13 Theories and Models of Speech Production

ANDERS LÖFQVIST

"The purpose of models is not to fit the data but to sharpen the questions"
Samuel Karlin
(11th R.A. Fisher Memorial Lecture, Royal Society, 20 April 1983)

1 The speech signal and its description

For the purpose of this chapter, it is convenient to view speech as audible gestures. A speaker creates variations in air pressure and air flow in the vocal tract by making valving actions with different parts of the vocal tract: the glottis, the velum, the tongue, the lips, and the jaw. The changes in pressure and flow give rise to the acoustic signal that we hear when perceiving speech. Most of the variations in the acoustic signal are made intentionally by the speaker to convey linguistic information. Other properties convey what is called paralinguistic information, such as attitudes and emotions, social and geographical dialect characteristics. In addition, there are properties reflecting biological features of the speaker such as sex and age. The resulting acoustic signal is thus shaped by contributions from many different sources that are all overlaid on each other. The fact that listeners can usually identify these different sources suggests that they are recoverable from the acoustic signal.

In describing speech and language, it is common to use one of two modes that can be referred to as the linguistic and the dynamic mode (see Pattee, 1977, for a further elaboration of this distinction). In the linguistic mode, the units of language are described without a temporal domain. For example, most phonological descriptions use a set of symbols that can be arranged in different ways to produce different messages. Although the primitives used for this type of analysis vary depending on the theoretical framework being adopted, the units are commonly described as being discrete and serially ordered. The dynamic mode is used for describing articulatory and acoustic properties of speech. Here, the focus is on the time-varying properties of articulatory movements and/or the spectral characteristics of the speech signal.

This necessarily implies a temporal domain. The linguistic units of speech can no longer be described as discrete, since a salient feature of speech production is that the units show considerable articulatory influence and overlap. This is commonly referred to as coarticulation, coproduction, blending or aggregation (cf. Farnetani, COARTICULATION AND CONNECTED SPEECH PROCESSES; Perkell, ARTICULATORY PROCESSES). Thus, the movements associated with different production units blend seamlessly with each other and in the articulatory record there are no boundaries between units. Consequently, the movements necessary for the production of a given unit differ according to its context, and likewise its acoustic properties vary according to context. A further result of this overlap is that at any one point in time, the vocal tract is an aggregate of different production units (cf. Fowler and Smith, 1986; Saltzman and Munhall, 1989; Löfqvist, 1990). The obvious acoustic consequence is that a single temporal slice of the signal contains influences from several production units (see Fant, 1962, for an early discussion).

Throughout the history of the study of speech, much effort has been devoted to arguments about these two modes of description (cf. Ohala, THE RELATION BETWEEN PHONETICS AND PHONOLOGY). One famous depiction of their different natures is provided by Hockett (1955), who makes an analogy between speech production and a row of raw Easter eggs on a conveyor belt, being smashed between the two rollers of a wringer. The implication is that the units of speech are distinct and serially ordered (perhaps also invariant and displaying their essential properties) before they are all smeared together in the process of articulation: "The flow of eggs before the wringer represents the impulses from the phoneme source; the *mess* that emerges from the wringer represents the output from the speech transmitter." (Hockett, 1955, p. 210; italics added.) We should note that Hockett does not imply that it is impossible to recover the original eggs that went into the mess. He duly comments that an inspector examining the passing mess could "decide, on the basis of the broken and unbroken yolks, the variously spread-out albumen, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer" (ibid.) and further notes that the inspector represents the hearer.

While Hockett's Easter egg analogy would seem to represent an extreme case, it is not unique. Rather, it represents a class of theories which have been called translation theories, because they view speech production as translating a mental representation into something completely different during the process of articulation (cf. Fowler, Rubin, Remez, and Turvey, 1980; Fowler, 1993). Hockett's view is also understandable from its epistemological context. The discovery of coarticulation was made around the beginning of this century, and Menzerath and de Lacerda (1933) published the first systematic treatise on the subject. Not only did they show large contextual variability for productions of the same sound, but they also showed, as had others before them, that it was impossible to draw boundaries between sounds in the articulatory record (cf. Hardcastle, 1981; Farnetani, COARTICULATION AND CONNECTED SPEECH

PROCESSES, for historical reviews). These findings caused some consternation among speech scientists, since it had often been assumed that the same sound would be articulated in the same way irrespective of its context – an assumption that seems to reappear at certain intervals over time. Hence, the search was on to find the invariant, or essential, properties of the phoneme in production (or acoustics). In a review of theories of speech production, Peter MacNeilage neatly sums up the shift in emphasis that has come to dominate work on speech motor control:

> ... it becomes clear that the more basic problem in speech production theory is not the one considered central to most theorists; namely, why articulators do not always reach the same position for a given phoneme. It is, How do articulators always come as close to reaching the same position as they do? One of the main conclusions of this paper is that the essence of the speech production process is not an inefficient response to invariant central signals, but an elegantly controlled variability of response to the demand for a relatively constant end. (MacNeilage, 1970, p. 184)

Before continuing, we should also note another shift of emphasis in the study of speech motor control. Much work in speech physiology was carried out within a paradigm in which two general issues dominated: chain versus comb models for the serial ordering of articulatory movements, and the role of peripheral feedback in speech production (Lashley, 1951; Kozhevnikov and Chistovich, 1965; Keele, 1968; see Kent, 1976b, for a review of these issues). Briefly, in a chain model, the central motor commands to the articulators to produce a segment were supposed to be triggered by feedback from the periphery upon the completion of the articulatory movements for the previous one. In a comb model, the commands to the articulators for successive segments were assumed to be sent according to a plan or temporal scheme. In practice, one limitation in this approach was a tendency to subsume the question of feedback under the question of serial order, and phrase the alternatives as either a chain model incorporating feedback or a comb model without feedback. Of the two remaining alternatives, one was perhaps automatically ruled out, i.e. a chain model without feedback, but the possibility of a comb model incorporating feedback was not generally explored, in spite of the wealth of physiological studies of sensorimotor mechanisms (e.g. Granit, 1970; Matthews, 1972). In such a model, the role of feedback would not necessarily be limited to the sequencing of movements but rather would be important in the shaping of movements as well. A further limitation was an insistence that signals from peripheral receptors go to higher centers with the resulting problem of apparently inadequate loop time. Another possibility could be that information from the periphery goes to lower levels of the nervous system such as the spinal cord or the brain stem; we will later explore the idea that these levels may play a crucial role in integrating signals from the periphery with signals from higher centers.

While coarticulation has been taken as a fundamental characteristic of speech and the basis for the rapidity with which information can be conveyed (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967), it is also likely to be a fundamental characteristic of most motor activities. Presumably, the reason why it has received so much attention in speech science is that speech can, at one level, be described as a succession of discrete segments, making it possible to study how the "canonical" forms of these segments are altered in the process of articulation. Very similar patterns of contextual variability can be found in typing. In typing, the goal is to produce a sequence of keystrokes, and it may thus be easier to define the targets in typing than in speech production. The movements of the fingers towards the keys show large contextual variability in both space and time (Salthouse, 1986). For example, successive keystrokes are made faster by fingers on alternate hands than by fingers on the same hand. The likely reason is that when alternate hands are used, there is no conflict between the fingers used for the strokes, since the two hands can operate independently. The time needed for a keystroke depends on the context in which a character occurs. The range of these contextual influences in typing appears to be limited to two or three characters. In contrast, coarticulatory influences in speech have been claimed to span up to six segments, but the size of the temporal window for coarticulatory influences remains under debate (cf. Farnetani, COARTICULATION AND CONNECTED SPEECH PROCESSES). The differences between strokes made by the same or alternating hands in typing are similar to coarticulation in speech, where the different parts of the vocal tract can operate relatively independently of each other.

2 Concepts and issues in movement control

During speech, parts of the vocal tract are briefly coupled in a functional manner to produce the acoustic characteristics of speech sounds. For example, the production of the bilabial voiceless stop /p/ requires the following set of actions. The lips are closed by joint activity of the jaw and the lips. The velum is elevated to seal off the entrance into the nasal cavity. The glottis is widened and the longitudinal tension of the vocal folds is often increased to prevent glottal vibrations. These articulatory and laryngeal actions all contribute to a period of silence in the acoustic signal and an increase in oral air pressure associated with the stop consonant. Speech production thus involves control and coordination of different parts of the vocal tract. How this is achieved is not well understood. Speech motor control should properly be seen as an instance of the control of coordinated movements in general. As a preliminary to this discussion, we shall briefly review a line of experiments on speech production that will provide a suitable empirical and experimental background. After this review, the remaining parts of this section discuss a number of issues in the control of movement and their implications for speech motor control.

2.1 What happens when speech movements are perturbed?

Daily activities such as walking and picking up and moving objects often require rapid actions to cope with unexpected events such as stumbling or hitting an object with the hand. One valuable experimental paradigm for understanding movement coordination and control is to introduce unexpected perturbations to motor acts in a systematic manner. In a standard experiment, a subject is attached to a small motor that can be activated during some trials to generate a brief load. The rationale for this research is that the nature and time course of the response to the load may reveal the motor organization and reflex structure of the motor act. This paradigm has been applied to different types of motor behavior in humans such as posture control (e.g. Nashner and McCollum, 1985), hand and finger movements (e.g. Traub, Rothwell, and Marsden, 1980; Rothwell, Traub, and Marsden, 1982; Cole, Gracco, and Abbs, 1984), and respiratory control (Newsom Davis and Sears, 1970). A number of studies have also used this method to study speech motor control (Folkins and Abbs, 1975; Folkins and Zimmermann, 1982; Abbs and Gracco, 1984; Kelso, Tuller, Vatikiotis-Bateson, and Fowler, 1984; Gracco and Abbs, 1985, 1988, 1989; Shaiman, 1989; Shaiman and Abbs, 1987; Kollia, 1994; Munhall, Löfqvist, and Kelso, 1994).

From these speech perturbation studies, some general conclusions can be drawn. First, compensations are rapid. Electromyographic responses can occur 20–30 ms after load onset. The latency is not fixed, however, but depends on when the load was applied with respect to onset of activity in the muscles responsible for the movement in question (Abbs, Gracco, and Cole, 1984). The short latencies suggest that the responses are not due to reaction time processes. Second, compensations are mostly task-specific. That is, they are neither stereotypic nor evident throughout the system, but rather tailored to the needs of the ongoing motor act. For example, when the jaw is loaded during the transition from a vowel to a bilabial stop, compensatory responses are made in the upper and lower lips to achieve the labial closure. On the other hand, when the jaw is loaded during the transition from a vowel to a dental fricative or a dental stop, a response is seen in the tongue (Kelso, Tuller, Vatikiotis-Bateson, and Fowler, 1984; Shaiman, 1989). We should add a word of caution here, however, since task specificity is not always consistent across speakers. In particular, one of the subjects in the study by Shaiman (1989) showed increased lower lip movement in addition to jaw and tongue compensatory movements when the jaw was perturbed during the utterance /ædæ/, which does not require lip activity. Similarly, the study by Kelso, Tuller, Vatikiotis-Bateson, and Fowler (1984) found increased upper lip EMG activity in perturbed productions of /bæz/. Third, compensations are flexible and distributed among articulators involved in a specific task. Thus, when the jaw is loaded in the production of a bilabial stop, responses can occur in the jaw

itself and/or in the upper and lower lips (Shaiman, 1989). Fourth, compensations are functional and effective in the sense that the intended goal is normally achieved. For example, Munhall, Löfqvist, and Kelso (1994) perturbed the lower lip at the transition from the first vowel to the medial bilabial voiceless stop in the utterance /i pip/. The system was able to overcome the load, making the intended closure of the vocal tract and increasing the air pressure in the oral cavity: recordings of oral pressure revealed no differences in pressure between load and control productions.

While the results of these studies clearly indicate that the articulatory system is capable of rapid and functional responses to external loads, such loads may, nevertheless, affect the timing between different articulatory systems (Löfqvist and Gracco, 1991; Saltzman, Löfqvist, Kinsella-Shaw, Rubin, and Kay, 1992). For example, Munhall, Löfqvist, and Kelso (1994) also examined laryngeal responses to lower lip perturbations during the production of a voiceless bilabial stop. In addition to lip and jaw actions to achieve the labial closure, a laryngeal response was evident by a delay of the onset of glottal abduction, measured relative to the onset of the preceding vowel. This delay was presumably made to maintain lip-larynx coordination at the onset of labial closure, and resulted in an increased acoustic duration of the preceding vowel. However, the period of bilabial closure for the stop was shortened by the perturbation while the laryngeal abduction–adduction movement increased in duration. The normal phasing between the oral and laryngeal movements was consequently disrupted at the release of the oral closure. As a result, Voice Onset Time increased in the perturbed trials since it depends in part on the timing between the oral and laryngeal events in stop production (e.g. Löfqvist and Yoshioka, 1984; Löfqvist, 1992).

2.2 *Planning and execution of movements*

While it is convenient to discuss movement control in terms of a plan and its execution, there is reason to believe that a clear separation between plan and execution is often not possible. One problem here concerns the representations used in speech planning. Current phonological representations would seem to require a great deal of detail to be filled in during the conversion into a phonetic representation, in particular temporal information (cf. Zsiga, 1993, for a recent discussion). Another issue is how much motoric detail a plan can contain, an issue that will be taken up in more detail in section 2.3. Theories of speech planning have often used cases of speech errors, slips of the tongue or spoonerisms, as evidence (see also Perkell, ARTICULATORY PROCESSES). An example of such an error is when someone says "queer old dean" instead of the intended "dear old queen". Based on analysis of such speech errors, several models of the speech planning process have been proposed (e.g. Garrett, 1980; Levelt, 1989; Dell, Juliano, and Govindjee, 1993). These findings obviously suggest that utterances are planned, since it would otherwise be difficult to

explain how an upcoming word could be exchanged with one that is preceding it. Both words would have to be activated at the same time for such an exchange to occur. Still, the nature of this plan is not clear. Using Hockett's Easter egg analogy, the plan in these models of speech production would seem to correspond to the organization of the eggs before they are smashed between the rollers. That is, the smashing process does not appear to be part of the plan.

2.3 *Distributed control*

The nervous system is made up of a complex network of interacting neurons and centers at different levels of the system. In motor control, one important function must involve integrating signals from higher centers with signals from the periphery which indicate the current state. This involves selecting the appropriate muscles, activating them to a suitable degree, and establishing a proper sequence of activation. The integrative function for limb control is located in the spinal cord (cf. Humphrey and Freund, 1991; McCrea, 1992). Only at this level is all the relevant information present. The activity of the neural pool in the spinal cord is constantly changing as a function of central and peripheral inputs. Hence, a given central command will have different results depending on the current state of the pool. From the perspective of movement planning and execution, the executive and integrative function is thus played by lower levels of the nervous system. Indeed, it is not entirely clear that a general division between central and peripheral processes is possible. A metaphor would be that an intended movement is realized successively in more motoric detail as it is passed down through the system until it reaches the final common path from the motor neurons to the muscles. Speech production would seem to share the same form of control, where the brain stem plays the integrative role (see Kent and Tjaden, BRAIN FUNCTIONS UNDERLYING SPEECH). The rapid and functional compensations following perturbations to articulators are in agreement with such a distributed system.

2.4 *Coordinate spaces*

One persistent problem in movement control concerns the coordinate space in which movements are planned and represented (see Hollerbach, 1990, for a general discussion, and Munhall, Ostry, and Flanagan, 1991, for discussion of speech movements). In unrestrained reaching movements, the hand usually traverses a relatively straight path in an extrinsic cartesian coordinate system. If the same movement is described in an intrinsic coordinate system represented by the joint angles of the shoulder and elbow, a plot of elbow angle versus shoulder angle typically shows a curved path. One can similarly compare articulator path shapes observed during speech production in extrinsic

versus intrinsic coordinates. For example, jaw movements can be represented in extrinsic or intrinsic coordinate space, where the latter involves at least rotation and translation of the jaw, possibly also yaw. For tongue movements, the situation is even more complex. Due to its mechanical linkage to the jaw, movements of the tongue are partly due to jaw rotation and translation, and partly to the activities of intrinsic and extrinsic tongue muscles. Moreover, the tongue has a hydrostatic skeleton, like an elephant's trunk, and not like the joints of the legs, the arms, and the jaw (cf. Smith and Kier, 1989).

Straight-path trajectories in extrinsic space have often been cited as evidence that movements are planned in extrinsic space. For speech, this argument can possibly be bolstered by the fact that the result of the speech production process is a time-varying acoustic signal. It has been argued that speech movements are controlled with respect to such acoustic effects. The acoustic effects depend on the transfer function of the vocal tract. Planning and control of speech movements in an acoustic coordinate system thus seems plausible. We should perhaps add, however, that tongue movements usually do not follow straight lines in extrinsic coordinate space but rather show curved paths, cf. Figure 13.1 (Houde, 1968; Perkell, 1969; Kent and Moll, 1972; Schönle, 1988; Munhall, Ostry, and Flanagan, 1991; Löfqvist, Gracco, and Nye, 1993; Löfqvist and Gracco, 1994; Perkell, ARTICULATORY PROCESSES).

However, one traditional cause for concern is that control in extrinsic space requires the motor control system to solve the so-called inverse problem. A solution to the inverse problem entails going backwards from the desired movement trajectory to the muscle forces required to produce the movements. In arm movement control, the inverse problem involves mapping backwards from the desired movement goal in extrinsic space to the required muscular forces. For speech, the same mappings would be involved, perhaps with the added step of going from acoustic coordinates to vocal tract coordinates. The inverse problem is mathematically ill-posed in the sense that it is unclear whether a solution exists, is unique, and depends continuously on initial conditions (Tikhonov and Arsenin, 1977). One component of the inverse problem for arm movements is that the arm has excess degrees of freedom – seven (e.g. Alexander, 1992). Excess degrees of freedom in this context imply that the number of controlled spatial variables for the arm is less than the number of controlled joint angular variables. In such a case, the mapping from spatial variables to joint variables is indeterminate, since the same final position of the hand can be achieved by very many possible combinations of joint movements and, consequently, of very many different combinations of muscle activity patterns (there are 22 distinct muscles in the arm). In speech, the problem is the one-to-many mapping from acoustic signal to vocal tract area function as well as the excess degrees of freedom of the articulatory system. Models of speech production arguing for acoustically-based targets would assume implicitly that speech movements are planned in acoustic space. Interestingly, when the acoustic properties of the vocal tract are changed experimentally, e.g. by having subjects wear a dental prosthesis, speakers do not compensate

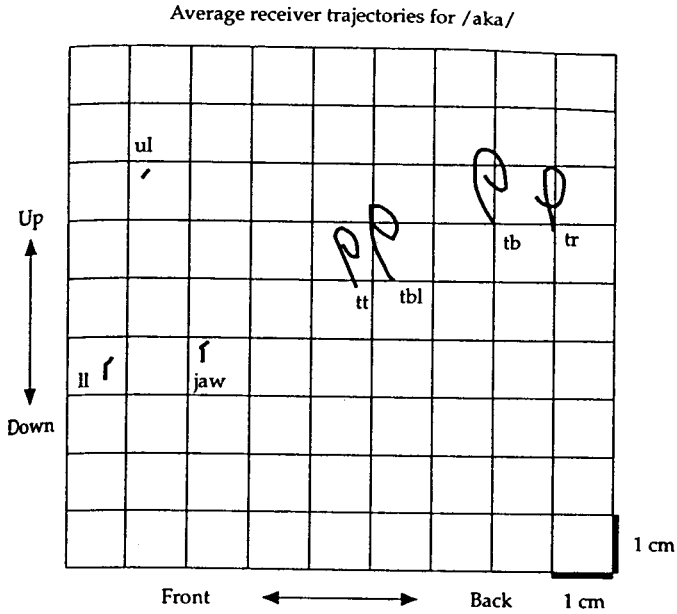

Figure 13.1a Trajectories of receivers placed on the upper and lower lips, the jaw, and on four positions on the tongue during production of the sequence /aka/. All four tongue receivers move counterclockwise. See Löfqvist, Gracco, and Nye (1993), and Löfqvist and Gracco (1994) for experimental details.

immediately for the induced changes (Hamlet and Stone, 1976, 1978). Rather, such a manipulation requires some time for adjustment, possibly indicating that an inverse mapping has to be solved anew. These results might superficially seem to contradict the finding of immediate compensations when jaw movements are constrained by a bite-block held between a subject's teeth (Lindblom, Lubker, and Gay, 1979; Lubker, 1979; Fowler and Turvey, 1980). Note that the bite-block does not necessarily change the transfer function of the vocal tract in the same way as a dental prosthesis.

Recent work using parallel processing (Jordan, 1990; Jordan and Rumelhart, 1992) suggests that the traditional computational concerns about the inverse problem may be exaggerated (see section 2.7). We should also remember that speech and most other skilled movements are highly learned motor activities. Depending on what definition we use for mastering speech, it takes human infants two or three years to acquire it. Thus, in many instances of movement control, the motor system may not have to perform an exhaustive inverse computation, since learning can reduce the number of possible actions. Furthermore, much of the discussion about the inverse problem has received

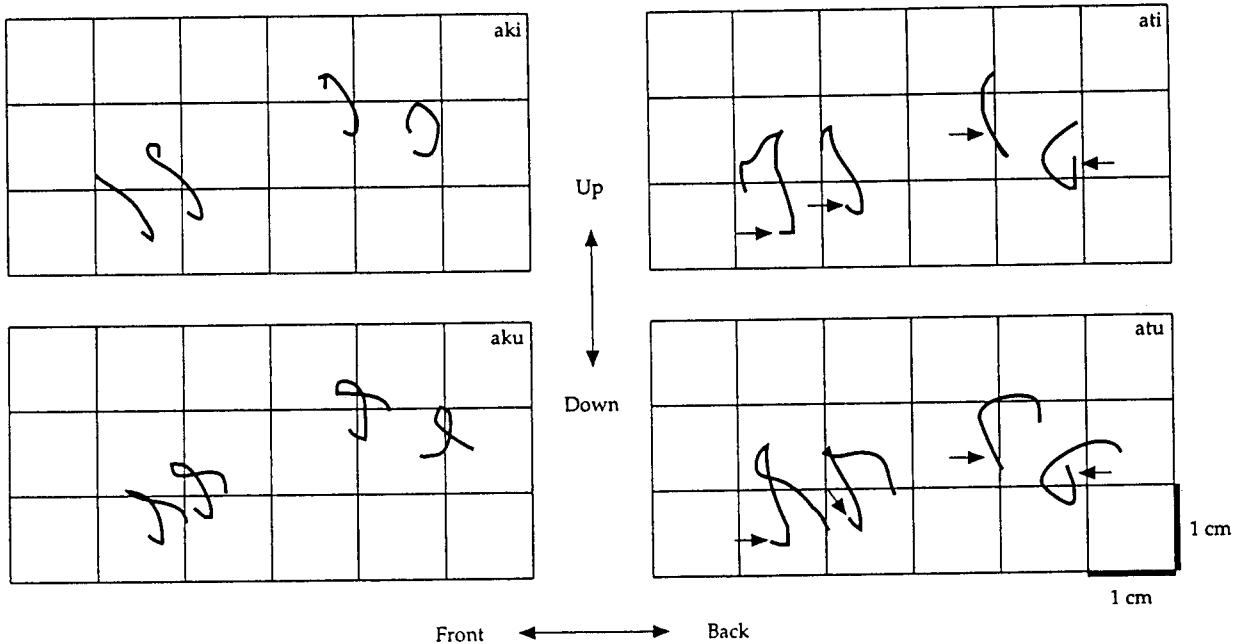

Figure 13.1b Trajectories of four tongue receivers during different VCV sequences. All four receivers move counterclockwise in the sequences with a velar consonant. The arrows identify the onset of the trajectories in the sequences with a dental stop.

input from the field of robotics, but natural systems may "take a loan" on physics and evolution to solve this problem. Brains and nervous systems are not general purpose devices, but rather special purpose devices that have evolved to solve ecologically significant problems in a world governed by relatively stable and predictable physical forces.

In an attempt to alleviate the inverse problem, some investigators have argued that motor control is formulated in terms of muscular coordinates. One such model is the equilibrium-point model (Asatryan and Feldman, 1965; Feldman, 1966; see Bizzi, Hogan, Mussa-Ivaldi, and Giszter, 1992, and commentaries for a review). According to the equilibrium-point model, the target of the movement is specified by the length and stiffness of agonist–antagonist muscle pairs working across a joint. This specification is made via central commands. We noted in section 2.4 that the tongue is a muscular hydrostat lacking joints. Equilibrium-point control of the tongue would nevertheless seem possible. By specifying relationships between the three major extrinsic tongue muscles the tongue can be moved up and down, forward and backward. Control of tongue shape can similarly be made by changing the relation between on the one hand the transverse and vertical muscles, and on the other hand the longitudinal muscles. There is some experimental evidence for such a control model. For example, Bizzi and colleagues (Polit and Bizzi, 1979) studied arm movements in monkeys who had been deprived of sensory information from the arm. The animals could still reach a visually presented target using the arm without kinesthetic or visual feedback about arm position. They could even do so when the arm was momentarily perturbed in the opposite direction, slowing the movement. Initially, it was thought that the target was set once and for all before the initiation of movement. In a later study (Bizzi, Accornero, Chapple, and Hogan, 1984), the perturbation was applied in the opposite direction during the reaching task. That is, the perturbation moved the arm towards the target position and thus assisted the movement. Contrary to expectations, this did not result in the arm reaching the target faster. Instead, after the perturbation had been released, the arm moved away from the target and returned to the position on its trajectory before the perturbation was applied. Hence, the target is apparently not specified at the onset of movement but rather continuously updated.

Using an equilibrium-point approach, the muscles controlling movement in a joint can be modeled as a mass-spring system. This has certain attractive features (cf. Cooke, 1980). One of them is that movement will proceed in the face of transient perturbations (cf. section 2.1). Another one is equifinality, i.e. the intended goal will be reached from different initial conditions (see also Perkell, ARTICULATORY PROCESSES).

2.5 *Coordinative structures*

The speech perturbation studies suggest another property of movement control. In coordinated action, the level of control is not the individual muscles

but rather task-dependent groupings of muscles. For example, perturbations to the jaw during the formation of a labial closure are compensated for by any combination of lip and jaw activity. It thus appears that individual articulators can be flexibly marshaled during speech to perform the intended closure in the vocal tract (cf. Gracco and Abbs, 1986).

Such task-dependent groupings of muscles have been called coordinative structures or synergies. This particular view of movement control owes much of its initial formulation to the Russian physiologist N. Bernstein (Bernstein, 1967). Further discussion and elaboration of these concepts are found in Greene (1972), Gelfand, Gurfinkel, Fomin, and Tsetlin (1971), Turvey (1977, 1990), Kelso, Holt, Kugler, and Turvey (1980), Kugler, Kelso, and Turvey (1980), and Lee (1984). A synergy is defined as "those classes of movements which have similar kinematic characteristics, coinciding active muscle groups and conducting types of afferentation" (Gelfand, Gurfinkel, Tsetlin, and Shik, 1971, p. 331). According to Lee (1984), synergies can be defined by coherent patterns of muscle activity and/or movement, and in terms of spatial, temporal and scaling properties. Spatially, the same set of muscles should be activated. In the temporal domain, synchronicity, stable order or stable phase relationships should hold between events. Relations among events should demonstrate a scaling relationship. Such a definition requires appropriate measurements for a synergy to be recognized. For speech, the arguments for synergies have mostly been based on temporal and spatial relationships between muscle and/or movement patterns. As will be discussed in section 3, there are some intriguing experimental problems in defining synergies using timing and scaling properties. The most convincing evidence for coordinative structures would appear to come from the perturbation studies reviewed in section 2.1. In particular, a theory of coordinative structures predicts task-specific responses.

Coordinative structures should be seen as linkages between muscles that are set up for the execution of specific tasks. For example, Kelso, Tuller, and Harris (1983) had subjects make flexion-extension movements of the index finger in synchrony with stressed and unstressed syllables. They noted a coupling between speech and finger movements. When producing a stressed syllable, the subjects also increased the amplitude of the finger movements. Similarly, when the finger was mechanically perturbed, a change in the acoustic speech signal was also observed. The authors argue that these findings can be accounted for by a coordinative structure comprising the vocal tract and the hand, set up for the execution of a specific task. Thus, when one member of the synergy was perturbed, other members also showed a change. We should note that the functionality of this particular coupled change is not entirely clear. Perhaps we should entertain another interpretation of these results.

Movements such as walking and swimming are rhythmic, and such movements can be effectively modeled by coupled oscillators producing many different patterns of organization (cf. Cohen, Rossignol, and Grillner, 1988; see also Stewart and Golubitsky, 1992, chapter 8, for a discussion of locomotion in terms of coupled oscillators). According to the oscillator model, a perturbation

to a coupled system would manifest itself throughout the system. This class of models is very powerful for simulating coordinated rhythmic movements such as those found in locomotion, swimming, and chewing. The question arises, however, whether such rhythmic patterns are a property of normal speech movements.

One attractive feature of coordinative structures is that they can provide a principled solution to the problem of controlling many degrees of freedom. We noted above that the arm has several degrees of freedom which, on the one hand, provides flexibility in the control of arm movements, but on the other hand introduces the problem of indeterminacy in managing all the degrees. A coordinative structure can be described as a set of constraints between muscles that are set up to make the set of muscles behave as a unit. Thus, control is simplified in the sense that the individual muscles need not be controlled independently of each other but rather as a functional unit. It is obvious, however, that while control may be simplified at one level, complexities arise on other levels. If coordinative structures are task-specific and set up for brief periods of time to execute a given movement, there must be a way for the system to keep track of these different coordinative structures, to put them together and break them apart at the appropriate time. For example, in the production of a sequence of a bilabial stop and a vowel, the jaw and the lips are engaged in making and releasing the labial closure for the consonant, whereas for the vowel, the lips may not be directly involved while the jaw and the tongue shape the vocal tract. Thus, the degree of coupling between the lips, the jaw, and the tongue is changing.

2.6 *A gestural approach to speech production*

Records of speech movements generally show a succession of opening and closing movements at different locations in the vocal tract. One approach to understanding speech motor control is to posit underlying gestures as the building blocks of speech. A gesture can be defined briefly as a class of functionally equivalent movement patterns (cf. Saltzman and Munhall, 1989). Again, a word of caution is in order, since introducing underlying representations always carries a certain risk – such representations have a tendency to show an unprincipled rate of multiplication. Parsimony and a judicious use of Occam's razor is often desirable in science, although we are also well advised to keep in mind that the famous razor has been described as an instrument used by scientists to cut their own throats. Still, using gestures as underlying representations has certain advantages. It can possibly bypass the translation problem by providing the underlying linguistic units with more motoric detail. In this view, a segment should be viewed as a set of gestures (see Löfqvist, 1990, for a defense of the segment).

Munhall and Löfqvist (1992) examined how the two successive laryngeal movements in the utterance "Kiss Ted", for the /s/ and the /t/, were affected

by variations in speaking rate. At a slow rate, two independent movements were found. At fast rates, a single movement was observed. Interestingly, at intermediate rates, a blend of the two gestures was seen. These blends could be reasonably well modeled by adding together two underlying gestures at different degrees of overlap. By varying speaking rate, it was thus possible to view the gestures both in isolation and as aggregates. Hence, the assumed underlying gestures could be readily observed. Similar effects of speaking rate on velar movements have been presented by Boyce, Krakow, Bell-Berti, and Gelfer (1990).

Using underlying gestures to account for movement control is not a new idea. Aiming movements have often been shown to be composed of a number of submovements (e.g. Woodworth, 1899). Here, a large initial movement is followed by smaller corrective movements. It has been suggested by Milner and Ijaz (1990) that irregularities in the tangential velocity of aiming movements can be accounted for by linearly superimposing submovements to create a single composite movement. Similarly, when a subject is suddenly required to switch to a new target after a reaching movement has started, the initial movement is not aborted. Rather, a second movement is blended with the first one (Flash, 1990), and the resulting tangential velocity of the movement can be modeled by adding two underlying movements. For speech, Öhman (1966, 1967) showed evidence of gestural blending in VCV sequences when the vowels and medial consonant shared the same articulator: In the sequences /aga/ and /igi/, the tongue shape during the closure for the /g/ is a blend of the gestures for the vowels and the consonant (cf. Saltzman and Munhall, 1989, for simulation of such patterns using gestural blending). Thus, blending of gestures may be a general strategy that the motor system applies in implementing successive elements of movements.

2.7 *Some new tools: connectionist models*

The nervous system consists of a very large number of individual processing units, neurons, with very many interconnections. Compared to the central processor of a modern computer, each of these processing units is slow. In contrast to most computers, though, the neurons of nervous systems perform their processing in parallel, making up for the slowness of the individual units. Models borrowing the parallel processing (neural network) approach have become more common in the last decade and offer potentially useful and interesting tools for modeling speech processes. (While there is a large and rapidly increasing literature on neural networks, the standard references are still Rumelhart and McClelland (1986) and McClelland and Rumelhart (1986).) Although such models have some similarities to natural nervous systems, they are not brain models – they are modeling tools. A typical neural network consists of one layer of input units, one layer of output units, and one layer of intermediate, or hidden, units. The units are interconnected with each other

and the connections can increase and decrease the activation level of the receiving unit. The activation level of a given unit thus depends on the weighted sum of its inputs. In a pattern recognition task, a task where neural nets have proved especially useful, the object is to make the network respond to a pattern applied to the input units with a certain pattern at the output units. For example, one task could be to have the network produce speech in response to an alphabetic input. During a training period, the network "learns" to associate input and output patterns by adjusting the weights of the connections between the units. After training, the network shows some similarities with natural systems such as generalization of responses to patterns not in the training set and graceful degradation of performance when units are eliminated. (For further description of neural nets, see Ainsworth, SOME APPROACHES TO AUTOMATIC SPEECH RECOGNITION.)

Jordan and Rumelhart (1992) describe the application of neural nets to the inverse problem in movement control (cf. section 2.4). This approach involves two steps. First a forward model of the system under study is learned. A forward model produces the consequence of a given action based on the current state, e.g. what happens with the hand when there is motion at the joints of the arm. The counterpart of a forward model is an inverse model. An inverse model produces an action as a function of the current state and a desired consequence. In the second step, the inverse model and the forward model are composed and an identity mapping is learned across the composed network. When learning the identity mapping, the network also finds an inverse model.

One important factor in learning in both natural and artificial systems is constraints. Jordan (1990) discusses some general constraints on movements that are most likely identical across systems. Such constraints can be taken as minimization of cost functions, defined in terms of, for example, energy, time, and smoothness. Principles of least-effort are intuitively attractive in motor control (cf. Nelson, 1983), and have often been invoked in reference to speech motor control (e.g. Lindblom, 1983). A persistent problem in speech has been to obtain the actual costs. Models may offer some insights in this area.

As a tool in studying processes of speech production, neural nets have been used to model articulatory movements from EMG signals (Hirayama, Vatikiotis-Bateson, Kawato, and Honda, 1992; Vatikiotis-Bateson, Hirayama, Honda, and Kawato, 1992). In this approach, a network is first trained to learn correlations between articulator position, velocity, EMG and articulator acceleration. This produces a forward model. The forward model is then incorporated into a recurrent network that is driven by EMG signals and predicts articulator position and velocity over time as output. Bailly, Laboissière, and Schwartz (1991) describe a similar system for speech synthesis, where a forward model of the articulatory system is incorporated into a control system that drives an articulatory synthesizer.

Guenther (1994) presents a connectionist model of speech production that integrates and provides a unified account for a number of experimental findings

regarding the effects of changes in speaking rate and phonetic context on speech movement kinematics. In this model, the movement targets of different articulators for the production of a given sound are expressed using convex hulls; a convex hull is a region in orosensory space. The size of the hulls serves as an index of the precision needed to produce a given sound. The place of constriction or closure for a consonant is associated with smaller hulls than those for vowels. A change in speaking rate can be modeled by making the hull for a sound larger or smaller; a shrinking hull is associated with a decrease in speaking rate. The change of the size of the hull is larger for vowels than for consonants. Consequently, in the model an increase in speaking rate is associated with an increased movement velocity for vowels but with no change, or even a decrease in velocity, for consonants, as has been found in several experimental studies.

Connectionism is a rapidly expanding field that cuts across many different approaches to modeling (cf. Farmer, 1990). At present, most connectionist models of speech production are limited to statistical patterns of input-output relationships. They are not "complete" models in the sense that they incorporate the biomechanics of the articulators but rather lump everything into a single network.

3 Serial control of speech movements

During normal speech production, movements of the articulators have to be made in the proper sequence to produce an acoustic signal that transmits the intended message. Figure 13.2 shows aerodynamic and articulatory records of three productions of the utterance "It's a papaya", spoken at a conversational rate. The top trace shows the air pressure in the oral cavity; there are three local increases in pressure associated with the voiceless consonants of the utterance. The middle trace shows the vertical movements of the lower lip; the lip moves upwards for the labial closure of the two stop consonants. The bottom trace shows the opening in the glottis; the glottis opens three times for the production of the voiceless consonants; the glottal movements for the stop and the fricative in /ts/ blend together. The signals for the three productions have been temporally aligned at the first peak glottal opening, associated with the cluster /ts/. The duration of the three productions differ. This can be seen by comparing the temporal location of the highest lower lip position for the second /p/ in "papaya" across productions. The order of occurrence of this point for the three productions is (in terms of line thickness) the thin line, the thick line, and the medium line. The difference in utterance duration is visible in all the three signals for that utterance, i.e. the movements of the lower lip and the glottis as well as the increase in oral air pressure all shift together. This is, in a sense, self-evident, since if it didn't occur, the intelligibility of speech would break down. The temporal coordination between articulatory move-

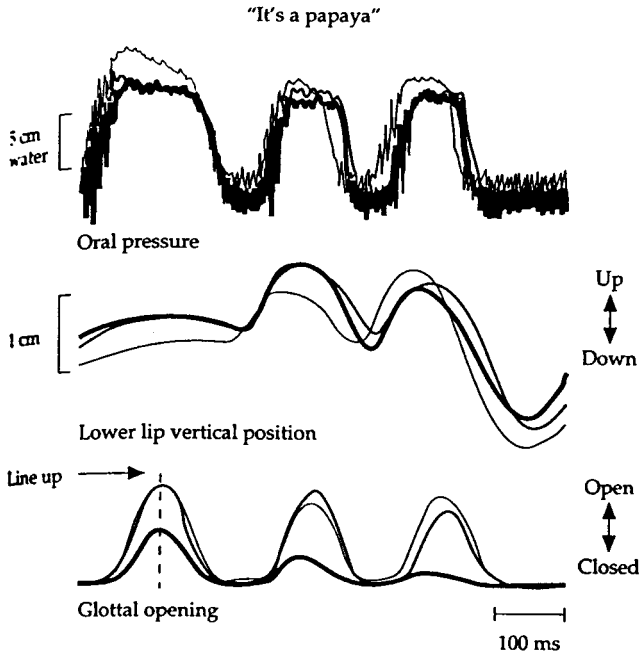

Figure 13.2 Records of three repetitions of the utterance "It's a papaya". The curves represent, from top to bottom, oral air pressure, lower lip vertical position, and glottal opening.

ments has to be maintained within certain limits for speech to be intelligible, across changes in speaking rate. How this temporal cohesion is achieved is not well understood, however. It has been suggested that variations in speaking rate result in a scaling between the different articulatory movements that are involved in the production process. This suggestion is based on the following theoretical view. If someone is writing a word on a paper with a pencil or on a blackboard with a piece of chalk, different parts of the body are used. When the word is written on paper, writing involves movements of the hand and around the wrist; when it is written on the blackboard, the arm moves around the shoulder joint. Since the written pattern on the blackboard can be seen as a scaled version of the one on paper, it has generally been argued that there is a single underlying representation of the movement pattern that is instantiated by different parts of the body using a scaling relation. The alternative view, that each pattern is stored as a separate entity, is at least intuitively implausible and inefficient. Thus, the claim is that the pattern is stored as a "generalized motor program" that can be reparameterized (see Schmidt, 1975).

A generalized motor program predicts that when variations in speed and amplitude of a movement complex occur, the relationship between the individual movements should remain virtually unchanged. The reason is that a submovement interval should maintain a constant proportion of the whole movement interval. Hence, the model is usually referred to as a proportional duration model (see Heuer, 1991, for a general discussion of such models). Initially, several studies claimed that proportional timing was indeed found for motor activities like locomotion, (Shapiro, Zernicke, Gregor, and Diestel, 1981) handwriting (Viviani and Terzuolo, 1980), typing (Terzuolo and Viviani, 1979), and speech (Tuller and Kelso, 1984).

Gentner (1987) proposed a stronger test of proportional duration by examining if the ratio between one movement interval and the duration of the whole movement sequence is unrelated to the duration of the whole movement sequence. The proportional duration model predicts that this should be the case, since the duration of all the components of a movement sequence should maintain a constant proportion of the overall duration. Studies applying this statistical analysis suggest that proportional timing does not occur in speech or any other motor activity that has been examined (cf. Sock, Ollila, Delattre, Zilliox, and Zohair, 1988; Wann and Nimmo-Smith, 1990; Löfqvist, 1991). The slope of the regression usually deviates from zero. One methodological uncertainty facing students of speech timing should be mentioned in this context. Studies of temporal phenomena by necessity have to break up the flow of articulatory movements into discrete intervals for measurement. To delimit these intervals, movement onset and offset, and peak velocity of movement are commonly used. It is, of course, possible that these events are not the ones that the nervous system uses for controlling movements. Kelso, Saltzman, and Tuller (1986) suggested that the proper metric for constant relative timing is phase as measured on a phase plane, rather than ratio of articulatory intervals, and presented some evidence in support of this notion. In a phase plane representation, position is plotted against velocity. In a vowel-labial consonant-vowel sequence, a phase plane plot of the jaw or the lower lip shows an elliptical orbit. Using this kind of representation, movement onsets for different articulators can be defined in terms of phase relationships. Further studies have, however, failed to replicate their findings (Lubker, 1986; Nittrouer, Munhall, Kelso, Tuller, and Harris, 1988; Nittrouer, 1991). These results have implications for theories of speech motor control based on coordinative structures. When discussing coordinative structures in section 2.5, we noted that a definition based on temporal relations requires fixed intervals or scaling among components. One interpretation of scaling is proportional timing which, as we have seen, does not appear to occur, or, at least the scaling is not linear. An important task for speech motor control is to define the metric that governs temporal relations among speech movements.

While constant proportionality thus does not appear to be a proper description of speech movement timing, movements still show temporal cohesion as exemplified by the material presented in Figure 13.2. Another way of analyzing

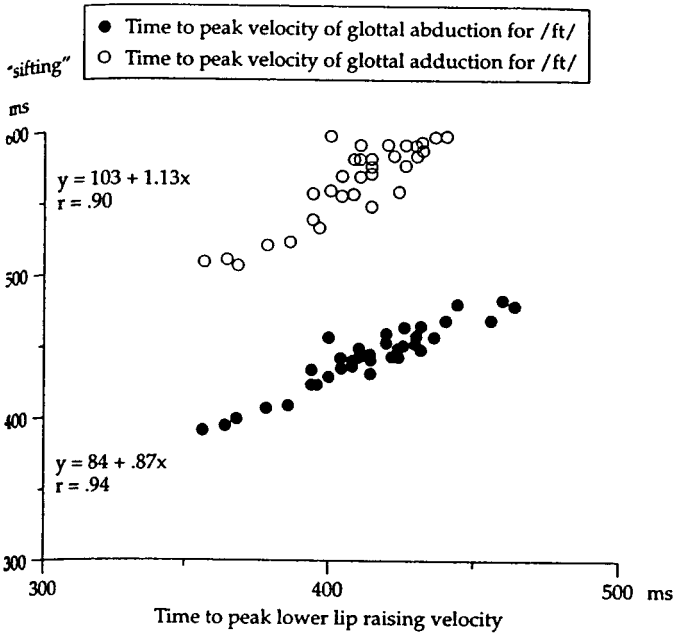

Figure 13.3a Plot of temporal intervals during productions of the word "sifting". The intervals are defined in Figure 13.3b.

the temporal properties between speech movements is illustrated in Figure 13.3. Figure 13.3a shows a plot of articulatory intervals during several normal productions of the utterance "It's a sifting again". Figure 13.3b shows how these intervals have been defined. The interval plotted along the x-axis in Figure 13.3a is measured from the peak glottal opening for the voiceless consonant cluster in "It's" (identified by the vertical dashed line in Figure 13.3b) to the peak velocity of the lower lip raising movement for the labiodental fricative /f/ in "sifting". On the y-axis are plotted two intervals that are both measured from the same instance of peak glottal opening. One of these intervals ends at the peak velocity of the glottal abduction movement for the fricative+stop sequence in "sifting", while the other one ends at the peak adduction velocity of the same laryngeal action. The movements of the lower lip and the glottis are both made to create a period of voiceless frication noise in the acoustic signal. The intervals plotted in Figure 13.3 are thus temporally related to each other in the production of the specific utterance, and one would expect that they should covary. As variations in the overall duration of the utterance occur between productions, the intervals measured for the lower lip

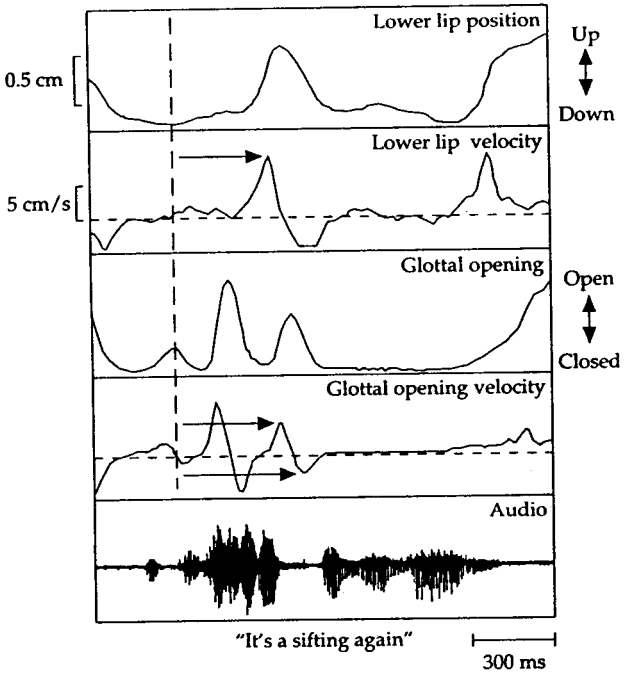

Figure 13.3b A single production of the utterance "It's a sifting again". The signals represent lower lip vertical movement, glottal opening, and audio. The vertical dashed line occurs at peak glottal opening for the voiceless consonant cluster in "It's" and serves as an anchor point for measuring articulatory intervals. The three intervals identified by the horizontal arrows represent (from top to bottom) time to peak velocity of lower lip raising for the /f/ in "sifting", time to peak velocity of glottal abduction for /ft/, and time to peak velocity of glottal adduction for /ft/.

and the larynx should change together; remember that they have been measured from the same temporal reference point, peak glottal opening for the voiceless consonants /ts/. As is evident from Figure 13.3a, this is indeed the case. Their covariation can be indexed by the high correlation between them. At the same time, it is also apparent from Figure 13.3a that they do not scale linearly, since the intercepts of the regressions are not at, or close to, zero. This type of analysis has been used to index temporal cohesion in speech production between the lips and the jaw (see Gracco, 1988, 1994; Gracco and Abbs, 1988) and also between oral and laryngeal movements (e.g. Löfqvist and Yoshioka, 1984; Gracco and Löfqvist, 1994). One possible statistical problem should be mentioned in this context. In using correlations, one has to be aware

of the possibility of correlating intervals that form a part-whole relationship. Such a relationship would in itself result in a correlation coefficient of about .7 (cf. Benoit, 1986; Munhall, 1985).

Speech movements thus show temporal cohesion even though they do not appear to follow a proportional duration model. What are the rules governing this cohesion? Admittedly, not very much is known about this problem, although a reasonable assumption is that intervals that are important for the integrity of the speech signal will show relatively less variability than others. A recent experiment by Saltzman, Löfqvist, Kinsella-Shaw, Kay, and Rubin (1992, 1995) tried to shed some light on this issue using the perturbation paradigm discussed above (see also Gracco and Abbs, 1989). As a subject was producing the pseudo-word "pæsæpæpple", a mechanical load was applied to the lower lip, pulling the lip downwards; the load was applied at different points in time during the production of the utterance. They found that the temporal intervals between the successive bilabial closing movements for the stop consonants were systematically affected by the perturbation. Most of the timing changes occurred during the lip opening phases of these intervals; these phases are associated with the production of the vowels. The closing phases were relatively resistant to temporal distortion, suggesting that their durations were, in some sense, more actively controlled.

In discussions about timing control of speech movements, a confusing issue has been whether timing is intrinsic or extrinsic. According to an extrinsic timing model, time is metered out by a central clock or time keeper that is, in a sense, outside the movement itself. Proponents of intrinsic timing argue that time may not be represented outside the movement but is rather "inside" it (cf. Fowler, 1980; Kelso and Tuller, 1987). It seems safe to conclude that the solution depends on the level of description being adopted. According to the equilibrium-point model of movement control, the end point of the movement is specified in terms of the relationship between the stiffness or activation thresholds of agonist and antagonist muscles. The duration of the movement trajectory thus depends on the dynamical system defined among muscles, and there may be no timing device keeping track of the progression of the movement. In this sense, time is not represented outside the movement by a timekeeper but is rather intrinsic to it. However, for movements to be properly executed and sequenced, the equilibrium points have to be reset continuously. These changes have to be made at the appropriate points in time and the system must have some time-keeping mechanism to make them. At this level, the time keeping should thus more properly be considered extrinsic.

What are the properties of the clock or time keeper? Again, applying mechanical perturbations to movements may provide some clues. For rhythmic movements, phase resetting analysis can be used (see Winfree, 1980, and Glass and Mackey, 1988, for general discussion). In this type of experiment, one measures the temporal shift that is introduced by a perturbation relative to the timing pattern of the pre-perturbation rhythm. If a phase shift is found, the implication is that a central clock does not drive the periphery in a unidirectional

manner. Rather, the central–peripheral coupling is bi-directional, since feedback from the periphery affects the clock. Studies of rhythmic finger movements (Kay, Saltzman, and Kelso, 1991) suggest that mechanical perturbations do introduce shifts in the phasing of such movements. Results reported by Saltzman (1992) and by Saltzman, Löfqvist, Kay, Rubin, and Kinsella-Shaw (1992, forthcoming) indicate that this is also the case for speech, at least when the speech task consists of the repetition of a single consonant-vowel syllable.

4 Summary

The theoretical and empirical approaches to speech production that we have discussed in this chapter converge in their focus on understanding how the different parts of the vocal tract are flexibly marshaled and coordinated to produce the acoustic signal that the speaker uses to convey a message. A variety of experimental paradigms are currently being applied to the problem of coordination and control in motor systems with excess degrees of freedom. Progress in speech motor control is likely to benefit from input from other areas of movement control and in using a combined strategy of empirical studies and mathematical modeling.

NOTE

I am grateful to Vincent L. Gracco, Laura L. Koenig and Elliot Saltzman for discussions and comments on earlier versions of this manuscript. This work was supported by Grant DC-00865 from the National Institute on Deafness and Other Communication Disorders.