractor layout at the tract-variable level, however, worked as it does in the current model. Active gestures "create" point attractors in the control landscape for the associated tract variables. In this landscape, the currently active target can be considered to lie at the bottom of a valley whose walls are slightly sticky. The resultant tract-variable motion is to slide stably down the valley wall from its current position toward the target, due to the nonzero driving influences associated with the system's attraction to the target position. Nonactive gestures, on the other hand, "create" only flat tract-variable control landscapes, in which no position is preferred over any other and the value of the tract-variable driving influences equals zero. Recall from the *Gestural Primitives* section (Figures 3 and 4) that the model includes a lower-tooth-height tract variable that maps one-to-one onto jaw angle. For the sequence /əbæbæ/, this tract variable is never active and, consequently, the corresponding component of the tract-variable driving influence vector is constantly equal to zero. When the gated pseudoinverse is full blown, this transformation embodies the kinematic relationships among the bilabial, tongue-dorsum, and lower-tooth-height tract variables that exist by virtue of the shared jaw. This means that the transformation treats the zero driving component for lower-tooth-height as a value that should be passed on to the articulators, in conjunction with the driving influences from the bilabial and tongue-dorsum constrictions. As a result, the jaw receives zero active driving, and because the jaw starts off at its neutral position for the initial schwa, it also receives zero driving from the neutral attractor (Equation [4]) throughout the sequence. The result is the observed flat trajectory for the jaw. Thus, if the current model is a sensible one, this nonacceptable "flat jaw" simulation tells us that the kinematic interrelationships embodied in the system's pseudoinverse at any given point in time must be gated functions of the currently active gesture set.

SERIAL DYNAMICS

The task-dynamic model defines, in effect, a selective pattern of coupling among the articulators that is specific to the set of currently active gestures. This coupling pattern is shaped according to three factors: a) the current state of the gestural activation matrix; b) the tract-variable parameter sets and articulator weights associated with the currently active gestures; and c) the geometry of the nonlinear kinematic mapping between articulatory and tract-variable coordinates (represented by J and J^T in Equations [A2] and [A3]) for all associated active gestures. The model provides an intrinsically dynamical account of multiarticulator coordination within the activation intervals of single (perturbed and unperturbed) gestures. It also holds promise for understanding the blending dynamics of coproduced gestures that share articulators in common. However, task-dynamics does not currently provide an intrinsically dynamic account of the intergestural timing patterns comprising even a simple speech sequence (see Figures 1 and 6). At the level of phonologically defined segments, the sequence might be a repetitive alternation between a given vowel and consonant, e.g., /bababa.../. At a more fine-grained level of description, the sequence might be a "constellation" (Browman & Goldstein, 1986, in press) of appropriately phased gestures, e.g., the bilabial closing-opening and the laryngeal opening-closing for word-initial /p/ in English. As discussed earlier, current simulations rely on explicit gestural scores to provide the layout of activation intervals over time and tract variables for such utterances.

The lack of an appropriate *serial dynamics* is a major shortcoming in our speech modeling to date. This shortcoming is linked to the fact that the most-studied and best-understood dynamical systems in the nonlinear dynamics literature are those whose behaviors are governed by point attractors, periodic attractors (limit cycles), and *strange* attractors. (Strange attractors underlie the behaviors of *chaotic* dynamical systems, in which seemingly random movement patterns have deterministic origins; e.g., Ruelle, 1980). For nonrepetitive and nonrandom speech sequences,

such attractors appear clearly inadequate. However, investigations in the computational modeling of connectionist (parallel distributed processing, neuromorphic, neural net) dynamical systems have focused on the problem of sequence control and the understanding of serial dynamics (e.g., Grossberg, 1986; Jordan, 1986, in press; Kleinfeld & Sompolinsky, 1988; Lapedes & Farber, cited in Lapedes & Farber, 1986; Pearlmutter, 1988; Rumelhart, Hinton, & Williams, 1986; Stornetta, Hogg, & Huberman, 1988; Tank & Hopfield, 1987). Such dynamics appear well-suited to the task of sequencing or orchestrating the transitions in activation among gestural primitives in a dynamical model of speech production.

Intergestural timing: A connectionist approach

Explaining how a movement sequence is generated in a connectionist computational network becomes primarily a matter of explaining the patterning of activity over time among the network's processing elements or *nodes*. This patterning occurs through cooperative and competitive interactions among the nodes themselves. Each node can store only a small amount of information (typically only a few *marker bits* or a single scalar *activity-level*) and is capable of only a few simple arithmetic or logical actions. Consequently, the interactions are conducted, not through individual programs or symbol strings, but through very simple messages—signals limited to variations in strength. Such networks, in which the transmission of symbol strings between nodes is minimal or nonexistent, depend for their success on the availability and attunement of the right connections among the nodes (e.g., Ballard, 1986; Fahlman & Hinton, 1987; Feldman & Ballard, 1982; Grossberg, 1982; Rumelhart, Hinton, & McClelland, 1986). The knowledge constraining the performance of a serial activity, including coarticulatory patterning, is embodied in these connections rather than stored in specialized memory banks. That is, the structure and dynamics of the network govern the movement as it evolves, and knowledge of the movement's time course never appears in an explicit, declarative form.

In connectionist models, the plan for a sequence is static and timeless, and is identified with a set of input units. Output units in the network are assumed to represent the control elements of the movement components and to affect these elements in direct proportion to the level of output-unit activation. One means of producing temporal ordering is to (a) establish an activation-level gradient through lateral inhibition among the output units so that those referring to earlier aspects of the sequence are more active than those referring to later aspects; and (b) inhibit output units once a threshold value of activation is achieved (e.g., Grossberg, 1978). Such connectionist systems, however, have difficulty producing sequences in which movement components are repeated (e.g., Rumelhart & Norman, 1982). In fact, a general awkwardness in dealing with the sequential control of network activity has been acknowledged as a major shortcoming of most current connectionist models (e.g., Hopfield & Tank, 1986). Some promising developments have been reported that address such criticisms (e.g., Grossberg 1986; Jordan, 1986, in press; Kleinfeld & Sompolinsky, 1988; Lapedes & Farber, cited in Lapedes & Farber, 1986; Pearlmutter, 1988; Rumelhart, Hinton, & Willams, 1986; Stornetta, Hogg, & Huberman, 1988; Tank & Hopfield, 1987). We now describe in detail one such development (Jordan, 1986, in press).

Serial dynamics: A representative model

Jordan's (1986, in press) connectionist model of serial order can be used to define a time-invariant dynamical system with an intrinsic time scale that spans the performance of a given output sequence. There are three levels in his model (see Figure 10). At the lowest level are *output* units. Even if a particular output unit is activated repeatedly in an intended sequence, it is represented by only one unit. Thus, the model adopts a *type* rather than token representation scheme for sequence elements. In the context of the present article, a separate output unit would exist for each distinct gesture in a sequence. The tuning and gating consequences of gestural activation described earlier (see the *Active Gestural Control* section) are consistent with Jordan's suggestion that "the output of the network is best thought of as influencing articulator trajectories indirectly, by setting parameters or providing boundary conditions for lower level processes which have their own inherent dynamics" (Jordan, 1986, p. 23). For example, in the repetitive sequence /bababa.../, there would be (as a first approximation) only two output units, even though each unit potentially could be activated an indefinite number of times as the sequence continues. In this example, the

output units define the activation coordinates for the consonantal bilabial gesture and the vocalic tongue-dorsum gesture, respectively. The values of the output units are the activation values of the associated gestures, and can vary continuously across a range normalized from zero (the associated gestural unit is inactive) to one (the associated gestural unit is maximally active).

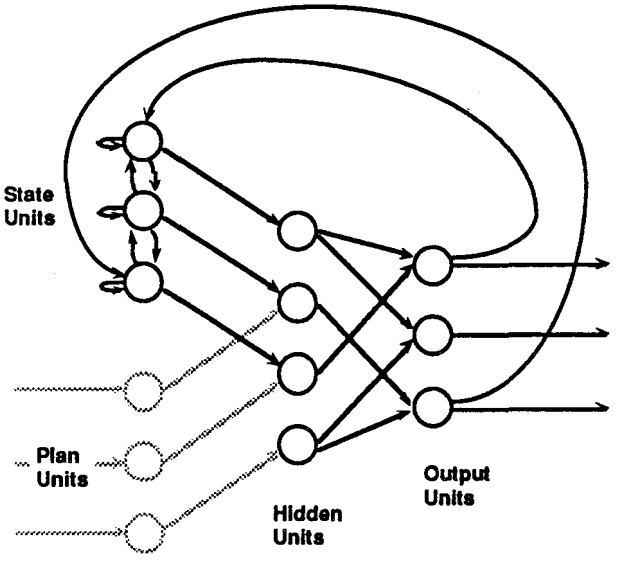

Figure 10. **Basic network architecture for Jordan's (1986, in press) connectionist model of serial order (not all connections are shown). The plan units and their connections (indicated in light grey) are not used in our proposed *hybrid* model for the serial dynamics of speech production (see text and footnote [8] for details).**

At the highest level of Jordan's model are the *state units* that, roughly speaking, define among themselves a dynamical flow with an intrinsic time scale specific to the intended sequence. These state-unit dynamics are defined by an equation of motion (the *next-state function*) that is implemented in the model by weighted recurrent connections among the state units themselves, and from the output units to the state units. Finally, at an intermediate level of the model are a set of *hidden units*. These units are connected to both the state units and the output units by two respective layers of weighted paths, thereby defining a nonlinear mapping or *output function* from state units to output units. The current vector of output activations is a function of the preceding state, which is itself a function of the previous state and previous output, and so on. Thus, the patterning over time of onsets and offsets for the output units does not arise as a consequence of direct connections among these units. Rather, such relative timing is an emergent property of the dynamics of the network as a whole. Temporal ordering among the output elements of a gestural sequence is an implicit consequence of the network architecture (i.e., the input-output functions of the system elements, and the pattern of connections among these elements) and the sequence-specific set of constant values for the weights associated with each connection path.[8]

The network can "learn" a different set of weight values for each intended utterance in Jordan's (1986; in press) model, using a "teaching" procedure that incorporates the *generalized delta rule* (*back propagation* method) of Rumelhart, Hinton, & Williams (1986). According to this rule, error signals generated at the output units (defined by the difference between the current output vector and a "teaching" vector of desired activation values) are projected back into the network to allow the hidden units to change their weights. The weights on each pathway are changed in proportion to the size of the error being back-propagated along these pathways, and error signals for each hidden unit are computed by adding the error signals arriving at these units. Rumelhart, Hinton, & Williams (1986) showed that this learning algorithm implements essentially a gradient search in weight space for the set of weights that allows the network to perform with a minimum sum of squared output errors.[9]

Jordan (1986; in press) reported simulation results in which activation values of the output units represented values of abstract phonetic features such as degree of voicing, nasality, or lip rounding. The serial network was trained to produce sequences of "phonemes", in which each phoneme was defined as a particular bundle of context-independent target values for the features. These features were not used to generate articulatory movement patterns, however. After training, the network produced continuous trajectories over time for the featural values. These trajectories displayed several impressive properties. First, the desired values were attained at the required positions in a given sequence. Second, the featural trajectories showed anticipatory and carryover coarticulatory effects for each feature that were contextually dependent on the composition of the sequence as a whole. This was due to the generalizing capacity of the network, according to which similar network states tend to produce similar outputs, and the fact that the network states during production of a given phoneme are similar to the states in which

nearby phonemes are learned. Finally, the coarticulatory temporal "spreading" of a given featural target value was not unlimited. Rather, it was restricted due to the dropoff in state similarity between a given phoneme and its surrounding context.

Toward a hybrid dynamical model

Jordan's (1986; in press) serial network has produced encouraging results for understanding the dynamics of intergestural timing in speech production. However, as already discussed, his speech simulations were defined with respect to a standard list of phonetic features, and were not related explicitly to actual articulatory movement patterns. We plan to incorporate such a serial network into our speech modeling as a means of patterning the gestural activation intervals in the task-dynamic model summarized in Equation (5). The resultant hybrid dynamical system (Figure 2) for articulatory control and coordination should provide a viable basis for further theoretical developments, guided by empirical findings in the speech production literature. For example, it is clear that the hybrid model must be able to accommodate data on the consequences for intergestural timing of mechanical perturbations delivered to the articulators during speaking. Without feedback connections that directly or indirectly link the articulators to the intergestural level, a mechanical perturbation to a limb or speech articulator could not alter the timing structure of a given movement sequence. Recent data from human subjects on unimanual oscillatory movements (Kay, 1986; Kay, Saltzman, & Kelso, 1989) and speech sequences (Gracco & Abbs, in press) demonstrate that transient mechanical perturbations induce systematic shifts in the timing of subsequent movement elements. In related animal studies (see footnote [3]), transient muscle-nerve stimulation during swimming movements of a turtle's hindlimb were also shown to induce phase shifts in the locomotor rhythm. Taken together, such data provide strong evidence that functional feedback pathways exist from the articulators to the intergestural level in the control of sequential activity. These pathways will be incorporated into our hybrid dynamical model (see the lighter pathway indicated in Figure 2).

Intrinsic vs. extrinsic timing: Autonomous vs. nonautonomous dynamics

As discussed earlier (see *Gestural Activation Coordinates* section) there are two time spans associated with every gesture in the current model. The first is the gestural settling time, defined as the time required for an idealized, temporally isolated gesture to reach a certain criterion percentage of the distance from initial to target location in tract-variable coordinates. This time span is a function of the gesture's intrinsic set of dynamic parameters (e.g., damping, stiffness). The second time-span, the gestural activation interval, is defined according to a gesture's sequence-specific activation function. In the present model, gestural activation is specified as an explicit function of time in the gestural score for a given speech sequence. In the hybrid model discussed in the previous section, these activation functions would emerge as implicit consequences of the serial dynamics intrinsic to a given sequence.

These considerations may serve to clarify certain aspects of a relatively longstanding and tenacious debate on the issue of intrinsic (e.g., Fowler, 1977, 1980) versus extrinsic (e.g., Lindblom, 1983; Lindblom et al., 1987) timing control in speech production. In the framework of the current model, intragestural temporal patterns (e.g., settling times, interarticulator asynchronies in peak velocities) can be characterized unambiguously, at least for isolated gestures, as intrinsic timing phenomena. These phenomena are emergent properties of the gesture-specific dynamics implicit in the coordinative structure spanning tract-variable and articulator coordinates (Figure 2, interarticulator level). In terms of intergestural timing, the issue is not so clear and depends on one's frame of reference. If one focuses on the interarticulatory level, then all activation inputs originate from the "outside", and activation timing must be considered extrinsic with reference to this level. Activation timing is viewed as being controlled externally according to whatever type of clock is assumed to exist or be instantiated at the system's intergestural level. However, if one considers both levels within the same frame of reference then, by definition, the timing of activation becomes intrinsic to the system as a whole. Whether or not this expansion of reference frame is useful in furthering our understanding of speech timing control depends, in part, on the nature of the clock posited at the intergestural level. This issue of clock structure leads us to a somewhat more technical consideration of the relationship between intrinsic and extrinsic timing on the one hand, and autonomous and nonautonomous dynamical systems on the other hand.

For speech production, one can posit that intrinsic timing is identified with autonomous dynamics, and extrinsic timing with nonautonomous dynamics. In an autonomous dynamical system, the terms in the corresponding equation of motion are explicit functions only of the system's "internal" state variables (i.e., positions and velocities). In contrast, a nonautonomous system's equation of motion contains terms that are explicit functions of "external" clock-time, t, such as f $(t) = \cos(\omega t)$ (e.g., Haken, 1983; Thompson & Stewart, 1986). However, the autonomous-nonautonomous distinction is just as susceptible to one's selected frame of reference as is the distinction between intrinsic and extrinsic timing. The reason is that any nonautonomous system of equations can be transformed into an autonomous one by adding an equation(s) describing the dynamics of the (formerly) external clock-time variable. That is, the frame of reference for defining the overall system equation can be extended to include the dynamics of both the original nonautonomous system as well as the formerly external clock. In this new set of equations, a state of *uni*directional coupling exists between system elements. The clock variable affects, but is unaffected by, the rest of the system variables. However, when such unidirectional coupling exists and the external clock meters out time in the standard, linear time-flow of everyday clocks and watches, we feel that its inclusion as an extra equation of motion adds little to our understanding of system behavior. In these cases, the nonautonomous description probably should be retained.

In earlier versions of the present model (Kelso et al., 1986a & 1986b; Saltzman, 1986; see also Appendices 1 & 2) only temporally isolated gestures or perfectly synchronous gesture pairs were simulated. In these cases, the equations of motion were truly autonomous, because the parameters at the interarticulatory level did not vary over the time course of the simulations. The parameters in the present model, however, are time-varying functions of the activation values specified at the intergestural level in the gestural score. Hence, the interarticulatory dynamics (Equation [5]) are currently nonautonomous. Because the gestural score specifies these activation values as explicit functions of standard clock-time, little understanding is to be gained by conceptualizing the system as an autonomous one that incorporates the unidirectionally coupled dynamics of standard clock-time and the interarticulatory level. Thus, the present model most sensibly should be considered as nonautonomous. This would not be true, however, for the proposed hybrid model in which: a) clock-time dynamics are nontrivial and intrinsic to the utterance-specific serial dynamics of the intergestural level; and b) the intergestural and interarticulator dynamics mutually affect one another. In this case, we posit that much understanding is to be gained by incorporating the dynamics of both levels into a single set of bidirectionally coupled, autonomous system equations.

INTERGESTURAL COHESION

As indicated earlier in this article (e.g., in Figure 1), speech production entails the interleaving through time of gestures defined across several different articulators and tract variables. In our current simulations, the timing of activation intervals for tract-variable gestures is controlled through the gestural score. Accordingly, gestures unfold independently over time, producing simulated speech patterns much like a player piano generates music. This rule-based description of behavior in the vocal tract makes no assumptions about coordination or functional linkages among the gestures themselves. However, we believe that such linkages exist, and that they reflect the existence of dynamical coupling within certain gestural subsets. Such coupling imbues these gestural "bundles" with a temporal cohesion that endures over relatively short (e.g., sublexical) time spans during the course of an utterance.

Support for the notion of intergestural cohesion has been provided by experiments that have focused on the structure of correlated variability evidenced between tract-variable gestures in the presence of externally delivered mechanical perturbations. Correlated variability is one of the oldest concepts in the study of natural variation, and it is displayed in a system if "when slight variations in any one part occur..., other parts become modified" (Darwin 1896, p. 128). For example, in unperturbed speech it is well known that a tight temporal relation exists between the oral and laryngeal gestures for voiceless obstruents (e.g., Löfqvist & Yoshioka, 1981a). For example, word-initial aspirated /p/ (in English) is produced with a bilabial closing-opening gesture and an accompanying glottal opening-closing gesture whose peak coincides with stop release. In a perturbation study on voiceless obstruents (Munhall et al., 1986), laryngeal compensations occurred when the lower lip was perturbed during

the production of the obstruent. Specifically, if the lower lip was unexpectedly pulled downward just prior to oral closure, the laryngeal abduction gesture for devoicing was delayed. Shaiman and Abbs (1987) have also reported data consistent with this finding. Such covariation patterns indicate a temporal cohesion among gestures, suggesting to us the existence of higher order, multigesture units in speech production.

How might intergestural cohesion be conceptualized? We hypothesize that such temporal stability can be accounted for in terms of dynamical coupling structure(s) that are defined among gestural units. Such coupling has been shown previously to induce stable intergestural phase relations in a model of two coupled gestural units whose serially repetitive (oscillatory) dynamics have been explored both experimentally and theoretically in the context of rhythmic bimanual movements (e.g., Haken, Kelso, & Bunz, 1985; Kay, Kelso, Saltzman, & Schöner, 1987; Scholz, 1986; Schöner, Haken, & Kelso, 1986). This type of model also provides an elegant account of certain changes in intergestural phase relationships that occur with increases in performance rate in the limbs and, by extension, the speech articulators. In speech, such stability and change have been examined for bilabial and laryngeal sequences consisting of either the repeated syllable /pi/ or /ip/ (Kelso, Munhall, Tuller, & Saltzman, 1985; also discussed in Kelso et al., 1986a, 1986b). When /pi/ is spoken repetitively at a self-elected "comfortable" rate, the glottal and bilabial component gestures for /p/ maintain a stable intergestural phase relationship in which peak glottal opening lags peak oral closing by an amount that results in typical (for English) syllable-initial aspiration of the /p/. For repetitive sequences of /ip/ spoken at a similarly comfortable rate, peak glottal opening occurred synchronously with peak oral closing as is typical (for English) of unaspirated (or minimally aspirated) syllable-final /p/. When /pi/ was produced repetitively at a self-paced increasing rate, intergestural phase remained relatively stable at its comfort value. However, when /ip/ was scaled similarly in rate, its phase relation was maintained at its comfort value until, at a critical speaking rate, an abrupt shift occurred to the comfort phase value and corresponding acoustic pattern for the /pi/.

In the context of the model of bimanual movement, the stable intergestural phase values at the comfort rate and the phase shift observed with rate scaling are reflections of the dynamical behavior of nonlinearly coupled, higher-order oscillatory *modes*. This use of modal dynamics parallels the identification of tract-variables with mode coordinates in the present model (see Appendix 1). Recall that the dynamics of these modal tract-variables serve to organize patterns of cooperativity among the articulators in a gesture-specific manner (see the earlier section entitled *Model Articulator and Tract Variable Coordinates*). Such interarticulator coordination is shaped according to a coupling structure among the articulators that is "provided by" the tract-variable modal dynamics. By extension, patterns of intergestural coordination are shaped according to inter-tract-variable coupling structures "provided by" a set of even higher-order multigesture modes. Because tract-variables are defined as uncoupled in the present model (Equation [A1]), it seems clear that (some sort of) inter-tract-variable coupling must be introduced to simulate the multigesture functional units evident in the production of speech.[10]

Such multigesture units could play (at least) three roles in speech production. One possibility is a hierarchical reduction of degrees of freedom in the control of the speech articulators beyond that provided by individual tract-variable dynamical systems (e.g., Bernstein, 1967). A second, related possibility is that multigesture functional units are particularly well suited to attaining articulatory goals that are relatively inaccessible to individual (or uncoupled) gestural units. For example, within single gestures the associated synergistic articulators presumably cooperate in achieving local constriction goals in tract-variable space, and individual articulatory covariation is shaped by these spatial constraints. Coordination between tract-variable gestures might serve to achieve more global aerodynamic/acoustic effects in the vocal tract. Perhaps the most familiar of such between-tract-variable effects is that of voice onset time (Lisker & Abramson, 1964), in which subtle variations in the relative timing of laryngeal and oral gestures contribute to perceived contrasts in the voicing and aspiration characteristics of stop consonants.

The third possible role for multigesture units is that of phonological primitives. For example, in Browman and Goldstein's (1986) *articulatory phonology*, the phonological primitives are gestural *constellations* that are defined as "cohesive bundles" of tract-variable gestures. Intergestural cohesion is conceived in terms of the stability of relative phasing or spatiotemporal relations among gestures within a given

constellation. In some cases, constellations correspond rather closely to traditional segmental descriptions: for example, a word-initial aspirated /p/ (in English) is represented as a bilabial closing-opening gesture and a glottal opening-closing gesture whose peak coincides with stop release; a word-initial /s/ (in English) is represented as a tongue tip raising-lowering and a glottal opening-closing gesture whose peak coincides with mid-frication. In other cases, however, it is clear that Browman and Goldstein offered a perspective that is both linguistically radical and empirically conservative. They rejected the traditional notion of segment and allowed as phonological primitives only those gestural constellations that can be observed directly from physical patterns of articulatory movements. Thus, in some instances, segmental and constellation representations diverge. For example, a word-initial /sp/ cluster (unaspirated in English) is represented as a constellation of two oral gestures (a tongue-tip and bilabial constriction-release sequence) and a *single* glottal gesture whose peak coincides with mid-frication. This representation is based on the experimental observation that for such clusters only one single-peaked glottal gesture occurs (e.g., Lisker, Abramson, Cooper, & Schvey, 1969), and thus captures the language-specific phonotactic constraint (for English) that there is no voicing contrast for stops following an initial /s/. The gestural constellation representation of /sp/ is consequently viewed as superior to a more traditional segmental approach which might predict two glottal gestures for this sequence.

Our perspective on this issue is similar to that of Browman and Goldstein in that we focus on the gestural structure of speech. Like these authors, we assume that the underlying phonological primitives are context-independent cohesive "bundles" or constellations of gestures whose cohesiveness is indexed by stable patterns of intergestural phasing. However, we adopt a position that, in comparison with theirs, is both more conservative linguistically and more radical empirically. We assume that gestures cohere in bundles corresponding, roughly, to traditional segmental descriptions, and that these segmental units maintain their integrity in fluent speech. We view many context-dependent modifications of the gestural components of these units as emergent consequences of the serial dynamics of speech production. For example, we consider the single glottal gesture accompanying English word-initial /sp/ clusters to be a within-tract-variable blend of separate glottal gestures associated with the underlying /s/ and /p/ segments (see the following section for a detailed discussion of the observable kinematic "traces" left by such underlying gestures).

INTERGESTURAL TIMING PATTERNS: EFFECTS OF SPEAKING RATE AND SEQUENCE COMPOSITION

One of the working assumptions in this article is that gestural coproduction is an integral feature of speech production, and that many factors influence the degree of gestural overlap found for a given utterance. For example, a striking phenomenon accompanying increases in speaking rate or degree of casualness is that the gestures associated with temporally adjacent segments tend to "slide" into one another with a resultant increase in temporal overlap (e.g., Browman & Goldstein, in press; Hardcastle, 1985; Machetanz, 1989; Nittrouer, Munhall, Kelso, Tuller, & Harris, 1988). Such intergestural sliding occurs both between and within tract-variables, and is influenced by the composition of the segmental sequence as well as its rate or casualness of production. We turn now to some examples of the effects on intergestural sliding and blending of changes in speaking rate and sequence composition.

Speaking rate

Hardcastle (1985) showed with electropalatographic data that the tongue gestures associated with producing the (British English) consonant sequence /kl/ tend to slide into one another and increase their temporal overlap with experimentally manipulated increases in speaking rate. Many examples of interarticulator sliding were also identified some years ago by Stetson (1951). Stetson was interested in studying the changes in articulatory timing that accompany changes in speaking rate and rhythm. Particularly interesting are his scaling trials in which utterances were spoken at increasing rates. Figure 11 is one of Stetson's figures showing the time course of lip (L), tongue (T), and air pressure (A) for productions of "sap" at different speaking rates. As can be seen, the labial gesture for /p/ and the tongue gesture for /s/ are present throughout the scaling trial but their relative timing varies with increased speaking rate. By syllable 4 the tongue gesture for /s/ and the labial gesture for /p/ from the preceding syllable completely overlap, and syllable identity is altered from then on in the trial.

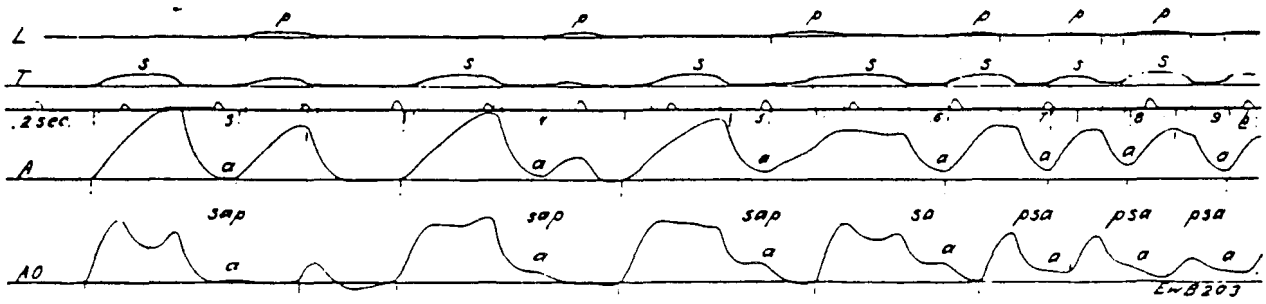

FIGURE 62. Abutting Consonants; Continuant with Stop

Syllables: *sap sap . . .*

L—Lip marker. Contact grows shorter and lighter as the rate increases and overlapping and coincidence occur.

T—Tongue marker. Well marked doubling from syl. 5-6; thereafter the single releasing compound form *ps*.

A—Air in mouth. Doubling forms, syl. 5-6.

AO—Air outside. Varied in appearance because of the high pressure during the continuant *s*. Plateau of *s* becomes mere point as compound form appears, syl. 6-7.

Figure 11. **Articulatory and aerodynamic records taken during productions of the syllable "sap" as speaking rate increases. (from Stetson, 1951; used with publisher's permission).**

In terms of the present theoretical framework, these instances of relative sliding can be described as occurring between the activation intervals associated with tongue dorsum gestures (for the velar consonant /k/), tongue tip gestures (for the alveolars /s/ and /l/), and lip aperture gestures (for the bilabial /p/). During periods of temporal overlap, the gestures sharing articulators in common are blended. Because the gestures are defined in separate tract variables, they are observably distinct in articulatory movement records. Such patterns of change might be interpretable as the response of the hybrid dynamical model discussed earlier (see *Hybrid Model* section) to hypothetically simple changes in the values of a control parameter or parameter set presumably at the model's intergestural level (see Figure 2). One goal of future empirical and simulation research is to test this notion, and if possible, to identify this parameter set and the means by which it is scaled with speaking rate.

Löfqvist & Yoshioka (1981b) have provided evidence for similar sliding and blending within tract-variables in an analysis of transillumination data on glottal devoicing gestures (abduction-adduction sequences) for a native speaker of Icelandic (a Germanic language closely related to Swedish, English, etc.). These investigators demonstrated intergestural temporal reorganization of glottal activity with spontaneous variation of speaking rate. For example, the cross-word-boundary sequence /t#k/ was accompanied by a two-peaked glottal gesture at a slow rate, but by a single-peaked gesture at fast rates. The interpretation of these data was that there were two underlying glottal gestures (one for /t/, one for /k/) at both the slow and fast rates. The visible result of only a single gesture at the fast rate appeared to be the simple consequence of blending and merging these two highly overlapping, underlying gestures defined within the same tract variable. These results have since been replicated for two speakers of North American English during experimentally controlled variations in the production rates of /s#t/ sequences (Munhall & Löfqvist, 1987).

Sequence composition: Laryngeal gestures, oral-laryngeal dominance

The rate scaling data described in the previous section for laryngeal gestures provide support for the hypothesis that the single-peaked gestures observed at fast speaking rates resulted from the sliding and blending of two underlying, sequentially adjacent gestures. In turn, this interpretation suggests a reasonable account of

glottal behavior in the production of segmental sequences containing fricative-stop clusters.

Glottal transillumination and speech acoustic data for word-final /s#ɛ/, /ks#ɛ/, and /ps#ɛ/ (unpublished data from Fowler, Munhall, Saltzman, & Hawkins, 1986a, 1986b) showed that the glottal opening-closing gesture for /s#/, in comparison to the other cases, was smaller in amplitude, shorter in duration, and peaked closer in time to the following voicing onset. These findings are consistent with the notion that a separate glottal gesture was associated with the cluster-initial stop, and that this gesture left its trace both spatially and temporally in blending with the following fricative gesture to produce a larger,[11] longer, and earlier-peaking single gestural aggregate. Other data from this experiment also indicate that the single-peaked glottal gestures observed in word-final clusters were the result of the blending of two overlapping underlying gestures. These data focus on the timing of peak glottal opening relative to the acoustic intervals (closure for /p/ or /k/, frication for /s/) associated with the production of /s#/, /ps#/, /ks#/, /sp#/, and /sk#/. For /s#/, the glottal peak occurred at mid-frication. However, for /ps#/ and /ks#/ it occurred during the first quarter of frication; for /sp#/ and /sk#/, it occurred during the third quarter of frication. These data indicate that an underlying glottal gesture was present for the /p/ or /k/ in these word-final clusters that blended with the gesture for the /s/ in a way that "pulled" or "perturbed" the peak of the gestural aggregate towards the /p/ or /k/ side of the cluster. The fact that the resultant glottal peak remained inside the frication interval for the /s/ may be ascribed, by hypothesis, to a relatively greater *dominance* over the timing of the glottal peak associated with /s/ compared to that associated with the voiceless stops /p/ or /k/.

Dominance refers to the strength of hypothesized coupling between oral acoustic/articulatory events (e.g., frication and closure intervals) and glottal events (e.g., peak glottal opening). The dominance for a voiceless consonant's oral constriction over its glottal timing appears to be influenced by (at least) two factors.[12] The first is the *manner* class of the segment: frication intervals (at least for /s/) dominate glottal behavior more strongly than stop closure intervals. This factor was highlighted previously for word-initial clusters by Browman and Goldstein (1986; cf., Kingston's [in press] related use of oral-laryngeal "binding" and Goldstein's [in press] reply to Kingston). As just discussed, this factor also appears to influence glottal timing in word-final clusters. The second factor is the presence of a preceding word-initial boundary: word-initial consonants dominate glottal behavior more strongly than the same non-word-initial consonants. These two factors appear to have approximately additive effects, as illustrated by the following examples of fricative-stop sequences defined word- or syllable-initially and across word boundaries. In these cases, as was the case word-finally, the notion of dominance can be invoked to suggest that the single-peaked glottal gestures observed for such clusters are also blends of two underlying, overlapping gestures.

Example 1. In English, Swedish, and Icelandic (e.g., Löfqvist, 1980; Löfqvist & Yoshioka, 1980, 1981a, 1981b, 1984; Yoshioka, Löfqvist, & Hirose, 1981), word-initial /s-(voiceless)stop/ clusters and /s/ are produced with a single-peaked glottal gesture that peaks at mid-frication. Word-initial /p/ is produced with a glottal gesture peaking at or slightly before closure release. Thus, the word-initial position of the /s/ in these clusters apparently bolsters the intrinsically high "segmental" dominance of the /s/, and elimi nates the displacement of the glottal peak toward the /p/ that was described earlier for word-final clusters.

Example 2. The rate scaling study for the cross-word boundary /s#t/ sequence described earlier (Munhall & Löfqvist 1987) showed two single-peaked glottal gestures for the slowest speaking rates, one double-peaked gesture for intermediate rates, and one single-peaked gesture at the fastest rate. At the slow and intermediate rates, the first peak occurred at mid-frication for the /s#/ and the second peak occurred at closure release for the /#t/. The single peak at the fastest rate occurred at the transition between frication offset and closure onset. These patterns indicate that when the two underlying glottal gestures merged into a single-peaked blend, the peak was located at a "compromise" position between the intrinsically stronger /s/ and the intrinsically weaker /t/ augmented by its word-initial status.

Example 3. In Dutch (Yoshioka, Löfqvist, & Collier, 1982), word-initial /#p/ (voiceless, unaspirated) is produced with a glottal gesture peaking at midclosure, and the glottal peak for /#s/ and /#sp/ occurs at mid-frication. However, for /#ps/ (an allowable sequence in Dutch), the glottal peak occurs at the transition between closure offset and frication onset. Again, this suggests that when the inherently stronger /s/ is augmented by word-initial status in /#sp/, the

glottal peak cannot be perturbed away from mid-frication by the following /p/. However, when the intrinsically weaker /p/ is word-initial, the glottal peak is pulled by the /p/ from mid-frication to the closure-frication boundary.

Example 4. In Swedish (e.g., Löfqvist & Yoshioka, 1980), some word-final voiceless stops are aspirated (e.g., /k#/), and are produced with glottal gestures peaking at stop release. Word-initial /#s/ is produced with glottal peak occurring at mid-frication (see Example 1). When the cross-word-boundary sequence /k#s/ is spoken at a "natural" rate, a single glottal gesture is produced with its peak occurring approximately at mid-frication. This is consistent with the high degree of glottal dominance expected for the intrinsically stronger /s/ in a word-initial position for the /k#s/ sequence.

These examples provide support for the hypothesis that fricative-stop sequences can be associated with an underlying set of two temporally overlapping but slightly offset component glottal gestures blended into a single gestural aggregate. These examples focused on the observable kinematic "traces" evident in the timing relations between the aggregate glottal peak and the acoustic intervals of the sequence. Durational data also suggest that such single observable gestures result from a two-gesture blending process. For example, the glottal gesture for the cluster /#st/ is longer in duration than either of the gestures for /#s/ and /#t/ (McGarr & Löfqvist, 1988). A similar pattern has also been found by Cooper (1989) for word-internal, syllable-initial /#s/, /#p/, and /#sp/.

SUMMARY

We have outlined an account of speech production that removes much of the apparent conflict between observations of surface variability on the one hand, and the hypothesized existence of underlying, invariant gestural units on the other hand. In doing so, we have described progress made toward a dynamical model of speech patterning that can produce fluent gestural sequences and specify articulatory trajectories in some detail. Invariant units are posited in the form of relations between context-independent sets of gestural parameters and corresponding subsets of activation, tract-variable, and articulatory coordinates in the dynamical model. Each gesture's influence over the valving and shaping of the vocal tract waxes and wanes according to the activation strengths of the units. Variability emerges in the unfolding tract-variable and articulatory movements as a result of both the utterance-specific temporal interleaving of gestural activations, and the accompanying patterns of blending or coproduction. The relative timing of the gestures and the interarticulator cooperativity evidenced for a currently active gesture set are governed by two functionally distinct but interacting levels in the model—the intergestural and interarticulatory coordination levels, respectively. At present, the dynamics of the interarticulatory level are sufficiently well developed to offer promising accounts of movement patterns observed during unperturbed and mechanically perturbed speech sequences, and during periods of coproduction. We have only begun to explore the dynamics of the intergestural level. Yet even these preliminary considerations, grounded in developments in the dynamical systems literature, have already begun to shed light on several longstanding issues in speech science, namely, the issues of intrinsic versus extrinsic timing, the nature of intergestural cohesion, and the hypothesized existence of segmental units in the production of speech. We find these results encouraging, and look forward to further progress within this research framework.

REFERENCES

Abbs, J. H., & Gracco, V. L. (1983). Sensorimotor actions in the control of multimovement speech gestures. *Trends in Neuroscience, 6*, 393-395.

Abraham, R., & Shaw, C. (1982). *Dynamics-The geometry of behavior. Part 1: Periodic behavior.* Santa Cruz, CA: Aerial Press.

Abraham, R., & Shaw, C. (1986). Dynamics: A visual introduction. In F. E. Yates (Ed.), *Self-organizing systems: The emergence of order.* New York: Plenum Press.

Arbib, M. A. (1984). From synergies and embryos to motor schemas. In H. T. A. Whiting (Ed.), *Human motor actions: Bernstein reassessed* (pp. 545-562). New York: North-Holland.

Ballard, D. H. (1986). Cortical connections and parallel processing: Structure and function. *The Behavioral and Brain Sciences, 9*, 67-120.

Ballieul, J., Hollerbach, J., & Brocket, R. (1984; December). Programming and control of kinematically redundant manipulators. *Proceedings of the 23rd IEEE Conference on Decision and Control.* Las Vegas, NV.

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.

Benati, M., Gaglio, S., Morasso, P., Tagliasco, V., & Zaccaria, R. (1980). Anthropomorphic robotics. I. Representing mechanical complexity. *Biological Cybernetics, 38*, 125-140.

Bernstein, N. A. (1967). *The coordination and regulation of movements.* London: Pergamon Press. Reprinted in H. T. A. Whiting (Ed.). (1984). *Human motor actions: Bernstein reassessed.* New York: North-Holland.

Boyce, S. E. (1988). *The influence of phonological structure on articulatory organization in Turkish and English: Vowel harmony and coarticulation.* Unpublished doctoral dissertation, Department of Linguistics, Yale University.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. (in press). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology: I. Between the Grammar and the Physics of Speech*. Cambridge, England: Cambridge University Press.

Browman, C. P., Goldstein, L., Kelso, J. A. S., Rubin, P., & Saltzman, E. L. (1984). Articulatory synthesis from underlying dynamics [Abstract]. *Journal of the Acoustical Society of America, 75* (Suppl. 1), S22-S23.

Browman, C. P., Goldstein, L., Saltzman, E. L., & Smith, C. (1986). GEST: A computational model for speech production using dynamically defined articulatory gestures [Abstract]. *Journal of the Acoustical Society of America, 80* (Suppl. 1), S97.

Bullock, D., & Grossberg, S. (1988a). Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review, 95*, 49-90.

Bullock, D., & Grossberg, S. (1988b). The VITE model: A neural command circuit for generating arm and articulator trajectories. In J. A. S. Kelso, A. J. Mandell, & M. F. Schlesinger (Eds.), *Dynamic patterns in complex systems*. Singapore: World Scientific Publishers.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Cohen, M. A., Grossberg, S., & Stork, D. G. (1988). Speech perception and production by a self-organizing neural network. In Y. C. Lee (Ed.), *Evolution, learning, cognition, and advanced architectures*. Hong Kong: World Scientific Publishers.

Coker, C. H. (1976). A model of articulatory dynamics and control. *Proceedings of the IEEE, 64*, 452-460.

Cooper, A. (1989). [An articulatory description of the distribution of aspiration in English]. Unpublished research.

Darwin, C. (1896). *The origin of species*. New York: Caldwell.

Fahlman, S. E., & Hinton, G. E. (1987). Connectionist architectures for artificial intelligence. *Computer, 20*, 100-109.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 9*, 205-254.

Folkins, J. W., & Abbs, J. H. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research, 18*, 207-220.

Fowler, C. A. (1977). *Timing control in speech production*. Bloomington, IN: Indiana University Linguistics Club.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics, 8*, 113-133.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: Human Perception and Performance, 112*, 386-412.

Fowler, C. A., Munhall, K. G., Saltzman, E. L., & Hawkins, S. (1986a). Acoustic and articulatory evidence for consonant-vowel interactions [Abstract]. *Journal of the Acoustical Society of America, 80* (Suppl. 1), S96.

Fowler, C. A., Munhall, K. G., Saltzman, E. L., & Hawkins, S. (1986b). [Laryngeal movements in word-final single consonants and consonant clusters]. Unpublished research.

Gay, T. J., Lindblom, B., & Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America, 69*, 802-810.

Goldstein, L. (in press). On articulatory binding: Comments on Kingston's paper. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology: I. Between the grammar and the physics of speech*. Cambridge, England: Cambridge University Press.

Gracco, V. L., & Abbs, J. H. (1986). Variant and invariant characteristics of speech movements. *Experimental Brain Research, 65*, 156-166.

Gracco, V. L., & Abbs, J. H. (1989). Sensorimotor characteristics of speech motor sequences. *Experimental Brain Research, 75*, 586-598.

Greene, P. H. (1971). Introduction. In I. M. Gelfand, V. S. Gurfinkel, S. V. Fomin, & M. L. Tsetlin (Eds.), *Models of the structural-functional organization of certain biological systems* (pp. xi-xxxi). Cambridge, MA: MIT Press.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology* (Vol. 5, pp. 233-374). New York: Academic Press.

Grossberg, S. (1982). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*. Amsterdam: Reidel Press.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines* (Vol. 1, pp. 187-294). New York: Academic Press.

Grossberg, S., & Mingolla, E. (1986). Computer simulation of neural networks for perceptual psychology. *Behavior Research Methods, Instruments, & Computers, 18*, 601-607.

Guckenheimer, J., & Holmes, P. (1983). *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. New York: Springer-Verlag.

Haken, H. (1983). *Advanced synergetics*. Heidelberg: Springer-Verlag.

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

Hardcastle, W. J. (1981). Experimental studies in lingual coarticulation. In R. E. Asher & E. J. A. Henderson (Eds.), *Towards a history of phonetics* (pp. 50-66). Edinburgh, Scotland: Edinburgh University Press.

Hardcastle, W. J. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication, 4*, 247-263.

Harris, K. S. (1984). Coarticulation as a component of articulatory descriptions. In R. G. Daniloff (Ed.), *Articulation assessment and treatment issues* (pp. 147-167). San Diego: College Hill Press.

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer simulation*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: A model. *Science, 233*, 625-633.

Jespersen, O. (1914). *Lehrbuch der Phonetik* [Textbook of Phonetics]. Leipzig: Teubner.

Joos, M. (1948). Acoustic phonetics. *Language, 24 (SM23)*, 1-136.

Jordan, M. I. (1985). *The learning of representations for sequential performance*. Unpublished doctoral dissertation, Department of Cognitive Science and Psychology, University of California, San Diego, CA.

Jordan, M. I. (1986). *Serial order in behavior: A parallel distributed processing approach* (Tech. Rep. No. 8604). San Diego: University of California, Institute for Cognitive Science.

Jordan, M.I. (in press). Serial order: A parallel distributed processing approach. In J.L. Elman & D.E. Rumelhart (Eds.), *Advances in connectionist theory: Speech*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kay, B. A. (1986). *Dynamic modeling of rhythmic limb movements: Converging on a description of the component oscillators*. Unpublished doctoral dissertation, Department of Psychology, University of Connecticut, Storrs, CT.

Kay, B. A., Kelso, J. A. S., Saltzman, E. L., & Schöner, G. (1987). Space-time behavior of single and bimanual rhythmical movements: Data and limit cycle model. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 178-192.

Kay, B. A., Saltzman, E. L., & Kelso, J. A. S. (1989). *Steady-state and perturbed rhythmical movements: A dynamical analysis.* Manuscript submitted for publication.

Keating, P. A. (1985). CV phonology, experimental phonetics, and coarticulation. *UCLA Working Papers in Phonetics, 62,* 1-13.

Kelso, J. A. S., Munhall, K. G., Tuller, B., & Saltzman, E. L. (1985). [Phase transitions in speech production]. Unpublished research.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986a). The dynamical theory on speech production: Data and theory. *Journal of Phonetics, 14,* 29-60.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986b). Intentional contents, communicative context, and task dynamics: A reply to the commentators. *Journal of Phonetics, 14,* 171-196.

Kelso, J.A.S. & Tuller, B. (in press). Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and performance, XIII.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 812-832.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & Kay, B. A. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America, 77,* 266-280.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics, 5,* 115-133.

Kingston, J. (in press). Articulatory binding. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology: I. Between the grammar and physics of speech.* Cambridge, England: Cambridge University Press.

Klein, C. A., & Huang, C. H. (1983). Review of pseudoinverse control for use with kinematically redundant manipulators. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13,* 245-250.

Kleinfeld, D. & Sompolinsky, H. (1988). Associative neural network model for the generation of temporal patterns: Theory and application to central pattern generators. *Biophysical Journal, 54, 1039-1051.*

Krakow, R. A. (1987; November). *Stress effects on the articulation and coarticulation of labial and velic gestures.* Paper presented at the meeting of the American Speech-Language-Hearing Association, New Orleans, LA.

Kugler, P. N., Kelso, J. A. S., & Turvey, M. T. (1980). On the concept of coordinative structures as dissipative structures: I. Theoretical lines of convergence. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 3-47). New York: North-Holland.

Kugler, P. N., Kelso, J. A. S., & Turvey, M. T. (1982). On the control and coordination of naturally developing systems. In J. A. S. Kelso & J. E. Clark (Eds.), *The development of movement control and coordination* (pp. 5-78). Chichester, England: John Wiley.

Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lapedes, A., & Farber, R. (1986). *Programming a massively parallel, computation universal system: Static behavior* (Preprint No. LA-UR 86-7179). Los Alamos, NM: Los Alamos National Laboratory, .

Lennard, P. R. (1985). Afferent perturbations during "monopodal" swimming movements in the turtle: Phase-dependent cutaneous modulation and proprioceptive resetting of the locomotor rhythm. *The Journal of Neuroscience, 5,* 1434-1445.

Lennard, P. R., & Hermanson, J. W. (1985). Central reflex modulation during locomotion. *Trends in Neuroscience, 8,* 483-486.

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 217-245). New York: Springer-Verlag.

Lindblom, B., Lubker, J., Gay, T., Lyberg, B., Branderud, P., & Holmgren, K. (1987). The concept of target and speech timing. In R. Channon & L. Shockey (Eds.), *In honor of Ilse Lehiste* (pp. 161-181). Dordrecht, Netherlands: Foris Publications.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384-422.

Lisker, L., Abramson, A. S., Cooper, F. S., & Schvey, M. H. (1969). Transillumination of the larynx in running speech. *Journal of the Acoustical Society of America, 45,* 1544-1546.

Löfqvist, A. (1980). Interarticulator programming in stop production. *Journal of Phonetics, 8,* 475-490.

Löfqvist, A., & Yoshioka, H. (1980). Laryngeal activity in Swedish obstruent clusters. *Journal of the Acoustical Society of America, 68,* 792-801.

Löfqvist, A., & Yoshioka, H. (1981a). Interarticulator programming in obstruent production. *Phonetica, 38,* 21-34.

Löfqvist, A., & Yoshioka, H. (1981b). Laryngeal activity in Icelandic obstruent production. *Nordic Journal of Linguistics, 4,* 1-18.

Löfqvist, A., & Yoshioka, H. (1984). Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. *Speech Communication, 3,* 279-289.

Macchi, M. (1985). *Segmental and suprasegmental features and lip and jaw articulators.* Unpublished doctoral dissertation, Department of Linguistics, New York University, New York, NY.

Machetanz, J. (1989, May). *Tongue movements in speech at different rates.* Paper presented at the Salk Laboratory for Language and Cognitive Studies, San Diego, CA.

MacNeilage, P. F. (1970). Motor control of serial ordering of speech. *Psychological Review, 77,* 182-196.

Mattingly, I. G. (1981). Phonetic representation and speech synthesis by rule. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 415-420). Amsterdam: North-Holland.

McGarr, N. S., & Löfqvist, A. (1988). Laryngeal kinematics in voiceless obstruents produced by hearing-impaired speakers. *Journal of Speech and Hearing Research, 31,* 234-239.

Miyata, Y. (1987). Organization of action sequences in motor learning: A connectionist approach. In Proceedings of the Ninth Annual Conference of the Cognitive Science Society (pp. 496-507). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Miyata, Y. (1988). *The learning and planning of actions* (Tech. Rep. No. 8802). San Diego, CA: University of California, Institute for Cognitive Science.

Munhall, K., & Löfqvist, A. (1987). Gestural aggregation in speech. *PAW Review, 2,* 13-15.

Munhall, K. G., & Kelso, J. A. S. (1985). Phase-dependent sensitivity to perturbation reveals the nature of speech coordinative structures [Abstract]. *Journal of the Acoustical Society of America, 78* (Suppl. 1), S38.

Munhall, K. G., Löfqvist, A., & Kelso, J. A. S. (1986). Laryngeal compensation following sudden oral perturbation [Abstract]. *Journal of the Acoustical Society of America, 80* (Suppl. 1), S109.

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America, 84,* 1653-1661.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39,* 151-168.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41*, 310-320.

Pearlmutter, B. A. (1988). Learning state space trajectories in recurrent neural networks. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufmann.

Perkell, J. S. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, MA: MIT Press.

Rubin, P. E., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70*, 321-328.

Ruelle, D. (1980). Strange attractors. *The Mathematical Intelligencer, 2*, 126-137.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 45-76). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 318-362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science, 6*, 1-36.

Saltzman, E. L. (1979). Levels of sensorimotor representation. *Journal of Mathematical Psychology, 20*, 91-163.

Saltzman, E. L. (1986). Task dynamic coordination of the speech articulators: A preliminary model. *Experimental Brain Research, Ser. 15*, 129-144.

Saltzman, E. L., Goldstein, L., Browman, C. P., & Rubin, P. (1988a). Dynamics of gestural blending during speech production [Abstract]. *Neural Networks, 1*, 316.

Saltzman, E. L., Goldstein, L., Browman, C. P., & Rubin, P. (1988b). Modeling speech production using dynamic gestural structures [Abstract]. *Journal of the Acoustical Society of America, 84* (Suppl.1), S146.

Saltzman, E. L., & Kelso, J. A. S. (1983). Toward a dynamical account of motor memory and control. In R. Magill (Ed.), *Memory and control of action* (pp. 17-38). Amsterdam: North Holland.

Saltzman, E. L., & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review, 94*, 84-106.

Saltzman, E. L., Rubin, P., Goldstein, L., & Browman, C. P. (1987). Task-dynamic modeling of interarticulator coordination [Abstract]. *Journal of the Acoustical Society of America, 82* (Suppl. 1), S15.

Schöner, G., Haken, H., & Kelso, J. A. S. (1986). Stochastic theory of phase transitions in human hand movement. *Biological Cybernetics, 53*, 1-11.

Scholz, J. P. (1986). *A nonequilibrium phase transition in human bimanual movement: Test of a dynamical model*. Unpublished doctoral dissertation, Department of Psychology, University of Connecticut, Storrs, CT.

Shaiman, S., & Abbs, J. H. (1987; November). *Phonetic task-specific utilization of sensorimotor actions*. Paper presented at the meeting of the American Speech-Language-Hearing Association, New Orleans, LA.

Smith, C. L., Browman, C. P., & McGowan, R. S. (1988). Applying the Program NEWPAR to extract dynamic parameters from movement trajectories [Abstract]. *Journal of the Acoustical Society of America, 84* (Suppl. 1),S128.

Stetson, R. H. (1951). *Motor phonetics: A study of speech movements in action*. Amsterdam: North Holland. Reprinted in J. A. S. Kelso & K. G. Munhall (Eds.). (1988). *R. H. Stetson's motor phonetics: A retrospective edition*. Boston: College-Hill.

Stornetta, W. S., Hogg, T., & Huberman, B. A. (1988). A dynamical approach to temporal information processing. In D. Z. Anderson (Ed.), *Neural information processing systems*. New York: American Institute of Physics.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *Journal of Speech and Hearing Research, 16*, 397-420.

Tank, D. W. & Hopfield, J. J. (1987). Neural computation by concentrating information in time. *Proceedings of the National Academy of Sciences, USA, 84*, 1896-1900.

Thompson, J. M. T., & Stewart, H. B. (1986). *Nonlinear dynamics and chaos: Geometrical methods for engineers and scientists*. New York: Wiley.

Tiede, M. K. & Browman, C. P. (1988). [Articulatory x-ray and speech acoustic data for CV_1CV_2C sequences]. Unpublished research.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 211-265). Hillsdale, NJ: Lawrence Erlbaum Associates.

Turvey, M. T., Rosenblum, L. D., Schmidt, R. C., & Kugler, P. N. (1986). Fluctuations and phase symmetry in coordinated rhythmic movements. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 564-583.

Vatikiotis-Bateson, E. (1988). *Linguistic structure and articulatory dynamics: A cross-language study*. Bloomington, IN: Indiana University Linguistics Club.

von Holst, E. (1973). *The behavioral physiology of animal and man: The collected papers of Erich von Holst* (Vol. 1; R. Martin, trans.). London: Methuen and Co., Ltd.

Whitney, D. E. (1972). The mathematics of coordinated control of prosthetic arms and manipulators. *ASME Journal of Dynamic Systems, Measurement and Control, 94*, 303-309.

Winfree, A. T. (1980). *The geometry of biological time*. New York: Springer-Verlag.

Wing, A. M. (1980). The long and short of timing in response sequences. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 469-486). New York: North-Holland.

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics, 14*, 5-12.

Yoshioka, H., Löfqvist, A., & Collier, R. (1982). Laryngeal adjustments in Dutch voiceless obstruent production. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, 16*, 27-35.

Yoshioka, H., Löfqvist, A., & Hirose, H. (1981). Laryngeal adjustments in the production of consonant clusters and geminates in American English. *Journal of the Acoustical Society of America, 70*, 1615-1623.

FOOTNOTES

*Ecological Psychology, 1989, 1(4), 333-382.

†Department of Communicative Disorders, Elborn College, University of Western Ontario, London, Ontario.

[1]The term *gesture* is used, here and elsewhere in this article, to denote a member of a family of functionally equivalent articulatory movement patterns that are *actively* controlled with reference to a given speech-relevant goal (e.g., a bilabial closure). Thus, in our usage gesture and movement have different meanings. Although gestures are composed of articulatory movements, not all movements can be interpreted as gestures or gestural components.

[2]For example, we and others have asserted that several coordinate systems (e.g., articulatory and higher-order, goal-oriented coordinates), and mappings among these coordinate systems, must be involved implicitly in the production of speech. We have adopted one method of representing these mappings explicitly in the present model (i.e., using Jacobians and Jacobian pseudoinverses; see Appendix 2). We make no strong claim, however, as to the neural or behavioral reality of these specific methods.

[3]An analogous functional partitioning has also been suggested in recent physiological studies by Lennard (1985) and Lennard and Hermanson (1985) on cyclic swimming motions of single hindlimbs in the turtle. In this work, the authors argued for a model of the locomotor neural circuit for turtle swimming that consists of two functionally distinct but interacting components. One component, analogous to the present interarticulator level, is a central intracycle pattern generator (CIPG) that organizes the patterning of muscular activity within each locomotor cycle. The second component, analogous to the present intergestural level, is an oscillatory central timing network (CTN) that is responsible for rhythmically activating or entraining the CIPG to produce an extended sequence of cycles (see also von Holst, 1973). A related distinction between "motor" and "clock" coordinative processes, respectively, has been proposed in the context of human manual rhythmic tasks consisting of either continuous oscillations at the wrist joints (e.g., Turvey, Rosenblum, Schmidt, & Kugler, 1986) or discrete finger tapping sequences (e.g., Wing, 1980; Wing & Kristofferson, 1973).

[4]We do not mean to imply that the production of vocal tract constrictions and the shaping of articulatory trajectories are the primary goals of speech production. The functional role of speech gestures is to control air pressures and flows in the vocal tract so as to produce distinctive patterns of sound. In this article, we emphasize gestural form and stability as phonetic organizing principles for the sake of relative simplicity. Ultimately, the gestural approach must come to grips with the aerodynamic sound-production requirements of speech.

[5]Since the preparation of this article, the task-dynamic model was extended to incorporate control of the tongue-tip (TTCL, TTCD), glottal (GLO), and velic (VEL) constrictions. These tract-variables and associated articulator sets are also shown in Figures 3 and 4. Results of simulations using these "new" gestures have been reported elsewhere in preliminary form (Saltzman, Goldstein, Browman, & Rubin, 1988a, 1988b).

[6]Gestural activation pulses are similar functionally to Joos's (1948) theorized "innervation waves", whose ongoing values reflected the strength of vocal tract control associated with various phonological segments or segmental components. They are also analogous to the "phonetic influence functions" used by Mattingly (1981) in the domain of acoustic speech synthesis-by-rule. Finally, the activation pulses share with Fowler's (1983) notion of segmental "prominence" the property of being related to the "extent to which vocal tract activity is given over to the production of a particular segment" (p. 392).

[7]Coarticulatory effects could also originate in two simpler ways. In the first case, "passive" coproduction could result from carryover effects associated with the cessation of active gestural control, due to the inertial sluggishness or time constants inherent in the articulatory subsystems (e.g., Coker, 1976; Henke, 1966). However, neither active nor passive coproduction need be involved in coarticulatory phenomena, at least in a theoretical sense. Even if a string of segments were produced as a temporally discrete (i.e., non-coproduced) sequence of target articulatory steady-states, coarticulatory effects on articulatory movement patterns would still result. In this second case, context-dependent differences in articulatory transitions to a given target would simply reflect corresponding differences in the interpolation of trajectories from the phonologically allowable set of immediately preceding targets. Both "sluggishness" and interpolation coarticulatory effects appear to be present in the production of actual speech.

[8]In Jordan's (1986; in press) model, a given network can learn a *single* set of weights that will allow it to produce *several* different sequences. Each such sequence is produced (and learned) in the presence of a corresponding constant activation pattern in a set of *plan* units (see Figure 10). These units provide a second set of inputs to the network's hidden layer, in addition to the inputs provided by the state units. We propose, however, to use Jordan's model for cases in which different sets of weights are learned for different sequences. In such cases, the plan units are no longer required, and we ignore them in this article for purposes of simplicity.

[9]To teach the network to perform a given sequence, Jordan (1986; in press) first initialized the network to zero, and then presented a sequence of teaching vectors (each corresponding to an element in the intended sequence), delivering one every fourth time step. At these times, errors were generated, back-propagated through the network, and the set of network weights were incrementally adjusted. During the three time steps between each teaching vector, the network was allowed to run free with no imposed teaching constraints. At the end of the teaching vector sequence the network was reinitialized to zero, and the entire weight-correction procedure was repeated until the sum of the squared output errors fell below a certain criterion. After training, the network's performance was tested starting with the state units set to zero.

[10]One possibility is to construct explicitly a set of serial mini-networks that could produce sequentially cohesive, multigesture units. Then a higher order net could be trained to produce utterance-specific sequences of such units (e.g., Jordan, 1985). It is also possible that multigesture units could arise spontaneously as emergent consequences of the learning-phase dynamics of connectionist, serial-dynamic networks that are trained to produce orchestrated patterns of the simpler gestural components (e.g., Grossberg, 1986; Miyata, 1987, 1988). This is clearly an important area to be explored in the development of our hybrid dynamical model of speech production (see the section entitled *Toward a Hybrid Dynamical Model*).

[11]Transillumination signals are uncalibrated in terms of spatial measurement scale. Consequently, amplitude differences in glottal gestures are only suggested, not demonstrated, by corresponding differences in transillumination signal size. Temporal differences (e.g., durations, glottal peak timing) and the spatiotemporal shape (e.g., one vs. two peaks) of transillumination signals are reliable indices/reflections of gestural kinematics.

[12]It is likely that rate and stress manipulations also have systematic effects on oral-glottal coordination. We make no claims regarding these potential effects in this article, however.

APPENDIX 1

Tract-variable dynamical system

The tract-variable equations of motion are defined in matrix form as follows:

$$\ddot{\mathbf{z}} = \mathbf{M}^{-1}(-\mathbf{B}\,\dot{\mathbf{z}} - \mathbf{K}\Delta\mathbf{z}), \qquad \text{(A1)}$$

where $\mathbf{z}$ = the $m \times 1$ vector of current tract-variable positions, with components z_i listed in Figure 4; $\dot{\mathbf{z}}, \ddot{\mathbf{z}}$ = the first and second derivatives of $\mathbf{z}$ with respect to time; M = a $m \times m$ diagonal matrix of inertial coefficients. Each diagonal element, m_{ii}, is associated with the i^{th} tract variable; B = a $m \times m$ diagonal matrix of tract-variable damping coefficients; K = a $m \times m$ diagonal matrix of tract-variable stiffness coefficients; and $\Delta\mathbf{z} = \mathbf{z} - \mathbf{z_o}$ where $\mathbf{z_o}$ = the target or rest position vector for the tract variables.

By defining the M, B, and K matrices as diagonal, the equations in (A1) are uncoupled. In this sense, the tract variables are assumed to represent independent *modes* of articulatory behavior that do not interact dynamically (see Coker, 1976, for a related use of articulatory modes). In current simulations, M is assumed to be constant and equal to the identity matrix (m_{ij} = 1.0 for $i = j$, otherwise m_{ij} = 0.0), whereas the components of B, K, and $\mathbf{z_o}$ vary during a simulated utterance according to the ongoing set of gestures being produced. For example, different vowel gestures are distinguished in part by corresponding differences in target positions for the associated set of tongue-dorsum point attractors. Similarly, vowel and consonant gestures are distinguished in part by corresponding differences in stiffness coefficients, with vowel gestures being slower (less stiff) than consonant gestures. Thus, Equation (A1) describes a linear system of tract-variable equations with time-varying coefficients, whose values are functions of the currently active gesture set (see the *Parameter Tuning* subsection of the text section *Active Gestural Control: Tuning and Gating* for a detailed account of this coefficient specification process). Note that simulations reported previously in Saltzman (1986) and Kelso et al. (1986a, 1986b) were restricted to either single "isolated" gestures, or synchronous pairs of gestures defined across different tract variables, e.g., single bilabial closures, or synchronous "vocalic" tongue-dorsum and "consonantal" bilabial gestures. In these instances, the coefficient matrices and vector parameters in Equation (A1) remained constant (time-invariant) throughout each such gesture set.

APPENDIX 2

Model articulator dynamical system; Orthogonal projection operator

A dynamical system for controlling the model articulators is specified by expressing tract variables ($\mathbf{z}$, $\dot{\mathbf{z}}$, $\ddot{\mathbf{z}}$) as functions of the corresponding model articulator variables ($\boldsymbol{\phi}$, $\dot{\boldsymbol{\phi}}$, $\ddot{\boldsymbol{\phi}}$). The tract variables of Equation (A1) are transformed into model articulator variables using the following direct kinematic relationships:

$$\mathbf{z} = \mathbf{z}(\boldsymbol{\phi}) \qquad \text{(A2a)}$$

$$\dot{\mathbf{z}} = \mathbf{J}(\boldsymbol{\phi})\,\dot{\boldsymbol{\phi}} \qquad \text{(A2b)}$$

$$\ddot{\mathbf{z}} = \mathbf{J}(\boldsymbol{\phi})\,\ddot{\boldsymbol{\phi}} + \dot{\mathbf{J}}(\boldsymbol{\phi}, \dot{\boldsymbol{\phi}})\dot{\boldsymbol{\phi}}, \qquad \text{(A2c)}$$

where $\boldsymbol{\phi}$ = the $n \times 1$ vector of current articulator positions, with components ϕ_j listed in Figure 4; $\mathbf{z}(\boldsymbol{\phi})$ = the current $m \times 1$ tract-variable position vector expressed as a function of the current model articulator configuration. These functions are specific to the particular geometry assumed for the set of model articulators used to simulate speech gestures or produce speech acoustics via articulatory synthesis. $\mathbf{J}(\boldsymbol{\phi})$ = the $m \times n$ *Jacobian* transformation matrix whose elements J_{ij} are partial derivatives, $\partial z_i/\partial \phi_j$, evaluated at the current $\boldsymbol{\phi}$. Thus, each row-i of the Jacobian represents the set of changes in the i^{th} tract variable resulting from unit changes in all the articulators; and $\dot{\mathbf{J}}(\boldsymbol{\phi}, \dot{\boldsymbol{\phi}}) = (d\mathbf{J}(\boldsymbol{\phi})/dt)$, a $m \times n$ matrix resulting from differentiating the elements of $\mathbf{J}(\boldsymbol{\phi})$ with respect to time. The elements of $\dot{\mathbf{J}}$ are functions of both the current $\boldsymbol{\phi}$ and $\dot{\boldsymbol{\phi}}$. The elements of J and $\dot{\mathbf{J}}$ thus reflect the geometrical relationships among motions of the model articulators and motions of the corresponding tract variables. Using the direct kinematic relationships in Equation (A2), the equation of motion derived for the actively controlled model articulators is as follows:

$$\ddot{\boldsymbol{\phi}}_A = \mathbf{J}^*(\mathbf{M}^{-1}[-\mathbf{B}\mathbf{J}\,\dot{\boldsymbol{\phi}} - \mathbf{K}\Delta\mathbf{z}(\boldsymbol{\phi})]) - \mathbf{J}^*\,\dot{\mathbf{J}}\dot{\boldsymbol{\phi}}, \qquad \text{(A3)}$$

where $\boldsymbol{\phi}_A$ = an articulatory acceleration vector representing the active driving influences on the model articulators; M, B, K, J, and $\dot{\mathbf{J}}$ are the same matrices used in Equations (A1) and (A2); $\Delta\mathbf{z}(\boldsymbol{\phi}) = \mathbf{z}(\boldsymbol{\phi}) - \mathbf{z_o}$, where $\mathbf{z_o}$ = the same constant vector used in Equation (A1); It should be noted that because $\Delta\mathbf{z}$ in Equations (A1) and (A3) is *not* assumed to be "small," a differential approximation $d\mathbf{z} = \mathbf{J}(\boldsymbol{\phi})d\boldsymbol{\phi}$ is not justified and,

therefore, Equation (A2a) was used instead for the kinematic displacement transformation into model articulator variables; J^* = a $n \times m$ *weighted Jacobian pseudoinverse* (e.g., Benati, Gaglio, Morasso, Tagliasco, & Zaccaria, 1980; Klein & Huang, 1983; Whitney, 1972). $J^* = W^{-1}J^T(JW^{-1}J^T)^{-1}$ where W is a $n \times n$ positive definite *articulatory weighting* matrix whose elements are constant during a given isolated gesture, and superscript T denotes the vector or matrix *transpose* operation. The pseudoinverse is used because there are a greater number of model articulator variables than tract variables for this task. More specifically, using J^* provides a unique, optimal least squares solution for the *redundant* (e.g., Saltzman, 1979) differential transformation from tract variables to model articulator variables that is weighted according to the pattern of elements in the W-matrix. In current modeling, the W-matrix is defined to be of diagonal form, in which element w_{jj} is associated with articulator ϕ_j. A given set of articulator weights implements a corresponding pattern of constraints on the relative motions of the articulators during a given gesture. The motion of a given articulator is constrained in direct proportion to the magnitude of the corresponding weighting element relative to the remaining weighting elements. Intuitively, then, the elements of W establish a gesture-specific pattern of relative "receptivities" among the articulators to the driving influences generated in the tract-variable state space. In the present model, J^* has been generalized to a form whose elements are gated functions of the currently active gesture set (see the *Transformation gating* subsection of the text section *Active gestural control: Tuning and gating* for details).

In Equation (A3), the first and second terms inside the inner parentheses on the right hand side represent the articulatory acceleration components due to system damping ($\ddot{\phi}_d$) and stiffness ($\ddot{\phi}_s$), respectively. The rightmost term on the right hand side represents an acceleration component vector ($\ddot{\phi}_{vp}$) that is nonlinearly proportional to the squares and pairwise products of current articulatory velocities (e.g., $(\dot{\phi}_2)^2$, $\dot{\phi}_2\dot{\phi}_3$, etc.; for further details, see Kelso et al., 1986a, 1986b; Saltzman, 1986; Saltzman & Kelso, 1987).

In early simulations of unperturbed discrete speech gestures (e.g., bilabial closure) it was found that, after a given gestural target (e.g., degree of lip compression) was attained and maintained at a steady value, the articulators continued to move with very small but non-negligible (and undesirable) velocities. In essence, the model added to the articulator movements just those patterns that resulted in no tract-variable (e.g., lip aperture) motion above and beyond that demanded by the task. The source of this residual motion was ascertained to reside in the nonconservative nature of the pseudoinverse (J^*; see Equation [A3]) of the Jacobian transformation (J) used to relate tract-variable motions and model articulator motions (Klein & Huang, 1983). By nonconservative, we mean that a closed path in tract-variable space does not imply generally a closed path in model articulator space.

These undesired extraneous model-articulator motions were eliminated by including supplementary dissipative forces proportional to the articulatory velocities. Specifically, the *orthogonal projection operator*, $(I_n - J^*J])$, where I_n is a $n \times n$ identity matrix (Ballieul, Hollerbach & Brocket, 1984; Klein & Huang, 1983) was used in the following augmented form of Equation (A3):

$$\ddot{\phi}_A = J^*(M^{-1}[- BJ\dot{\phi} - K\Delta z(\phi)]) - J^*\dot{J}\dot{\phi} + (I_n - J^*J))\ddot{\phi}_d, \quad \text{(A4)}$$

where $\ddot{\phi}_d = B_N\dot{\phi}$ represents an acceleration damping vector, and B_N is a $n \times n$ diagonal matrix whose components, b_{Njj}, serve as constant damping coefficients for the j^{th} component of $\ddot{\phi}$. The subscript N denotes the fact that B_N is the same damping matrix as that used in the articulatory *neutral* attractor (see the text section on *Nonactive Gestural Control,* Equations [4] and [5]).

Using Equation (A4), the model generates movements to tract-variable targets with no residual motions in either tract-variable or model-articulator coordinates. Significantly, the model works equally well for both the case of unperturbed gestures, and the case in which gestures are perturbed by simulated external mechanical forces (see the text section *Gestural primitives*). In the present model, the identity matrix (I_n) in Equation (A4) has been generalized, like J^*, to a form whose elements are gated functions of the currently active gesture set (see the *Transformation gating* subsection of the text section *Active gestural control: Tuning and gating*).

The damping coefficients of B_N are typically assigned equal values for all articulators. This results in synchronous movements (varying in

amplitudes) for the tract variables and articulators involved in isolated gestures. Interesting patterns emerge, however, if the coefficients are assumed to be unequal for the various articulators (Saltzman et al., 1987). For example, the relatively sluggish rotations of the jaw or horizontal motions of the lips may be characterized by larger time constants than the lips' relatively brisk vertical motions. Implementing these asymmetries into B_N, interarticulator asynchronies within single speech gestures are generated by the model that mirror, partially, some patterns reported in the literature. For example, Gracco and Abbs (1986) showed that, during bilabial closing gestures for the first /p/ in /sæpæpl/, the raising onsets and peak velocities of the component articulatory movements occur in the order: upper lip, lower lip, and jaw. The peak velocities conform to this order more closely than the raising onsets. In current simulations of isolated bilabial gestures, the asynchronous pattern of the peak velocities (but not the movement onsets) emerges naturally when the elements of B_N are unequal. Interestingly, the tract-variable trajectories are identical to those generated when B_N's elements are equal. Additional simulations have revealed that patterns of closing onsets may become asynchronous, however, depending on several factors, e.g., the direction and magnitude of the jaw's velocity prior to the onset of the closing gesture.

APPENDIX 3

Competitive network equations for parameter tuning

The *post*blending activation strengths (p_{Tik} and p_{Wikj}) defined in text Equation (2) are given by the steady-state solutions to a set of feedforward, competitive-interaction-network dynamical equations (e.g., Grossberg, 1986) for the preblending activation strengths (a_{ik}) in the present model. These equations are expressed as follows:

$$p_{Tik} = -a_{ik}\left(p_{Tik} - B_p\right) - \left(\left[B_a - a_{ik}\right] + \beta_{ik} \sum_{\substack{\ell \in Z_i \\ \ell \neq k}} \left[\alpha_{i\ell} a_{i\ell}\right] \right) p_{Tik} \text{ , and} \tag{A5a}$$

$$p_{Wikj} = -a_{ik}\left(p_{Wikj} - B_p\right) - \left(\left[B_a - a_{ik}\right] + \beta_{ik} \sum_{i \in \Phi_j} \left[\sum_{\substack{\ell \in Z_i \\ \ell \neq k}} a_{i\ell} a_{i\ell} \right] \right) p_{Wikj} \text{ ,} \tag{A5b}$$

where B_p and B_a denote the maximum values allowed for the pre-blending and post-blending activation strengths, respectively. In current modeling, B_p and B_a are defined to equal 1.0; and a_{il} and β_{ik} are the lateral inhibition and "gatekeeper" coefficients, respectively, defined in text Equation (2).

The solutions to Equations (A5a) and (A5b) are obtained by setting their left-hand sides to zero, and solving for p_{Tik} and p_{Wikj}, respectively. These solutions are expressed in Equations (2a) and (2b). The dynamics of Equation (A5) are assumed to be "fast" relative to the dynamics of the interarticulator coordination level (Equations [A3] and [A4]). Consequently, incorporating the solutions of Equation (A5) directly into Equation (1) is viewed as a justified computational convenience in the present model (see also Grossberg & Mingolla, 1986, for a similar computational simplification).