7 Dynamics and Articulatory Phonology

Catherine P. Browman and Louis Goldstein

EDITORS' INTRODUCTION

Linguists studying the sound *of utterances distinguish between the strictly* physical *aspects of speech and its production, on the one hand, and its basic* linguistic *properties on the other. The difference here can be illustrated by two utterances of* Here it is, *one produced by Lurch, the laconic butler from The Addams Family, and the other by a child discovering an Easter egg. At the* phonetic *level—the level of the physical sounds—these differ enormously, but at a higher and more abstract* phonological *level they consist of the same sound units (known as* phonemes*) assembled in the same order.*

Developing good theories of phonetics, phonology, and the relation between them are central parts of linguistics, but these efforts are important to cognitive science as well. Somehow we manage to produce *utterances—to speak—and how we can do this cries out for explanation. The standard assumption is that the* phonological *level is basic as far as cognitive processes are concerned; the output of the cognitive system is a phonological specification of what it is one wants to say. Actually speaking involves using one's vocal mechanisms to translate a phonological specification into a stream of sound.*

Mainstream computational cognitive science assumes that cognitive processes are a matter of processing symbols inside the head. Consequently, it makes the assumption that phonemes are represented in the mind/brain by symbols *of basically the same kind as those used by linguists when they write about phonemes. Thus, linguists represent the phoneme /t/ by means of the symbol [t^h]; computational cognitive science assumes that when you produce an utterance involving this sound, the cognitive system delivers a similar symbol (though in "mentalese") to the motor system, which drives the vocal apparatus to produce the actual sound. (In more detailed versions, the phonemic symbol is more complex; it is a data structure specifying the presence or absence of more basic* features.*)*

This approach turns out to have some deep problems, grounded in the fact that the symbols of phonology are so different from the actual physical processes that constitute speaking. One problem is figuring out the nature of the relationship between phonological specifications and the resulting sounds that the motor system must somehow implement. Another problem is in the nature of the implementation

device itself. How does it translate from the sequence of static *symbols, which are output by the cognitive system, into the* dynamical *processes, which constitute speaking? How does it get from atemporal symbols to real speech, which has an extraordinarily subtle and complex temporal character?*

Browman and Goldstein do not solve *these problems; rather, they* avoid *them by offering a fundamentally different picture of phonology and its relationship with the physical processes of speaking. In their approach, known as* articulatory phonology, *the fundamental units are not abstract units of sound, represented by mental symbols, but rather basic coordinated* gestures *of the speech system. These gestures are high-level descriptions of a single complex dynamical system whose behaviors, at a lower level, constitute the articulatory processes of sound production. Consequently in articulatory phonology there is no deep incommensurability between the phonological and phonetic levels to be overcome. The basic units of phonology are themselves dynamic events of the same kind (though at a higher level) as the physical processes of speech production.*

In this chapter, Browman and Goldstein give an overview of the articulatory phonology approach, and describe its implementation in a speech production system for English. In this system a high-level gestural score drives a dynamical system which organizes movements of components of the articulatory system (in the manner described by Elliot Saltzman in chapter 6). The specifications of these movements are then fed into a sound synthesizer which produces the physical sound itself. (Note that in this chapter they describe this system as a computational model, but by this they mean simulated on a computer *rather than* a model of computational processes.*)*

This work illustrates a number of general characteristics of the dynamical approach to cognition. For example, it rejects the traditional assumptions that cognitive processes and bodily processes are fundamentally different in kind, and that cognition is "inner" while bodily movement is "outer." Articulatory phonology breaks down the difference in kind by reconceptualizing the basic units of cognition as behaviors of a dynamical system, and so as essentially temporal in nature. By making this move, this dynamical approach overcomes problems of embeddedness that plague standard computational cognitive science.

7.1 INTRODUCTION

Traditionally, the study of human speech and its patterning has been approached in two different ways. One way has been to consider it as mechanical or biomechanical activity (e.g., of articulators or air molecules or cochlear hair cells) that changes continuously in time. The other way has been to consider it as a linguistic (or cognitive) structure consisting of a sequence of elements chosen from a closed inventory. Development of the tools required to describe speech in one or the other of these approaches has proceeded largely in parallel, with one hardly informing the other at all (some notable exceptions are discussed below). As a result, speech has been seen as having two structures, one considered physical, and the other cognitive, where the

relation between the two structures is generally not an intrinsic part of either description. From this perspective, a complete picture requires "translating" between the intrinsically incommensurate domains (as argued by Fowler, Rubin, Remez, et al. 1980).

The research we have been pursuing (Browman and Goldstein, 1986, 1989, 1990a,b, 1992) ("articulatory phonology") begins with the very different assumption that these apparently different domains are, in fact, the low- and high-dimensional descriptions of a single (complex) system. Crucial to this approach is identification of phonological units with dynamically specified units of articulatory action, called *gestures*. Thus, an utterance is described as an act that can be decomposed into a small number of primitive units (a low-dimensional description), in a particular spatiotemporal configuration. The same description also provides an intrinsic specification of the high-dimensional properties of the act (its various mechanical and biomechanical consequences).

In this chapter, we briefly examine the nature of the low- and high-dimensional descriptions of speech, and contrast the dynamical perspective that unifies these with other approaches in which they are separated as properties of mind and body. We then review some of the basic assumptions and results of developing a specific model incorporating dynamical units, and illustrate how it provides both low- and high-dimensional descriptions.

7.2 DIMENSIONALITY OF DESCRIPTION

Human speech events can be seen as quite complex, in the sense that an individual utterance follows a continuous trajectory through a space defined by a large number of potential degrees of freedom, or dimensions. This is true whether the dimensions are neural, articulatory, acoustic, aerodynamic, auditory, or otherwise describable. The fundamental insight of phonology, however, is that the pronunciation of the words in a given language may differ from (i.e., contrast with) one another in only a restricted number of ways: the number of degrees of freedom actually employed in this contrastive behavior is far fewer than the number that is mechanically available. This insight has taken the form of the hypothesis that words can be decomposed into a small number of primitive units (usually far fewer than a hundred in a given language) which can be combined in different ways to form the large number of words required in human lexicons. Thus, as argued by Kelso, Saltzman, and Tuller (1986), human speech is characterized not only by a high number of potential (microscopic) degrees of freedom but also by a low-dimensional (macroscopic) form. This macroscopic form is usually called the "phonological" form. As suggested below, this collapse of degrees of freedom can possibly be understood as an instance of the kind of self-organization found in other complex systems in nature (Haken, 1977; Kugler and Turvey, 1987; Madore and Freedman, 1987; Schöner and Kelso, 1988; Kauffmann, 1991).

Historically, however, the gross differences between the macroscopic and microscopic scales of description have led researchers to *ignore one or the*

other description, or to assert its irrelevance, and hence to generally separate the cognitive and the physical. Anderson (1974) describes how the development of tools in the 19th and early 20th centuries led to the quantification of more and more details of the speech signal, but "with such increasingly precise description, however, came the realization that much of it was irrelevant to the central tasks of linguistic science" (p. 4). Indeed, the development of many early phonological theories (e.g., those of Saussure, Trubetzkoy, Sapir, Bloomfield) proceeded largely without any substantive investigation of the measurable properties of the speech event at all (although Anderson notes Bloomfield's insistence that the smallest phonological units must ultimately be defined in terms of some measurable properties of the speech signal). In general, what was seen as important about phonological units was their *function*, their ability to distinguish utterances.

A particularly telling insight into this view of the lack of relation between the phonological and physical descriptions can be seen in Hockett's (1955) familiar Easter egg analogy. The structure serving to distinguish utterances (for Hockett, a sequence of letter-sized phonological units called phonemes) was viewed as a row of colored, but unboiled, Easter eggs on a moving belt. The physical structure (for Hockett, the acoustic signal) was imagined to be the result of running the belt through a wringer, effectively smashing the eggs and intermixing them. It is quite striking that, in this analogy, the cognitive structure of the speech event cannot be seen in the gooey mess itself. For Hockett, the only way the hearer can respond to the event is to infer (on the basis of obscured evidence, and knowledge of possible egg sequences) what sequence of eggs might have been responsible for the mess. It is clear that in this view, the relation between cognitive and physical descriptions is neither systematic nor particularly interesting. The descriptions share color as an important attribute, but beyond that there is little relation.

A major approach that did take seriously the goal of unifying the cognitive and physical aspects of speech description was that presented in the *Sound Pattern of English* (Chomsky and Halle, 1968), including the associated work on the development of the theory of distinctive features (Jakobson, Fant, and Halle, 1951) and the quantal relations that underlie them (Stevens, 1972, 1989). In this approach, an utterance is assigned two representations: a "phonological" one, whose goal is to describe how the utterance functions with respect to contrast and patterns of alternation, and a "phonetic" one, whose goal is to account for the grammatically determined physical properties of the utterance. Crucially, however, the relation between the representations is quite constrained: both descriptions employ exactly the same set of dimensions (the features). The phonological representation is coarser in that features may take on only binary values, while the phonetic representation is more fine-grained, with the features having scalar values. However, a *principled* relation between the binary values and the scales is also provided: Stevens's quantal theory attempts to show how the potential continuum of scalar feature values can be intrinsically partitioned into categorical regions, when the mapping from articulatory dimensions to auditory properties is considered.

Further, the existence of such quantal relations is used to explain why languages employ these particular features in the first place.

Problems raised with this approach to speech description soon led to its abandonment, however. One problem is that its phonetic representations were shown to be inadequate to capture certain systematic physical differences between utterances in different languages (Ladefoged, 1980; Port, 1981; Keating, 1985). The scales used in the phonetic representations are themselves of reduced dimensionality, when compared to a complete physical description of utterances. Chomsky and Halle (1968) hypothesized that such further details could be supplied by universal rules. However, the above authors (also Browman and Goldstein, 1986) argued that this would not work—the same phonetic representation (in the Chomsky and Halle sense) can have different physical properties in different languages. Thus, more of the physical detail (and particularly details having to do with timing) would have to be specified as part of the description of a particular language. Ladefoged's (1980) argument cut even deeper. He argued that there is a system of scales that is useful for characterizing the measurable articulatory and acoustic properties of utterances, but that these scales are very different from the features proposed by Chomsky and Halle.

One response to these failings has been to hypothesize that descriptions of speech should include, in addition to phonological rules of the usual sort, rules that take (cognitive) phonological representations as input and convert them to physical parameterizations of various sorts. These rules have been described as rules of "phonetic implementation" (e.g., Klatt, 1976; Port, 1981; Keating, 1985; Liberman and Pierrehumbert, 1984; Keating, 1990; Pierrehumbert, 1990). Note that in this view the description of speech is divided into two separate domains involving distinct types of representations: the phonological or cognitive structure and the phonetic or physical structure. This explicit partitioning of the speech side of linguistic structure into separate phonetic and phonological components which employ distinct data types that are related to one another only through rules of phonetic implementation (or "interpretation") has stimulated a good deal of research (e.g., Liberman and Pierrehumbert, 1984; Fourakis and Port, 1986; Keating, 1988; Cohn, 1990; Coleman, 1992). However, there is a major price to be paid for drawing such a strict separation: it becomes very easy to view phonetic and phonological (physical and cognitive) structures as essentially independent of one another, with no interaction or mutual constraint. As Clements (1992) describes the problem: "The result is that the relation between the phonological and phonetic components is quite unconstrained. Since there is little resemblance between them, it does not matter very much for the purposes of phonetic interpretation what the form of the phonological input is; virtually any phonological description can serve its purposes equally well" (p. 192). Yet, there is a constrained relation between the cognitive and physical structures of speech, which is what drove the development of feature theory in the first place.

In our view, the relation between the physical and cognitive, i.e., the phonetic and phonological, aspects of speech is inherently constrained by their being simply two levels of description—the microscopic and macroscopic—of the same system. Moreover, we have argued that the relation between microscopic and macroscopic properties of speech is one of *mutual* or *reciprocal* constraint (Browman and Goldstein, 1990b). As we elaborated, the existence of such reciprocity is supported by two different lines of research. One line has attempted to show how the macroscopic properties of contrast and combination of phonological units arise from, or are constrained by, the microscopic, i.e., the detailed properties of speech articulation and the relations between speech articulation, aerodynamics, acoustics, and audition (e.g., Stevens, 1972, 1989; Lindblom, MacNeilage, and Studdert-Kennedy, 1983; Ohala, 1983). A second line has shown that there are constraints running in the opposite direction, such that the (microscopic) detailed articulatory or acoustic properties of particular phonological units are determined, in part, by the macroscopic system of contrast and combination found in a particular language (e.g., Wood, 1982; Ladefoged, 1982; Manuel and Krakow, 1984; Keating, 1990). The apparent existence of this bidirectionality is of considerable interest, because recent studies of the generic properties of complex physical systems have demonstrated that reciprocal constraint between macroscopic and microscopic scales is a hallmark of systems displaying "self-organization" (Kugler and Turvey, 1987; see also discussions by Langton in Lewin, 1992, pp. 12–14, 188–191; and work on the emergent properties of "co-evolving" complex systems: Hogeweg, 1989; Kauffman, 1989; Kauffman and Johnsen, 1991; Packard, 1989).

Such self-organizing systems (hypothesized as underlying such diverse phenomena as the construction of insect nests and evolutionary and ecological dynamics) display the property that the "local" interactions among a large number of microscopic system components can lead to emergent patterns of "global" organization and order. The emergent global organization also places constraints on the components and their local interactions. Thus, self-organization provides a principled linkage between descriptions of different dimensionality of the same system: the high-dimensional description (with many degrees of freedom) of the local interactions and the low-dimensional description (with few degrees of freedom) of the emergent global patterns. From this point of view, then, speech can be viewed as a single complex system (with low-dimensional macroscopic and high-dimensional microscopic properties) rather than as two distinct components.

A different recent attempt to articulate the nature of the constraints holding between the cognitive and physical structures can be found in Pierrehumbert (1990), in which the relation between the structures is argued to be a "semantic" one, parallel to the relation that obtains between concepts and their real-world denotations. In this view, macroscopic structure is constrained by the microscopic properties of speech and by the principles guiding human cognitive category formation. However, the view fails to account for the

apparent bidirectionality of the constraints. That is, there is no possibility of constraining the microscopic properties of speech by its macroscopic properties in this view. (For a discussion of possible limitations to a dynamic approach to phonology, see Pierrehumbert and Pierrehumbert, 1990.)

The articulatory phonology that we have been developing (e.g., Browman and Goldstein, 1986, 1989, 1992) attempts to understand phonology (the cognitive) as the low-dimensional macroscopic description of a physical system. In this work, rather than rejecting Chomsky and Halle's constrained relation between the physical and cognitive, as the phonetic implementation approaches have done, we have, if anything, increased the hypothesized tightness of that relation by using the concept of different dimensionality. We have surmised that the problem with the program proposed by Chomsky and Halle was instead in their choice of the elementary units of the system. In particular, we have argued that it is wrong to assume that the elementary units are (1) static, (2) neutral between articulation and acoustics, and (3) arranged in nonoverlapping chunks. Assumptions (1) and (3) have been argued against by Fowler et al. (1980), and (3) has also been rejected by most of the work in "nonlinear" phonology over the past 15 years. Assumption (2) has been, at least partially, rejected in the "active articulator" version of "feature geometry" (Halle, 1982; Sagey, 1986; McCarthy, 1988.)

7.3 GESTURES

Articulatory phonology takes seriously the view that the units of speech production are actions, and therefore that (1) they are dynamic, not static. Further, since articulatory phonology considers phonological functions such as contrast to be low-dimensional, macroscopic descriptions of such actions, the basic units are (2) not neutral between articulation and acoustics, but rather are articulatory in nature. Thus, in articulatory phonology, the basic phonological unit is the *articulatory gesture*, which is defined as a dynamical system specified with a characteristic set of parameter values (see Saltzman, chapter 6). Finally, because the actions are distributed across the various articulator sets of the vocal tract (the lips, tongue, glottis, velum, etc.), an utterance is modeled as an ensemble, or constellation, of a small number of (3) potentially overlapping gestural units.

As is elaborated below, contrast among utterances can be defined in terms of these gestural constellations. Thus, these structures can capture the low-dimensional properties of utterances. In addition, because each gesture is defined as a dynamical system, no rules of implementation are required to characterize the high-dimensional properties of the utterance. A time-varying pattern of articulator motion (and its resulting acoustic consequences) is lawfully entailed by the dynamical systems themselves—they are self-implementing. Moreover, these time-varying patterns automatically display the property of context dependence (which is ubiquitous in the high-dimensional description of speech) even though the gestures are defined in a context-

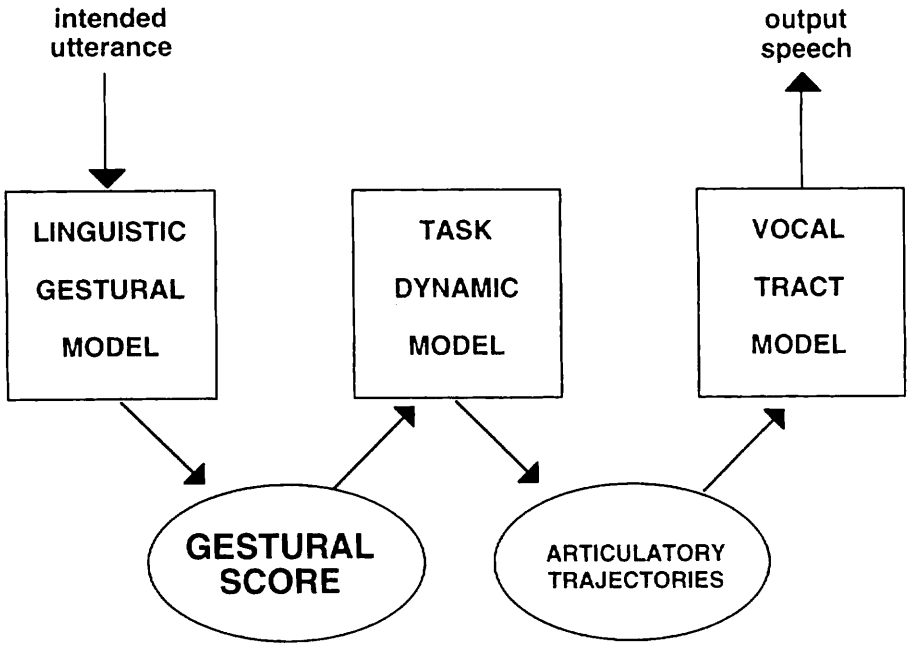


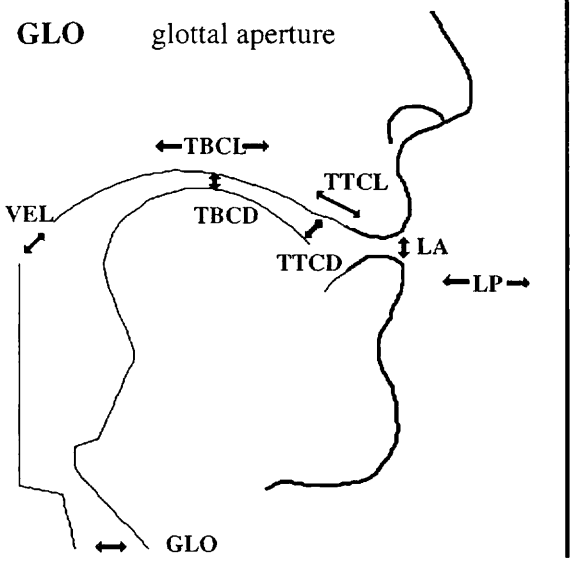
Figure 7.1 Computational system for generating speech using dynamically defined articulatory gestures.

independent fashion. The nature of the articulatory dimensions along which the individual dynamical units are defined allows this context dependence to emerge lawfully.

The articulatory phonology approach has been incorporated into a computational system being developed at Haskins Laboratories (Browman, Goldstein, Kelso, et al., 1984; Saltzman, 1986; Saltzman and Munhall, 1989; Browman and Goldstein, 1990a,c). In this system, illustrated in figure 7.1, utterances are organized ensembles (or *constellations*) of units of articulatory action called *gestures*. Each gesture is modeled as a dynamical system that characterizes the formation (and release) of a local constriction within the vocal tract (the gesture's functional goal or "task"). For example, the word "ban" begins with a gesture whose task is lip closure. The formation of this constriction entails a change in the distance between the upper and lower lips (or *lip aperture*) over time. This change is modeled using a second-order system (a "point attractor," Abraham and Shaw, 1982), specified with particular values for the equilibrium position and stiffness parameters. (Damping is, for the most part, assumed to be critical, so that the system approaches its equilibrium position and doesn't overshoot it.) During the activation interval for this gesture, the equilibrium position for lip aperture is set to the goal value for lip closure; the stiffness setting, combined with the damping, determines the amount of time it will take for the system to get close to the goal of lip closure.

The set of task or *tract* variables currently implemented in the computational model are listed at the top left of figure 7.2, and the sagittal vocal tract shape below illustrates their geometric definitions. This set of tract variables

	tract variable	articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis



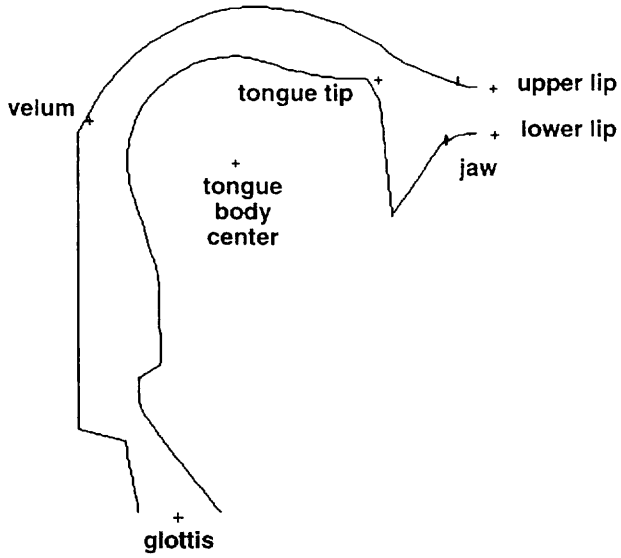


Figure 7.2 Tract variables and their associated articulators.

is hypothesized to be sufficient for characterizing most of the gestures of English (exceptions involve the details of characteristic shaping of constrictions; see Browman and Goldstein, 1989). For oral gestures, two paired tract-variable regimes are specified, one controlling the constriction degree of a particular structure, the other its constriction location (a tract-variable regime consists of a set of values for the dynamic parameters of stiffness, equilibrium position, and damping ratio). Thus, the specification for an oral gesture includes an equilibrium position, or goal, for each of two tract variables, as well as a stiffness (which is currently yoked across the two tract variables). Each functional goal for a gesture is achieved by the coordinated action of a set of articulators, i.e., a coordinative structure (Turvey, 1977; Fowler et al., 1980; Kelso et al., 1986; Saltzman, 1986); the sets of articulators used for each of the tract variables are shown on the top right of figure 7.2, with the articulators indicated on the outline of the vocal tract model below. Note that the same articulators are shared by both of the paired oral tract variables, so that altogether there are five distinct articulator sets, or coordinative structure types, in the system.

In the computational system the articulators are those of a vocal tract model (Rubin, Baer, and Mermelstein, 1981) that can generate speech waveforms from a specification of the positions of individual articulators. When a dynamical system (or pair of them) corresponding to a particular gesture is imposed on the vocal tract, the task-dynamic model (Saltzman, 1986; Saltzman and Kelso, 1987; Saltzman and Munhall, 1989; Saltzman, chapter 6) calculates the time-varying trajectories of the individual articulators constituting that coordinative structure, based on the information about values of the dynamical parameters, and phasing information (see section 7.4), contained in its input. These articulator trajectories are input to the vocal tract model, which then calculates the resulting global vocal tract shape, area function, transfer function, and speech waveform (see figure 7.1).

Defining gestures dynamically can provide a principled link between macroscopic and microscopic properties of speech. To illustrate some of the ways in which this is true, consider the example of lip closure. The values of the dynamical parameters associated with a lip closure gesture are macroscopic properties that define it as a phonological unit and allow it to contrast with other gestures such as the narrowing gesture for [w]. These values are definitional, and remain invariant as long as the gesture is active. At the same time, however, the gesture intrinsically specifies the (microscopic) patterns of continuous change that the lips can exhibit over time. These changes emerge as the lawful consequences of the dynamical system, its parameters, and the initial conditions. Thus, dynamically defined gestures provide a lawful link between macroscopic and microscopic properties.

While tract-variable goals are specified numerically, and in principle could take on any real value, the actual values used to specify the gestures of English in the model cluster in narrow ranges that correspond to contrastive categories: for example, in the case of constriction degree, different ranges are found for gestures that correspond to what are usually referred to as stops, fricatives, and approximants. Thus, paradigmatic comparison (or a density distribution) of the numerical specifications of all English gestures would reveal a macroscopic structure of contrastive categories. The existence of such narrow ranges is predicted by approaches such as the quantal theory (e.g., Stevens, 1989) and the theory of adaptive dispersion (e.g., Lindblom et al., 1983), although the dimensions investigated in those approaches are not identical to the tract-variable dimensions. These approaches can be seen as accounting for how microscopic continua are partitioned into a small number of macroscopic categories.

The physical properties of a given phonological unit vary considerably depending on its context (e.g., Öhman, 1966; Liberman, Cooper, Shankweiler, et al., 1967; Kent and Minifie, 1977). Much of this context dependence emerges lawfully from the use of task dynamics. An example of this kind of context dependence in lip closure gestures can be seen in the fact that the three independent articulators that can contribute to closing the lips (upper lip, lower lip, and jaw) do so to different extents as a function of the vowel

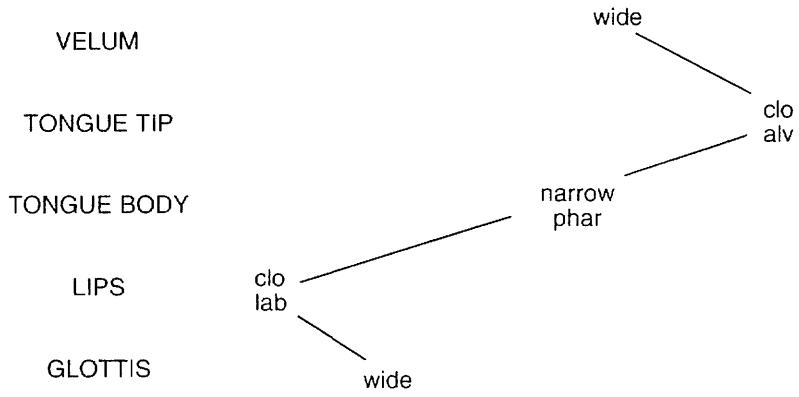
environment in which the lip closure is produced (Sussman, MacNeilage, and Hanson, 1973; Macchi, 1988). The value of lip aperture achieved, however, remains relatively invariant no matter what the vowel context. In the task-dynamic model, the articulator variation results automatically from the fact that the lip closure gesture is modeled as a coordinative structure that links the movements of the three articulators in achieving the lip closure task. The gesture is specified invariantly in terms of the tract variable of lip aperture, but the closing action is distributed across component articulators in a context-dependent way. For example, in an utterance like [ibi], the lip closure is produced concurrently with the tongue gesture for a high front vowel. This vowel gesture will tend to raise the jaw, and thus less activity of the upper and lower lips will be required to effect the lip closure goal than in an utterance like [aba]. These microscopic variations emerge lawfully from the task-dynamic specification of the gestures, combined with the fact of overlap (Kelso, Saltzman, and Tuller, 1986; Saltzman and Munhall, 1989).

7.4 GESTURAL STRUCTURES

During the act of talking, more than one gesture is activated, sometimes sequentially and sometimes in an overlapping fashion. Recurrent patterns of gestures are considered to be organized into gestural constellations. In the computational model (see figure 7.1), the linguistic gestural model determines the relevant constellations for any arbitrary input utterance, including the *phasing* of the gestures. That is, a constellation of gestures is a set of gestures that are coordinated with one another by means of phasing, where for this purpose (and this purpose only), the dynamical regime for each gesture is treated as if it were a cycle of an undamped system with the same stiffness as the actual regime. In this way, any characteristic point in the motion of the system can be identified with a phase of this virtual cycle. For example, the movement onset of a gesture is at phase 0 degrees, while the achievement of the constriction goal (the point at which the critically damped system gets sufficiently close to the equilibrium position) occurs at phase 240 degrees. Pairs of gestures are coordinated by specifying the phases of the two gestures that are synchronous. For example, two gestures could be phased so that their movement onsets are synchronous (0 degrees phased to 0 degrees), or so that the movement onset of one is phased to the goal achievement of another (0 degrees phased to 240 degrees), etc. Generalizations that characterize some phase relations in the gestural constellations of English words are proposed in Browman and Goldstein (1990c). As is the case for the values of the dynamical parameters, values of the synchronized phases also appear to cluster in narrow ranges, with onset of movement (0 degrees) and achievement of goal (240 degrees) being the most common (Browman and Goldstein, 1990a).

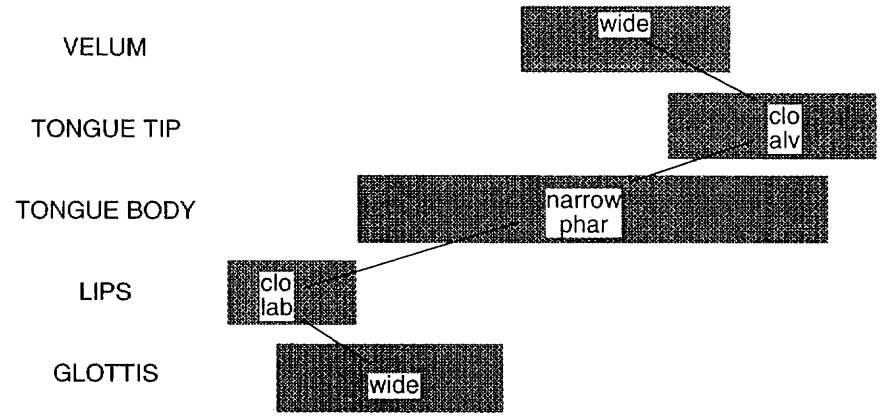
An example of a gestural constellation (for the word "pawn" as pronounced with the back unrounded vowel characteristic of much of the United States) is

ˈpɑn



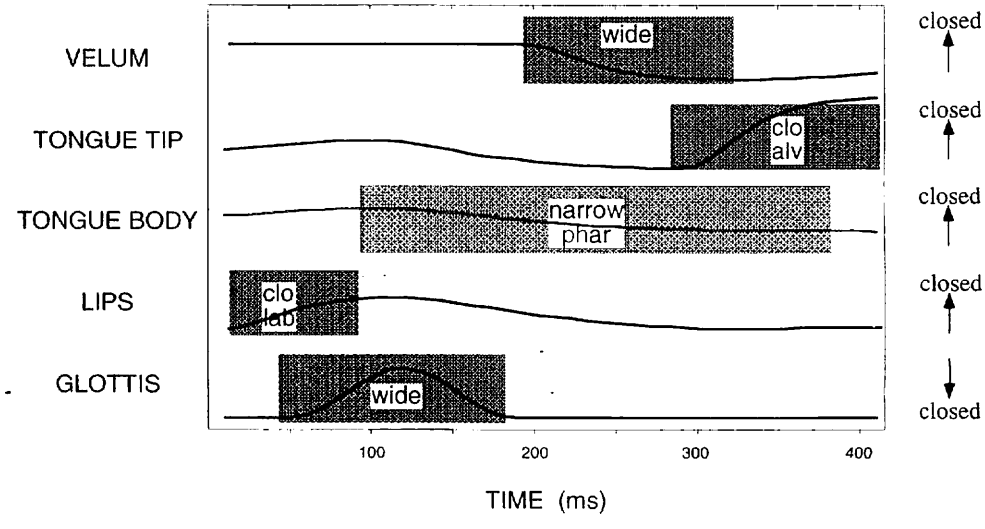
(a)

ˈpɑn



(b)

ˈpɑn



shown in figure 7.3a, which gives an idea of the kind of information contained in the gestural dictionary. Each row, or tier, shows the gestures that control the distinct articulator sets: velum, tongue tip, tongue body, lips, and glottis. The gestures are represented here by descriptors, each of which stands for a numerical equilibrium position value assigned to a tract variable. In the case of the oral gestures, there are two descriptors, one for each of the paired tract variables. For example, for the tongue tip gesture labeled {clo alv}, {clo} stands for −3.5 mm (the negative value indicates compression of the surfaces), and {alv} stands for 56 degrees (where 90 degrees is vertical and would correspond to a midpalatal constriction). The association lines connect gestures that are phased with respect to one another. For example, the tongue tip {clo alv} gesture and the velum {wide} gesture (for nasalization) are phased such that the point indicating 0 degrees—onset of movement—of the tongue tip closure gesture is synchronized with the point indicating 240 degrees—achievement of goal—of the velic gesture.

Each gesture is assumed to be active for a fixed proportion of its virtual cycle (the proportion is different for consonant and vowel gestures). The linguistic gestural model uses this proportion, along with the stiffness of each gesture and the phase relations among the gestures, to calculate a *gestural score* that specifies the temporal activation intervals for each gesture in an utterance. One form of this gestural score for "pawn" is shown in figure 7.3b, with the horizontal extent of each box indicating its activation interval, and the lines between boxes indicating which gesture is phased with respect to which other gesture(s), as before. Note that there is substantial overlap among the gestures. This kind of overlap can result in certain types of context dependence in the articulatory trajectories of the invariantly specified gestures. In addition, overlap can cause the kinds of acoustic variation that have been traditionally described as allophonic variation. For example, in this case, note the substantial overlap between the velic lowering gesture (velum {wide}) and the gesture for the vowel (tongue body {narrow pharyngeal}). This will result in an interval of time during which the velopharyngeal port is open and the vocal tract is in position for the vowel, i.e., a nasalized vowel. Traditionally, the fact of nasalization has been represented by a rule that changes an oral vowel into a nasalized one before a (final) nasal consonant. But viewed in terms of gestural constellations, this nasalization is just the lawful consequence of how the individual gestures are coordinated. The vowel gesture itself has not changed in any way: it has the same specification in this word and in the word "pawed" (which is not nasalized).

Figure 7.3 Various displays from the computational model for "pawn." (*a*) Gestural descriptors and association lines. (*b*) Gestural descriptors and association lines plus activation boxes. (*c*) Gestural descriptors and activation boxes plus generated movements of (*from top to bottom*): velic aperture; vertical position of the tongue tip (with respect to the fixed palate and teeth); vertical position of the tongue body (with respect to the fixed palate and teeth); lip aperture; glottal aperture.

The parameter value specifications and activation intervals from the gestural score are input to the task-dynamical model (see figure 7.1), which calculates the time-varying response of the tract variables and component articulators to the imposition of the dynamical regimes defined by the gestural score. Some of the time-varying responses are shown in figure 7.3c, along with the same boxes indicating the activation intervals for the gestures. Note that the movement curves change over time even when a tract variable is not under the active control of some gesture. Such motion can be seen, for example, in the LIPS panel, after the end of the box for the lip closure gesture. This motion results from one or both of two sources. (1) When an articulator is not part of *any* active gesture, the articulator returns to a neutral position. In the example, the upper lip and lower lip articulators both are returning to a neutral position after the end of the lip closure gesture. (2) One of the articulators linked to the inactive tract variable may also be linked to some active tract variable, and thus cause passive changes in the inactive tract variable. In the example, the jaw is part of the coordinative structure for the tongue-body vowel gesture, as well as part of the coordinative structure for the lip closure gesture. Therefore, even after the lip closure gesture becomes inactive, the jaw is affected by the vowel gesture, and its lowering for the vowel causes the lower lip to also passively lower.

The gestural constellations not only characterize the microscopic properties of the utterances, as discussed above, but systematic differences among the constellations also define the macroscopic property of phonological contrast in a language. Given the nature of gestural constellations, the possible ways in which they may differ from one another is, in fact, quite constrained. In other papers (e.g., Browman and Goldstein, 1986, 1989, 1992) we have begun to show that gestural structures are suitable for characterizing phonological functions such as contrast, and what the relation is between the view of phonological structure implicit in gestural constellations, and that found in other contemporary views of phonology (see also Clements, 1992, for a discussion of these relations). Here we simply give some examples of how the notion of contrast is defined in a system based on gestures, using the schematic gestural scores in figure 7.4.

One way in which constellations may differ is in the presence vs. absence of a gesture. This kind of difference is illustrated by two pairs of subfigures in figure 7.4: (*a*) vs. (*b*) and (*b*) vs. (*d*); (*a*) "pan" differs from (*b*) "ban" in having a glottis {wide} gesture (for voicelessness), while (*b*) "ban" differs from (*d*) "Ann" in having a labial closure gesture (for the initial consonant). Constellations may also differ in the particular tract-variable or articulator set controlled by a gesture within the constellation, as illustrated by (*a*) "pan" vs. (*c*) "tan," which differ in terms of whether it is the lips or tongue tip that performs the initial closure. A further way in which constellations may differ is illustrated by comparing (*e*) "sad" with (*f*) "shad," in which the value of the constriction location tract variable for the initial tongue-tip constriction is the only difference between the two utterances. Finally, two constellations may

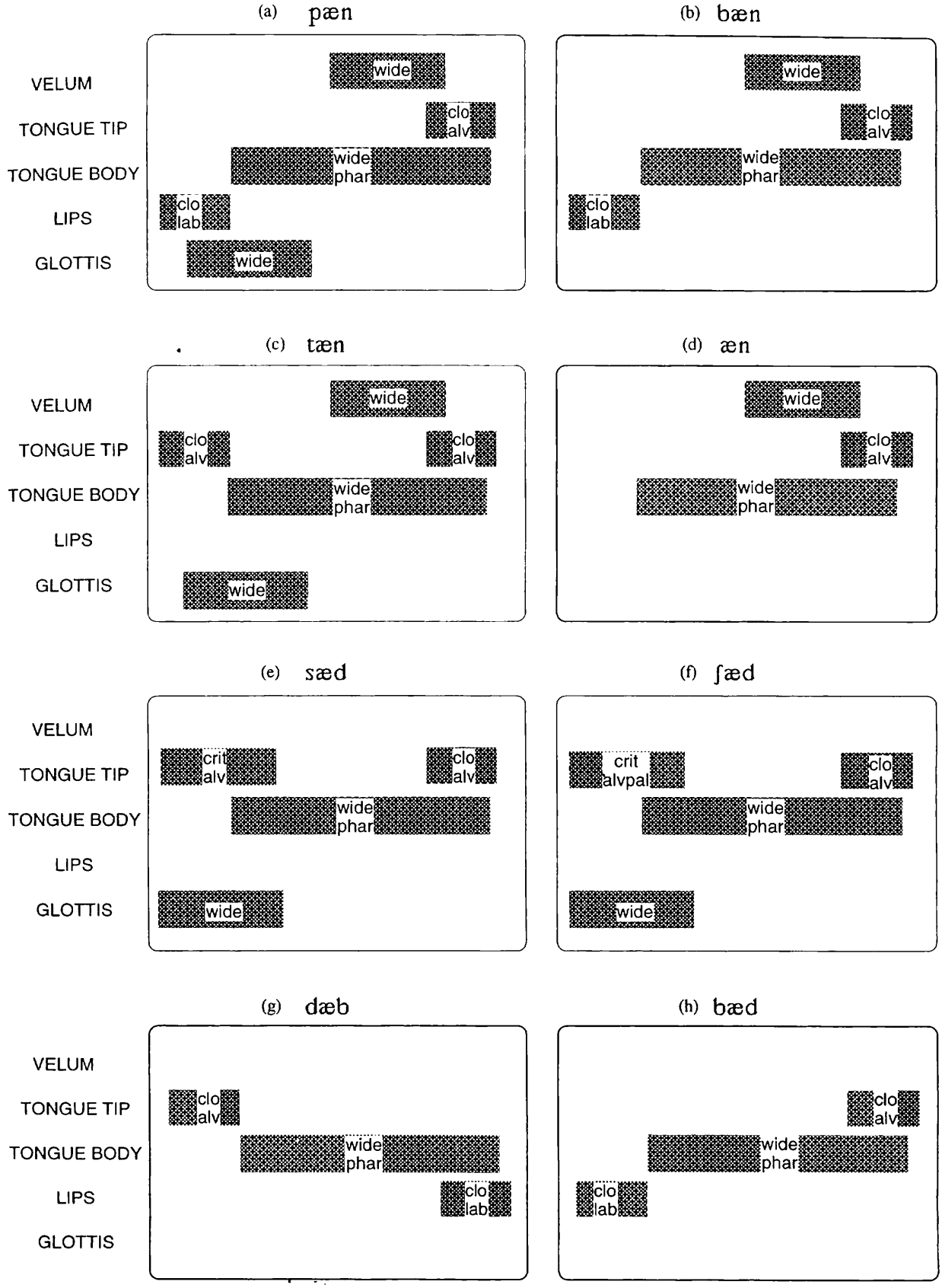


Figure 7.4 Schematic gestural scores exemplifying contrast. (*a*) "pan"; (*b*) "ban"; (*c*) "tan"; (*d*) "Ann"; (*e*) "sad"; (*f*) "shad"; (*g*) "dab"; (*h*) "bad".

contain the same gestures and differ simply in how they are coordinated, as can be seen in (*g*) "dab" vs. (*h*) "bad."

7.5 CONCLUSIONS

This chapter describes an approach to the description of speech in which both the cognitive and physical aspects of speech are captured by viewing speech as a set of actions, or dynamical tasks, that can be described using different dimensionalities: low-dimensional or macroscopic for the cognitive, and high-dimensional or microscopic for the physical. A computational model that instantiates this approach to speech was briefly outlined. It was argued that this approach to speech, which is based on dynamical description, has several advantages over other approaches. First, it captures both the phonological (cognitive) and physical regularities that must be captured in any description of speech. Second, it does so in a way that unifies the two descriptions as descriptions of different dimensionality of a single complex system. The latter attribute means that this approach provides a principled view of the reciprocal constraints that the physical and phonological aspects of speech exhibit.

ACKNOWLEDGMENTS

This work was supported by NSF grant DBS-9112198 and NIH grants HD-01994 and DC-00121 to Haskins Laboratories. We thank Alice Faber and Jeff Shaw for comments on an earlier version.

REFERENCES

Abraham, R. H., and Shaw, C. D. (1982). *Dynamics—the geometry of behavior*. Santa Cruz, CA: Aerial Press.

Anderson, S. R. (1974). *The organization of phonology*. New York: Academic Press.

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

Browman, C. P., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201–251.

Browman, C. P., and Goldstein, L. (1990a). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299–320.

Browman, C. P., and Goldstein, L. (1990b). Representation and reality: physical systems and phonological structure. *Journal of Phonetics, 18*, 411–424.

Browman, C. P., and Goldstein, L. (1990c). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. E. Beckman (Eds.), *Papers in laboratory phonology I: between the grammar and physics of speech* (pp. 341–376). Cambridge: Cambridge University Press.

Browman, C. P., and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica, 49*, 155–180.

Browman, C. P., Goldstein, L., Kelso, J. A. S., et al. (1984). Articulatory synthesis from underlying dynamics (abstract). *Journal of the Acoustical Society of America, 75,* S22–S23.

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

Clements, G. N. (1992). Phonological primes: features or gestures?, *Phonetica, 49,* 181–193.

Cohn, A. C. (1990). Phonetic and phonological rules of nasalization. *UCLA Working Papers in Phonetics, 76.*

Coleman, J. (1992). The phonetic interpretation of headed phonological structures containing overlapping constituents. *Phonology, 9,* 1–44.

Fourakis, M., and Port, R. (1986). Stop epenthesis in English. *Journal of Phonetics, 14,* 197–221.

Fowler, C. A., Rubin, P., Remez, R. E., et al. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language Production.* New York: Academic Press.

Haken, H. (1977). *Synergetics: an introduction.* Heidelberg: Springer Verlag.

Halle, M. (1982). On distinctive features and their articulatory implementation., *Natural Language and Linguistic Theory, 1,* 91–105.

Hockett, C. (1955). *A manual of phonology.* Chicago: University of Chicago.

Hogeweg, P. (1989). MIRROR beyond MIRROR, puddles of LIFE. In C. Langton (Ed.), *Artificial life* (pp. 297–316). New York: Addison-Wesley.

Jakobson, R., Fant, C. G. M., and Halle, M. (1951). *Preliminaries to speech analysis: the distinctive features and their correlates.* Cambridge, MA: MIT Press.

Kauffman, S. (1989). Principles of adaptation in complex systems. In D. Stein (Ed.), *Sciences of complexity* (pp. 619–711). New York: Addison-Wesley.

Kauffman, S. (1991). Antichaos and adaptation. *Scientific American, 265,* 78–84.

Kauffman, S., and Johnsen, S. (1991). Co-evolution to the edge of chaos: coupled fitness landscapes, poised states, and co-evolutionary avalanches. In C. Langton, C. Taylor, J. D. Farmer, et al. (Eds.), *Artificial life II* (pp. 325–369). New York: Addison-Wesley.

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

Keating, P. A. (1988). Underspecification in phonetics. *Phonology, 5,* 275–292.

Keating, P. A. (1990). Phonetic representations in a generative grammar. *Journal of Phonetics, 18,* 321–334.

Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics, 14,* 29–59.

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

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208–1221.

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

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56,* 485–502.

Ladefoged, P. (1982). *A course in phonetics,* 2nd ed. New York: Harcourt Brace Jovanovich.

Lewin, R. (1992). *Complexity.* New York: Macmillan.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., et al. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Liberman, M., and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, R. T. Oehrle, F. Kelley, et al., (Eds.), *Language sound structure* (pp. 157–233). Cambridge, MA: MIT Press.

Lindblom, B., MacNeilage, P., and Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, and O. Dahl (Eds.), *Explanations of linguistic universals* (pp. 181–203). The Hague: Mouton.

Macchi, M. (1988). Labial articulation patterns associated with segmental features and syllable structure in English. *Phonetica, 45,* 109–121.

Madore, B. F., and Freedman, W. L. (1987). Self-organizing structures. *American Scientist, 75,* 252–259.

Manuel, S. Y., and Krakow, R. A. (1984). Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Report on Speech Research, 77/78,* 69–78.

McCarthy, J. J. (1988). Feature geometry and dependency: a review. *Phonetica, 45,* 84–108.

Ohala, J. (1983). The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 189–216). New York: Springer Verlag.

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

Packard, N. (1989). Intrinsic adaptation in a simple model for evolution. In C. Langton (Ed.), *Artificial life* (pp. 141–155). New York: Addison-Wesley.

Pierrehumbert, J. (1990). Phonological and phonetic representation. *Journal of Phonetics, 18,* 375–394.

Pierrehumbert, J. B., and Pierrehumbert, R. T. (1990). On attributing grammars to dynamical systems. *Journal of Phonetics, 18,* 465–477.

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America, 69,* 262–274.

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

Sagey, E. C. (1986). *The representation of features and relations in non-linear phonology.* Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer and C. Fromm (Eds.), *Experimental brain research,* Series 15 (pp. 129–144). New York: Springer-Verlag.

Saltzman, E., and Kelso, J. A. S. (1987). Skilled actions: a task dynamic approach. *Psychological Review, 94,* 84–106.

Saltzman, E. L., and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333–382.

Schöner, G., and Kelso, J. A. S. (1988). Dynamic pattern generation in behavioral and neural systems. *Science, 239,* 1513–1520.

Stevens, K. N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In E. E. David and P. B. Denes (Eds.), *Human communication: a unified view* (pp. 51–66). New York: McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3–45.

Sussman, H. M., MacNeilage, P. F., and 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.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw and J. Bransford (Eds.), *Perceiving, acting and knowing: toward an ecological psychology.* Hillsdale, NJ: Erlbaum.

Wood, S. (1982). X-ray and model studies of vowel articulation (Vol. 23). Working papers, Lund University, Lund, Sweden.