Biol. Cybern. 72, 43–53 (1994)

Biological Cybernetics
© Springer-Verlag 1994

A neural network model of speech acquisition and motor equivalent speech production

Frank H. Guenther

Boston University, Center for Adaptive Systems and Department of Cognitive and Neural Systems, 111 Cummington Street, Boston, MA 02215, USA

Received: 8 September 1993/Accepted in revised form: 30 May 1994

Abstract. This article describes a neural network model that addresses the acquisition of speaking skills by infants and subsequent motor equivalent production of speech sounds. The model learns two mappings during a babbling phase. A phonetic-to-orosensory mapping specifies a vocal tract target for each speech sound; these targets take the form of *convex regions* in orosensory coordinates defining the shape of the vocal tract. The babbling process wherein these convex region targets are formed explains how an infant can learn phoneme-specific and language-specific limits on acceptable variability of articulator movements. The model also learns an orosensory-to-articulatory mapping wherein cells coding desired movement directions in orosensory space learn articulator movements that achieve these orosensory movement directions. The resulting mapping provides a natural explanation for the formation of coordinative structures. This mapping also makes efficient use of redundancy in the articulator system, thereby providing the model with motor equivalent capabilities. Simulations verify the model's ability to compensate for constraints or perturbations applied to the articulators automatically and without new learning and to explain contextual variability seen in human speech production.

1 Introduction

Speech production is perhaps the most complex motor control task performed by humans. In addition to the amazing rapidity with which words and phonemes are spoken, producing speech sounds requires intricate interactions among information in many different reference frames. These include, but are not limited to, acoustic, somatosensory, and motor frames. Acoustic signals form the actual medium through which speech is communicated; the true job of the speech production mechanism is the creation of an appropriate set of acoustic signals to convey linguistic units from the speaker to listeners. Somatosensory signals from tactile and proprioceptive receptors provide information about the shape of the vocal tract, which determines the sounds being produced. Finally, motor reference frames are used to issue the commands to individual articulators and muscles to produce the movements that result in speech. Normal speech production results from the effortless use of fine-tuned interactions between these very different reference frames. Understanding these frames and their interactions constitutes a difficult task for speech production modelers.

Additional factors further complicate the formulation of a computational model of speech production. First, the interactions between the different reference frames are language-specific. For example, English listeners distinguish between the sounds /r/ and /l/, but Japanese listeners do not. Corresponding differences are seen in the articulator movements of the two groups (Miyawaki et al. 1975). Thus, the precise nature of mappings between acoustic goals and articulator movements depends on the language being spoken. Interactions between reference frames must also be time-varying. As an infant grows, physical characteristics such as the length of the vocal tract and the shapes of articulators change. Temporary or permanent damage to the articulators may also occur. Such changes will affect the acoustic signal that is produced with a given set of motor commands. Maintaining the ability to produce important acoustic features properly thus requires that parameters governing the mappings between acoustic, somatosensory, and motor frames change with time.

The language-specific and time-varying aspects of mappings between reference frames implies that the speech production system must be adaptive; that is, the parameters governing these mappings must be tuned to appropriate values for the infant's native language(s) and must be kept tuned as the infant grows. In infants, babbling comprises an action-perception cycle that can be used to tune the parameters of the production system. Similarly, a complete computational model of speech production should be capable of using an action-perception cycle to tune the parameters governing its performance.

Supported in part by AFOSR F49620-92-J-0499

Speech production is also inherently motor equivalent, i.e., many different motor actions can be used to produce the same speech sound. For example, a speaker may speak normally, using upward and downward movements of the jaw, or he/she can speak with the jaw clenched on a pipe. Production of a given speech sound in these two cases requires a completely different set of articulator positions and movements, yet humans automatically compensate for such constraints (e.g., Abbs and Gracco 1984; Folkins and Abbs 1975; Kelso et al. 1984; Lindblom et al. 1979). Computational models of speech production should also produce such immediate and automatic compensation for perturbations or constraints on the articulators.

Computational modeling of speech production is thus a daunting task. Nonetheless, important computational models have been formulated. The dynamic articulatory model of Henke (1966) represented the first use of computer technology to generate complex movements of model articulators. This model provided an explanation for a wide range of speech production data, and central concepts of the model such as the look-ahead theory of coarticulation are still actively being discussed in the speech production literature (e.g., Boyce et al. 1990; Wood 1991). More recently, Saltzman and Munhall (1989) described the most complete computational model of speech production to date. This *task-dynamic* model, developed at Haskins Laboratories, has been used to explain a wide range of coarticulation and motor equivalence data.

However, these models do not deal with the problem of adaptive organization of their parameters. In fact, MacNeilage and Davis (1990) lament that 'there is at present no unified view of how [speech] motor control develops' due to the lack of attention to speech acquisition in the speech production literature (p. 454).

The remainder of this article describes a computational model that confronts the problem of speech acquisition while providing a unified account of many aspects of speech production in humans. The model is called DIVA because an important aspect of the model is a mapping from Directions (in an orosensory space) Into Velocities of Articulators, and has been briefly introduced in Guenther (1992, 1993). The model is self-organizing; that is, all model parameters are learned via an action-perception cycle, occurring during a babbling phase, rather than handcrafted by the modeler. To this end, the model is formulated as an adaptive neural network. Two learning processes are carried out during babbling: (1) learning of acceptable ranges of orosensory variables for each phoneme, and (2) learning of the redundant mapping between orosensory variables and articulator movements. The learning processes use only information available to an infant (i.e., there are no 'training sets' for the system's mappings as in back-propagation algorithms), and all learning laws governing the model's synapses use only information directly available from the pre- and postsynaptic cells. The self-organizing process is described in detail in Sect. 3. Other issues addressed by the model include the following.

1. The role of orosensory and acoustic feedback in acquisition and production of speech. In DIVA, acoustic feedback is used for acquiring the orosensory targets corresponding to speech sounds, and orosensory feedback (Perkell 1980) is used for both acquisition of speaking skills and normal speech production. This is described in Sect. 2, which gives an overview of the DIVA model.

2. The form of vocal tract 'targets'. Data indicate that the target shape of the vocal tract corresponding to a phoneme is not a specific configuration, but is instead a range of vocal tract configurations that all produce acceptable sounds (e.g., Keating 1990; Lindblom 1983). The DIVA model takes the novel approach of learning regions in orosensory space for each phoneme. These regions, rather than specific configurations of the vocal tract, act as the vocal targets. From a dynamical systems viewpoint, this corresponds to using convex region attractors rather than point attractors (cf. Saltzman and Munhall 1989). These issues are discussed in Sect. 4.

3. Motor equivalence. An appropriate mapping from vocal tract targets to articulator movements is required to achieve automatic compensation for unexpected or unusual conditions. In the task-dynamic model of Saltzman and Munhall (1989), this is accomplished through a complex dynamical system. The complexity of this dynamical system is largely due to the redundant nature of the mapping between vocal tract configurations and articulator positions; that is, many different combinations of articulator positions can be used to produce a single vocal tract configuration. The DIVA model uses a much simpler redundant mapping between desired directions of movement in vocal tract configuration space and velocities of the articulators, detailed in Sect. 5. Furthermore, the parameters of this mapping are learned during babbling, and coordinative structures (e.g., Easton 1972; Fowler 1980; Kelso et al. 1984; Saltzman and Kelso 1987) arise naturally in this learning process.

4. Coarticulation and speaking rate effects. The use of convex region targets in the DIVA model provides a natural explanation for coarticulation and speaking rate effects. These concepts are beyond the scope of the present article and are described in detail elsewhere (Guenther 1994). Although not discussed in the present article, the coarticulation and speaking rate aspects of the model were in place for the model simulations described in Sect. 6.

2 Overview of the DIVA model

A block diagram of the DIVA model is shown in Fig. 1. The model uses two different kinds of neural structure to represent information: vectors and maps. A vector is a set of cells that each code a different dimension in the space being represented (i.e., the input space); the pattern of activity across these cells codes the current position in this space. The term map describes a set of cells wherein each cell codes a small region in the input space. Only

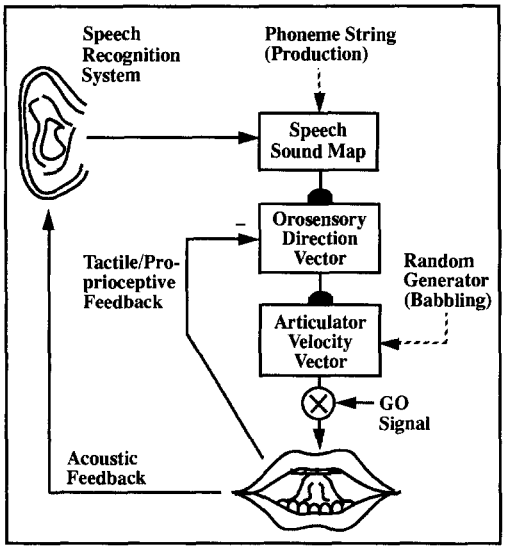

Fig. 1. Overview of the DIVA model. Learned mappings are indicated by *filled semicircles*

one cell can be maximally active in a map, and this cell alone codes the current position in the input space. Both vector and map representations have been widely reported in the neurophysiological literature; see Grobstein (1991) and Penfield and Rasmussen (1950) for examples of these neural structures.

Three main levels of representation are used in the model: a speech sound (phonetic) level, an orosensory (somatosensory) level, and an articulatory level. There are two learned mappings between these levels (shown as filled semicircles in Fig. 1): a phonetic-to-orosensory mapping and an orosensory-to-articulatory mapping. The parameters of these mappings are tuned during the babbling phase described in the next section.

The components of the DIVA model are outlined in the following paragraphs. For clarity of exposition, this discussion will start at the articulator velocity vector block and move clockwise around Fig. 1.

2.1 Articulator velocity vector

The articulator velocity vector (AVV) consists of a set of cells[1] that command movements of the articulators. The activity of each cell is meant to correspond roughly to a commanded contraction of a single muscle or a group of muscles in a fixed synergy. The cells are formed into antagonistic pairs, with each pair corresponding to a different degree of freedom of the articulatory mechanism. Appendix A tabulates the articulatory degrees of freedom used in the model.

[1] Each cell, or neuron, in the model corresponds only loosely to an hypothesized population of neurons in the nervous system; the model should thus be considered as a set of hypothesized stages of neural computation rather than as an attempt to identify specific neurons in the brain

During babbling, AVV cells are randomly activated to produce movements of the articulators. During performance, activation of the AVV cells occurs through the phonetic-to-orosensory and orosensory-to-articulatory mappings.

2.2 GO signal

The GO signal (Bullock and Grossberg 1988) is used to gate the movement commands multiplicatively at the AVV before sending them to the motoneurons controlling the contractile state of the muscles. This signal corresponds to volitional control of movement onset and speed in a human being. Application of the model to speaking rate effects is carried out in Guenther (1994); for the simulations of this paper, a fixed GO signal value of 1 is used.

2.3 Speech recognition system

Infants as young as 1 month of age have been shown to exhibit categorical perception of speech sounds (Eimas et al. 1971). The ability to identify a phoneme in different contexts and across speakers has been demonstrated at 6 months of age (Kuhl 1979), and language-specific categorical perception is evident by 10–12 months of age (Werker and Tees 1984). These abilities are represented by the speech recognition system in the DIVA model. During babbling, this system interprets the infant's speech signal, activating appropriate cells in the speech sound map whenever the infant produces a speech sound from his/her native language. For simplicity, speech sounds in the present implementation are equated to phonemes; the main concepts of the model remain valid, however, for different choices of sound units such as extrinsic allophones or auditory distinctive features.

The process of speech recognition is a very complex one and is beyond the scope of this model. Thus, even though the speech recognition system is conceptualized as interpreting acoustic signals, no acoustic signal is used in the present implementation. Instead, the speech recognition system is implemented as an expert system that looks at key constrictions of the vocal tract to determine which, if any, speech sounds would be produced. If the system recognizes a configuration corresponding to a known speech sound, it activates the corresponding cell in the speech sound map. This activation drives learning in the phonetic-to-orosensory mapping. This corresponds to the assumption that an infant learns when a match occurs between the acoustic effects of his/her own productions and sound categories established by listening to other's productions.

2.4 Speech sound map

Each cell in this map codes a different speech sound. During babbling, cells in the map are inactive except when the speech recognition system determines that the model has produced a speech sound; when this happens, the activity of the corresponding cell in the speech sound map is set to 1. During production, a higher-level brain

center is assumed to sequentially activate the speech sound cells for the desired phoneme string.

2.5 *Orosensory direction vector*

The term *orosensory* was used to describe tactile, proprioceptive, and more complex sensory information about the state of the vocal tract by Perkell (1980), who noted the importance of this kind of information for planning speech movements. This kind of information is key of the DIVA model, both for specifying the targets of speech and for activating appropriate articulator movements to reach these targets.

Cells in the orosensory direction vector (ODV) receive inhibitory tactile and proprioceptive feedback about the state of the vocal tract. The present implementation uses 15 different orosensory dimensions, corresponding to proprioceptive information from individual articulators, tactile information from pressure receptors, and higher-level combinations of information such as the sizes of important constrictions in the vocal tract. A complete list of the orosensory dimensions used in the model is given in Appendix B.

One of the main tasks of the model during babbling is to differentiate between important and unimportant orosensory cues for a sound. To verify that the model can successfully perform this task, the orosensory dimensions used herein correspond to a wide range of available sensory information, including not only important cues about vocal tract shape but also relatively unimportant cues such as the positions of individual articulators (Abbs 1986; Fowler 1990). As discussed in Sect. 4, the model successfully extracts the important information for each speech sound from this very general set of available sensory information. Thus, DIVA relies far less on assumptions about the form of available sensory information than most models of speech production.

ODV cells also receive excitatory input via the learned phonetic-to-orosensory mapping. When a cell in the speech sound map is activated for performance of the corresponding sound, this input to the ODV acts as a target in orosensory space for producing that sound.

During babbling, changes in the configuration of the vocal tract will cause changes in the ODV activities. These changes drive learning in the orosensory-to-articulatory mapping. During performance, the ODV represents the difference between the learned orosensory target for the desired sound and the current configuration; this value thus specifies a desired movement direction in orosensory space that is then mapped into a set of articulator velocities to move the vocal tract in this direction.

3 Acquisition of speaking skills in DIVA

Acquisition of speaking skills in the DIVA model consists of finding appropriate parameters, or synaptic weights, for the phonetic-to-orosensory and orosensory-to-articulatory mappings during a babbling phase. Two different methods of babbling were used for simulations. In the first method, separate babbling stages were used to train the orosensory-to-articulatory mapping and the phonetic-to-orosensory mapping. This corresponds to an early stage of infant learning where the sounds of speech are essentially ignored (and are largely absent) while learning sensory-motor relationships, followed by a stage in which the speech sounds are produced more frequently and drive learning of appropriate orosensory targets for each sound. This is consistent with the stages of babbling commonly reported in infants (e.g., Kaplan and Kaplan 1971; Oller 1980; Sachs 1976; Stark 1980), in which non-speech vocalizations and articulator movements occur well before the onset of frequent speech sounds. Although it is likely that sensory-motor learning starts before the learning of speech sound targets, it is also necessary for this sensory-motor learning to continue when speech sounds become prevalent so that the increasingly complex articulatory movements involved can be learned. In DIVA, the production of speech sounds results in activation of cells in the speech sound map and consequent changes in the ODV; this amounts to the addition of 'noise' to the orosensory-to-articulatory map learning. To verify that proper learning occurs despite this noise, the second learning method involved learning both mappings simultaneously. Since the two methods yield the same major results, only the steps involved in the latter method will be detailed, as follows (refer to Fig. 1):

1. *Randomly activate an AVV.* In DIVA, babbling is produced by superimposing random movements of the speech articulators on an oscillatory movement of the jaw. This corresponds to the phase in infant babbling known as variegated or nonreduplicated babbling (Oller 1980; Stark 1980) which starts at an age of approximately 10 months; this phase has been hypothesized as the stage during which infants learn to produce the various phonemes of their native language (MacNeilage and Davis 1990). With the exception of the AVV cells coding jaw movement, each AVV cell is activated to a value of 1 with probability 0.1; otherwise, its value is 0.
2. *For each of 10 time steps, repeat the following:*
   a. *Carry out learning in the phonetic-to-orosensory mapping.* The synaptic weights in the pathways projecting from a speech sound map cell to the ODV cells represent a vocal tract target for the corresponding speech sound in orosensory space. When the changing vocal tract configuration is identified by the speech recognition system as producing a speech sound during babbling, the appropriate speech sound map cell's activity is set to 1. This gates on learning in the synaptic weights of the phonetic-to-orosensory pathways projecting from that cell [see (2) in Sect. 4], resulting in a modification of the target associated with the active speech sound. This modification is appropriate because it expands the target to include the current configuration of the vocal tract, which is available through orosensory feedback at the ODV cells. See Sect. 4 for details of this learning process.

b. *Carry out learning in the orosensory-to-articulatory mapping.* Random activation of AVV cells produces movements of the articulators which are transmitted through orosensory feedback to the ODV stage, resulting in changes of the ODV cell activities. These changes in activity gate on learning in the synaptic weights of the orosensory-to-articulatory pathways [see (4) in Sect. 5]. If an ODV cell's activity is decreasing, the synaptic weights in pathways projecting from this cell to active AVV cells will increase; in this way, each ODV cell learns a set of articulator movements which will reduce its own activity.

3. Go to (1).

With the model simulation operating approximately in 'real-time' (as evidenced by the speed of articulator movements visible in a computer animation), the entire babbling sequence takes approximately 30 min, during which 5000 random movements of the articulators are carried out. During this time the model learns to produce a set of 29 English phonemes; a complete set of phonemes is not possible due to simplifications in the articulatory structure.

The next two sections motivate and detail the phonetic-to-orosensory and orosensory-to-articulatory mappings.

4 The targets of speech: ranges vs. canonical positions

One of the most active debates in the speech production literature over the past 30 years concerns the nature of the 'targets' as specified for the production mechanism. Henke (1966) posited targets consisting of desired spatial positions of the articulators. The target for a flexible articulator such as the tongue consisted of a series of spatial target positions for small segments of the articulator. MacNeilage (1970) also proposed the control of spatial positions of articulators, suggesting that an articulator's target could be specified as a set of desired muscle lengths. Muscle length targets have been proposed more recently by Cohen et al. (1988).

These spatial and muscle length target models suffer from the same shortcoming: they cannot account for compensatory movements of one articulator when another articulator cannot reach its 'normal' position. For example, Lindblom et al. (1979) showed that subjects immediately compensate for unnatural jaw positions imposed by a bite block when producing vowels, presumably by adjusting the position of the tongue. Compensation was evident even on the first glottal pulse. This eliminates the possibility that acoustic feedback played a role in producing the compensatory movements, but not the possibility that orosensory feedback played a role. Other studies (e.g., Abbs and Gracco 1984; Folkins and Abbs 1975; Kelso et al. 1984) demonstrated similar compensatory actions during lip and jaw perturbations, and the ability to produce intelligible speech with a pipe clenched in one's mouth provides an everyday example of this phenomenon.

Their results led Lindblom et al. (1979) to hypothesize that the targets were not spatial positions of individual articulators, but instead more abstract functions of vocal tract shape that correspond more closely to the speech signal. Specifically, they suggested that 'the target of a vowel segment is coded neurophysiologically in terms of its [vocal tract] area function by means of corresponding sensory information' (p. 157). Similarly, Perkell (1980) proposed that the targets were 'orosensory features' such as proprioceptive and tactile patterns that corresponded directly to distinctive features in the acoustic waveform. More recently, targets in Saltzman and Munhall (1989) are specified in terms of vocal tract variables that define aspects of key constrictions in the vocal tract. These vocal tract variables form a relatively low-dimensional representation of the acoustically important aspects of the vocal tract shape.

A common assumption of these models is that targets correspond to (possibly context-dependent or time-varying) canonical *positions* of articulators or vocal tract variables. There is significant evidence, however, for an alternative hypothesis: The targets of the speech production mechanism are instead *ranges* of articulator positions. For example, English speakers/hearers do not differentiate between velar and palatal stop consonants; as a result, wide anteroposterior variability is seen in the place of constriction for the stop consonants /k/ and /g/ in different vowel contexts (e.g., Daniloff et al. 1980; Kent and Minifie 1977). Kent and Minifie point out that if the target position for /k/ or /g/ is very concrete and positionally well-defined, then the variation cannot be explained by a target position model. Furthermore, if the target positions are only loosely defined, the possibility exists for too much variation that can destroy phonemic identity. Since large anteroposterior variation is seen in /k/ and /g/ but little or no variation is allowable in the vertical position of the tongue body (i.e., the tongue body must contact the palate), it appears that neither a well-defined nor a loosely defined target position will suffice. A more parsimonious explanation is that the tongue body target is an anteroposterior range of positions, and the actual position that is realized depends on contextual influences. This explanation holds also for vowels, where a wider variation of tongue body position is seen along acoustically important dimensions as compared with acoustically less important dimensions (Perkell and Nelson 1985).

More evidence for target ranges rather than positions comes from Keating (1990). Production of vowels in different consonant contexts results in large, but not complete, variability in velum position during the vowel (Kent et al. 1974). For example, if a vowel is produced between two nonnasal consonants as in the word 'dad', the velum remains completely closed throughout the utterance. When a vowel is produced between a nasal and a nonnasal consonant as in the word 'dan', the velum smoothly transitions from closed to open during the vowel. Thus, it would appear that no fixed target velum position is specified for vowels. However, Kent et al. (1974) report that for a vowel between two nasal consonants, a slight but incomplete raising of the velum occurs

during the vowel, followed by a lowering of the velum for the final nasal consonant. It thus appears that the velar target for vowels is a range of positions from maximally closed to largely, but not completely, open.

To explain these data, Keating (1990) hypothesized a 'window theory' of coarticulation wherein the target for each articulator is not a fixed position, but a range of possible positions. When producing a sequence of phonemes, an unspecified procedure might then be used to find an optimal path of the articulator through the sequence of target ranges.

As Fowler (1990) points out, however, in many cases the position of a single articulator may vary because this articulator is used in concert with other articulators to produce a higher-level goal which does not show much variability. For example, Abbs (1986) reports that whereas large variability is seen in lower lip height and jaw height during production of the vowel /a/, the quantity [lower lip height + jaw height] remains relatively constant. Variability is also seen in lower lip and upper lip heights used to produce bilabial closure (e.g., Kelso et al. 1984). In this case, it is insufficient to simply move the articulators to the acceptable ranges for upper lip height and lower lip height; in addition one must insure that the resulting lip aperture is zero. A simple window theory as proposed by Keating (1990) cannot explain these data.

The present work proposes a 'convex region[2] theory' that handles these shortcomings. According to this convex region theory, the target for a speech sound is specified within a high-dimensional orosensory space. This orosensory space includes not only the positions of individual articulators, but also other forms of orosensory information including tactile information from pressure receptors and more complex information corresponding to higher-order combinations of tactile and proprioceptive information such as the degree of constriction at different points along the vocal tract (see Appendix B for a full list of the orosensory variables used in the present implementation). Each dimension of the orosensory target specifies a range of acceptable positions. For example, the target for the vowel /a/ would include relatively large ranges of positions along the orosensory dimensions corresponding to lower lip height and jaw height, but a very small range of positions for the orosensory dimension corresponding to [lower lip height + jaw height].

A very important aspect of this work concerns how the nervous system extracts the appropriate forms of orosensory information that define the different speech sounds. How is it that the nervous system 'knows' that it is lip aperture, and not lower lip height or upper lip height, that is the important articulatory variable for stop consonant production? How does the nervous system know that whereas lip aperture must be strictly controlled for bilabial stops, it can be allowed to vary over a large range for many other speech sounds, including

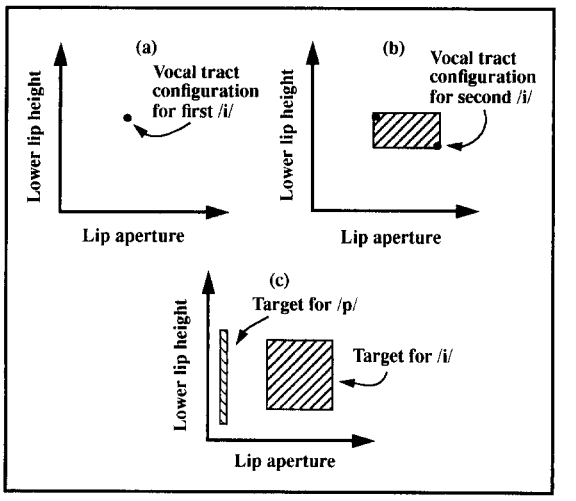

Fig. 2a–c. Learning of the orosensory convex region targets along orosensory dimensions corresponding to lip aperture and lower lip height. **a** The first time /i/ is produced during babbling, the learned target is simply the configuration of the vocal tract when the sound was produced. **b** The second time /i/ is babbled, the convex region target is expanded to encompass both vocal tract configurations used to produce the sound. **c** Convex regions for /i/ and /p/ after many productions of each sound during babbling. Whereas the target for /i/ allows large variation along the dimension of lip aperture, the target for the bilabial stop /p/ requires strict control of this dimension, indicating that the model has learned that lip aperture is an important aspect of /p/ but not /i/

not only vowels but also velar, alveolar, and dental stops? Perhaps even more telling, how does the nervous system of a Japanese speaker know that tongue tip location during production of /r/ can often vary widely, while the nervous system of an English speaker knows to control tongue tip location more strictly when producing /r/ so that /l/ is not produced instead?

The manner in which targets are learned in DIVA provides a unified answer to these questions. Figure 2 schematizes the learning sequence for the vowel /i/ along two dimensions (corresponding to lip aperture and lower lip height) of orosensory space. Every time the model produces the vowel /i/ from any vocal tract configuration during babbling, the speech recognition system activates the cell for /i/ in the speech sound map. The first time that the phoneme is produced during babbling, the corresponding cell in the speech sound map learns the orosensory position that caused the phoneme. This corresponds to a point in orosensory position space, schematized in Fig. 2a. The next time the phoneme is babbled, the speech sound map cell expands its learned target to be a convex region that encompasses both the previous orosensory position and the current orosensory position, as shown in Fig. 2b; this occurs via the simple and biological plausible learning law of (2) below.[3] In this way, the model is

[2] A convex region is a region in space such that for any two points in the region, all points on a line segment connecting these two points are also in the region. A cube is an example of a convex region in three-dimensional space

[3] Note that for reasons of parsimony the present implementation of the model learns hyperrectangles, which are not generally the minimal convex regions that encompass all experienced orosensory positions. A decision as to whether the model must be modified to learn the minimal convex region requires further investigation

constantly expanding its convex region target for /i/ to encompass all of the various vocal tract configurations that can be used to produce /i/.

Now we can address the questions posed above. Consider the regions that result after many instances of producing the vowel /i/ and the bilabial stop /p/ (Fig. 2c). The convex region for /p/ does not vary over the dimension of lip aperture but varies largely over the dimension of lower lip height; this is because all bilabial stops that the model has produced have the same lip aperture, but lower lip height has varied. In other words, the model has learned that bilabial aperture is the important orosensory invariant for producing the bilabial stop /p/. Furthermore, whereas lip aperture is the important orosensory dimension for /p/, the model has learned that this dimension is not very important for /i/, as indicated by the wide range of lip aperture in the target for /i/ in Fig. 2c. Finally, since convex region learning relies on language-specific recognition of phonemes by the infant, the shapes of the resulting regions will vary from language to language.

The mechanism used to learn the convex region targets in DIVA is related to the vector associative map detailed in Gaudiano and Grossberg (1991), and works as follows. The activity of an ODV cell is governed by the following equation:

$$d_i = \sum_j s_j z_{ji} - f_i \tag{1}$$

where d_i is the activity of the ith ODV cell, f_i is the orosensory feedback signal coding position along the ith dimension of orosensory space, s_j is the activity of the jth speech sound map cell, and z_{ji} is the synaptic weight of the pathway from the jth speech sound map cell of the ith ODV cell. The learning law governing modification of the synaptic weight is:

$$\frac{\mathrm{d}}{\mathrm{d}t} z_{ji} = \varepsilon_1 s_j (\alpha_1 z_{ji} - [d_i]^+) \tag{2}$$

where ε_1 and α_1 are learning parameters ($0 < \alpha_1 \ll 1$) and $[x]^+$ is a rectification function such that $[x]^+ = 0$ for $x < 0$ and $[x]^+ = x$ for $x \geqslant 0$. The learning law of (2) ensures that modification of a given phoneme's orosensory target only occurs when that phoneme is being produced. The weights start out large (initialized to 2.0) and primarily decrease with learning; this decrease in the weights corresponds to an increase in the size of the orosensory convex region target.

To see why this is the case, refer to Fig. 3a, which schematizes the mapping from a speech sound map cell to the antagonistic pair coding one dimension of the ODV. The orosensory feedback signal antagonistic pairs (f_{i+}, f_{i-}) each sum to a constant value of 1; this kind of push-pull relationship between cell activities is often found in the nervous system (e.g., Sakata et al. 1980). Assume a large value of ε_1 and a very small value of α_1 in (2). The first time the speech sound corresponding to s_j is produced during babbling, the weight pair (z_{ji+}, z_{ji-}) will converge to the value of (f_{i+}, f_{i-}) when this sound occurs; this is a direct consequence of the learning law defined in

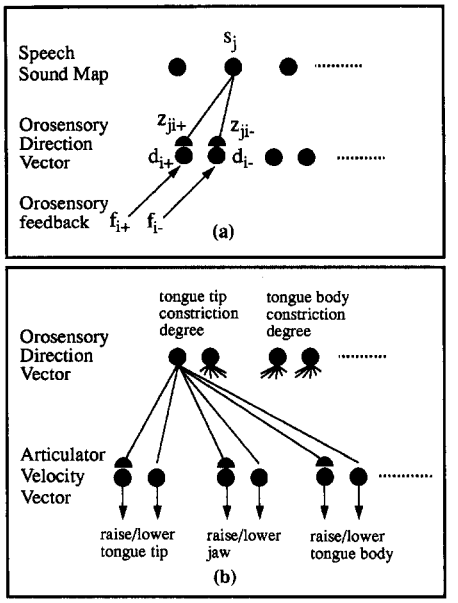

Fig. 3. a Portion of the phonetic-to-orosensory mapping from a speech sound map cell to the antagonistic pair coding one dimension of orosensory space. **b** Schematized view of orosensory-to-articulatory mapping after babbling. Orosensory direction vector (ODV) cells, each coding a desired movement direction in orosensory space, project with large weights to articulator velocity vector (AVV) cells that move the vocal tract in the appropriate direction. Projections to other AVV cells have withered away to zero during learning. Activity at an ODV cell during performance will propagate through the large weighted pathways and activate the corresponding set of articulator movements; this set of articulator movements constitutes a coordinative structure

(2). Assume that this occurred with $(f_{i+}, f_{i-}) = (0.4, 0.6)$. The equation governing AVV cell activities a_i during performance is:

$$a_i = \sum_j [d_j]^+ w_{ji} \tag{3}$$

where w_{ji} are synaptic weights governing the orosensory-to-articulatory mapping. Therefore, during performance, only positive d_i will activate articulator movements. With $(z_{ji+}, z_{ji-}) = (0.4, 0.6)$, from (1) we can see that any value of (f_{i+}, f_{i-}) other than (0.4, 0.6) will derive an articulator movement when s_j is activated to 1. This corresponds to a point attractor or point target at (0.4, 0.6) for (f_{i+}, f_{i-}).

Now consider what happens if the sound corresponding to s_j is produced a second time, with $(f_{i+}, f_{i-}) = (0.5, 0.5)$. Learning will drive the weights (z_{ji+}, z_{ji-}) to (0.4, 0.5). With this weight pair, we see from (1) that a positive d_i will only result if (f_{i+}, f_{i-}) is outside the range $(0.4 \leqslant f_{i+} \leqslant 0.5,\ 0.5 \leqslant f_{i-} \leqslant 0.6)$. This range thus defines a convex region attractor. Further decreases in the weight values will result in further increases in the size of the convex region attractor.

This section has outlined how the DIVA model learns a convex region target in orosensory space for each speech sound and has shown that during performance of the sound, positive activities at the ODV will only arise

when the current vocal tract configuration is outside the convex region. These positive ODV activities code the desired movement direction in orosensory space. The next section describes the mapping which transforms this desired orosensory movement direction into an appropriate set of articulator movements.

5 Mapping orosensory targets into articulator movements

The problem of mapping from orosensory space to articulator space in DIVA is analogous to the inverse kinematics problem of arm movement control: Given a desired three dimensional (3D) spatial trajectory of the hand, calculate an appropriate set of joint angle time courses to move the hand along this trajectory. For speech production, we want to map from a desired trajectory in orosensory space into movements of the speech articulators. To afford motor equivalence, such a mapping must be redundant; that is, it must be possible to reach a given orosensory target with an infinite number of different articulator configurations. A common method for solving the inverse kinematics problem for a redundant manipulator is through the use of the Jacobian pseudoinverse; this kind of solution is used to map from vocal tract variables to articulator variables in the speech production model of Saltzman and Munhall (1989). However, pseudoinverse methods can lead to spurious movements of the articulators well after a vocal tract target has been reached or when no target at all has been specified. Overcoming these problems leads to a very complex dynamical system in Saltzman and Munhall (1989); in fact, Munhall et al. (1991) state that this computational complexity 'should encourage us to seek alternative solutions' to the problem of articulator movement planning (p. 305). The DIVA model posits a much simpler solution to this redundant mapping problem, and the parameters governing this mapping are learned by the model during babbling, not handcrafted by the modeler.

The DIRECT model of reaching (Bullock et al. 1993, Guenther 1992) proposes a solution to the problem of redundant inverse kinematics that uses a learned mapping between desired movement directions in 3-D space and movement directions in joint space (i.e., joint rotations). Mathematical analysis and simulation results show that the parameters in this mapping can be learned in an action-perception cycle, and that the resulting direction-to-rotation mapping results in immediate, automatic compensation for unpracticed events such as reaches with a blocked joint or reaches using a tool rather than the hand.

Analogously, the DIVA model learns a mapping from directions in orosensory space to movement directions in articulator space (i.e., articulator velocities). Learning of the orosensory-to-articulatory mapping occurs as follows. Randomly activated AVV cells cause movements of the speech articulators which are reflected through orosensory feedback as changes in activity of the ODV cells. It is these changes in ODV activity, rather than the magnitude of the activity, that drives learning in the orosensory-to-articulatory pathways according to the following equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} w_{ji} = \varepsilon_2 a_i \left(-\alpha_2 w_{ji} - \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{d}_i \right) \tag{4}$$

where ε_2 and α_2 are learning parameters ($0 < \alpha_2 \ll 1$). Thus, a decrease in an ODV cell's activity results in an increase in the weights projecting from the ODV cell to active AVV cells; these AVV cells are responsible for the movements that resulted in the initial decrease of ODV activity. In this way, each ODV cell learns a set of articulator velocities that cause movements to decrease the ODV cell's activity, i.e. movements that move the vocal tract in the desired direction. The resulting mapping requires only $N \times M$ parameters (synaptic weights), where N is the number of ODV cells and M is the number of AVV cells, and learning of a complete set of parameters can occur very rapidly, minimally requiring only one random activation of each AVV cell (i.e., M total learning trials).

The orosensory-to-articulatory mapping in DIVA is closely related to the coordinative structure modeling concept (e.g., Easton 1972; Fowler 1980; Kelso et al. 1984; Saltzman and Kelso 1987). A coordinative structure is a task-specific grouping of articulators; such groupings arise naturally in the DIVA self-organizing process. Figure 3b schematizes the results after babbling for the ODV cell coding an increase in tongue tip constriction degree. This cell now projects through large weights to AVV cells that raise the tongue tip, the jaw, and the tongue body; the weights for projections to other AVV cells have withered to zero. During performance, a positive activity at this ODV cell will arise when the 'task' is to increase tongue tip constriction degree, as for a dental stop. This positive activity will propagate through the pathways with large weights [see (3)], resulting in the simultaneous raising of the tongue tip, tongue body, and jaw; this task-specific grouping of articulators constitutes a coordinative structure. Furthermore, if one of these three movements is blocked (e.g., a bite block could be used to prevent jaw movement), the other movements continue to decrease tongue tip constriction degree, resulting in the automatic compensation demonstrated in the model simulations of Sect. 6.

Investigation of (3) reveals two additional important properties of the orosensory-to-articulatory mapping. First, nonzero a_i activities can only occur during performance when there are positive d_j, i.e., when the current vocal tract configuration is outside the convex region target. Thus, no spurious movements of the articulators can occur after the vocal tract target has been reached (cf. pseudoinverse methods discussed earlier). Second, a_i (articulator velocity) varies directly with d_j (distance from the target);[4] such a direct variation of articulator velocity

[4] The present implementation of the model simply assumes that the actual articulator velocity is equal to a_i; this is unrealistic in that it can result in infinite acceleration. The direct relationship between articulator velocity and distance from the target will hold, however, for more realistic relationships between commanded and actual velocity

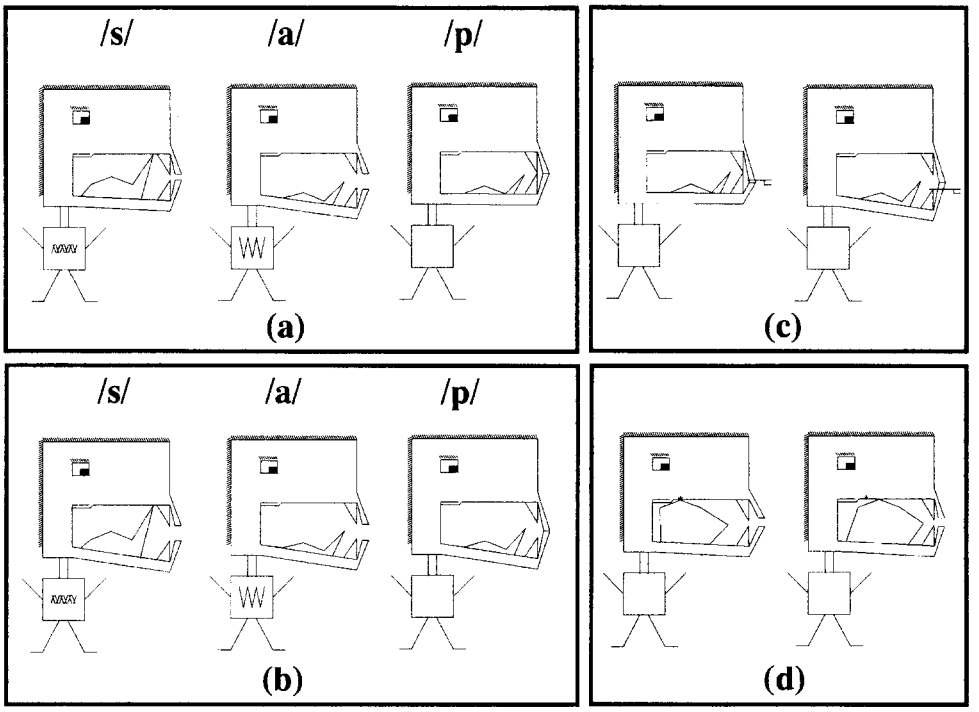

Fig. 4a–d. Simulation results. **a** Production of 'sap' under normal conditions. **b** Production of 'sap' with jaw fixed. **c** Production of /p/ in 'sap' with lip and jaw perturbations. **d** Production of /k/ in 'luke' and 'leak'

with movement distance has been widely reported (e.g., Kozhevnikov and Chistovich 1965; MacNeilage 1970; Sussman and Smith 1971) and is credited with producing nearly constant movement durations.

6 Model performance

After the babbling phase, arbitrary phoneme strings can be specified to the model for production. The model simulation produces an animation sequence showing the movements of the articulators as the string is being produced. Production of a phoneme string occurs as follows:

1. Activate the speech sound map cell corresponding to the next phoneme to be produced.
2. Repeat the following for each time step until the orosensory target for this phoneme is reached:
   a. Update the ODV based on the current vocal tract configuration.
   b. Map this into an AVV and update articulator positions accordingly.
3. If more phonemes remain, deactivate the speech sound map cell and go to step (1).

Several variations of this process were studied. In some simulations the convex region targets were modified based on future phonemes in the string to investigate coarticulation data. Other simulations modulated the size of the targets based on speaking rate to investigate speaking rate data. These variations are discussed in detail in Guenther (1994). Finally, some simulations used the speech recognition system rather than orosensory feedback to the ODV stage to detect phoneme completion. The major results reported here hold for all variations of the simulations.

Figure 4a shows three frames of the animation corresponding to the word 'sap'. Figure 4b shows three frames of 'sap' with the jaw fixed; this simulates bite-block experiments such as those of Lindblom et al. (1979). Despite the removal of the articulatory degree of freedom corresponding to the jaw, the model successfully reaches the orosensory configurations necessary for producing the phoneme string. This is particularly evident for the tongue tip position of the fricative /s/ and the lip closure of the bilabial stop /p/. Figure 4c shows the model producing the /p/ of 'sap' despite application of a bottom lip perturbation (left side) or a jaw perturbation (right side) during upward movement of the bottom lip. This simulates perturbation experiments such as those performed by Folkins and Abbs (1975), Abbs and Gracco (1984), and Kelso et al. (1984). Finally, Fig. 4d shows the model producing the velar stop /k/ in 'luke' (left side) and 'leak' (right side). the ' + ' marks the front-back position of the stop for 'luke'. Comparison of the stop location during 'leak' reveals the anteroposterior variation reported for human subjects when producing these words (e.g., Daniloff et al. 1980; Kent and Minifie 1977). Variability results in DIVA because the vocal tract configuration for /k/ moves to the closest point on the convex region target. When the preceding phoneme is a back vowel such as /u/, this results in a relatively posterior stop location, and when it is a front vowel such as /i/, this results in a relatively anterior stop location. Thus, the model reproduces the 'economy of effort' seen in human speech (Lindblom 1983) by moving from the vocal tract configuration for the vowel to the closest acceptable configuration for the sound /k/.

The simulation results presented here verify the model's ability to perform speech in a motor equivalent matter. All compensation for perturbations and constrained articulators occurs automatically, with no learning required under the constrained conditions.

7 Concluding remarks

This article has shown that study of the process by which infants learn to control their speech articulators can lead to many important theoretical contributions to the ongoing process of understanding speech production. By addressing the question of how the nervous system learns which orosensory information is important for a particular speech sound, a new convex region theory of the targets of speech was formulated. This theory generalizes and extends the window theory of coarticulation posited by Keating (1990), addressing shortcomings pointed out by Fowler (1990) and Keating herself, who offered no procedure for constructing articulator paths through window targets. The present article showed how the convex region theory explains data on variability in speech production; a detailed description of how the theory provides natural explanations for data on coarticulation and speaking rate effects is given in Guenther (1994). Investigating how an infant can learn a mapping from desired movement trajectories formulated in a sensory coordinate frame into the motor coordinate frame of articulator movements led to a simplified solution to the inverse kinematics problem for a redundant system. This solution provides a natural explanation for the formation of coordinative structures, and simulations verified motor equivalent properties seen in human speech such as automatic compensation for articulator constraints and perturbations.

Finally, the model as posited here does not address many important issues concerning the control of timing in speech production (e.g., Fowler 1980). Future work on the model will include an investigation of these timing issues as well as the incorporation of true acoustic information into the action-perception cycle.

Appendix

A. Articulatory degrees of freedom

The following articulatory degrees of freedom are presently used in the DIVA model:

1. Raise/lower jaw
2. Raise/lower tongue body with respect to jaw
3. Raise/lower tongue tip with respect to tongue body
4. Raise/lower upper lip
5. Raise/lower bottom lip with respect to jaw
6. Raise/lower velum
7. Forward/backward extension of tongue body with respect to jaw
8. Forward/backward extension of tongue tip with respect to tongue body
9. Forward/backward extension of both lips simultaneously

B. Orosensory dimensions

The following orosensory dimensions are presently used in the DIVA model. Several of these dimensions are closely related to the vocal tract variables of Saltzman and Munhall (1989).

1. Jaw height with respect to skull
2. Tongue body horizontal position with respect to skull
3. Tongue body height with respect to jaw
4. Tongue body height with respect to skull
5. Tongue body pressure receptors
6. Tongue tip horizontal position with respect to skull
7. Tongue tip height with respect to tongue body
8. Tongue tip height with respect to skull
9. Tongue tip pressure receptors
10. Lip protrusion
11. Lip aperture
12. Lower lip height with respect to jaw
13. Lower lip pressure receptors
14. Upper lip height with respect to skull
15. Upper lip pressure receptors

References

Abbs JH (1986) Invariance and variability in speech production: a distinction between linguistic intent and its neuromotor implementation. In: Perkell JS, Klatt DH (eds) Invariance and variability in speech processes. Erlbaum, Hillsdale, pp 202–219

Abbs JH, Gracco VL (1984) Control of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech. J Neurophysiol 51:705–723

Boyce SE, Krakow RA, Bell-Berti F, Gelfer CE (1990) Converging sources of evidence for dissecting articulatory movements into core gestures. J Phonetics 18:173–188

Bullock D, Grossberg S (1988) Neural dynamics of planned arm movements: emergent invariants and speed-accuracy properties during trajectory formation. Psychol Rev 95:49–90

Bullock D, Grossberg S, Guenther FH (1993) A self-organizing neural network model for redundant sensory-motor control, motor equivalence, and tool use. J Cogn Neurosci 5:408–435

Cohen MA, Grossberg S, Stork DG (1988) Speech perception and production by a self-organizing neural network. In: Lee YC (ed) Evolution, learning, cognition, and advanced architectures. World Scientific Publishers, Hong Kong

Daniloff R, Schuckers G, Feth L (1980) The physiology of speech and hearing: an introduction. Prentice-Hall, Englewood Cliffs

Easton TA (1972) On the normal use of reflexes. Am Sci 60:591–599

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J (1971) Speech perception in infants. Science 171:303–306

Folkins JW, Abbs JH (1975) Lip and jaw motor control during speech: responses to resistive loading of the jaw. J Speech Hearing Res 18:207–220

Fowler CA (1980) Coarticulation and theories of extrinsic timing. J Phonetics 8:113–133

Fowler CA (1990) Some regularities of speech are not consequences of formal rules: comments on Keating's paper. In: Kingston J, Beckman ME (eds) Papers in laboratory phonology. I. Between the grammar and physics of speech. Cambridge University Press, Cambridge, UK, pp, 476–487

Gaudiano P, Grossberg S (1991) Vector associative maps: Unsupervised real-time error-based learning and control of movement trajectories. Neural Networks 4:147–183

Grobstein P (1991) Directed movement in the frog: a closer look at a central representation of spatial location. In: Arbib MA, Ewert JP (eds) Visual structures and integrated functions. Springer, Berlin Heidelberg New York, pp 125–138

Guenther FH (1992) Neural models of adaptive sensory-motor control for flexible reaching and speaking. PhD dissertation, Boston University

Guenther FH (1993) A self-organizing neural model for motor equivalent phoneme production. In: Proceedings of the World Congress on Neural Networks, Portland. Erlbaum, Hillsdale, pp III-6–9

Guenther FH (1994) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Boston University Center for Adaptive Systems Technical Report CAS/CNS-94–012

Henke WL (1966) Dynamic articulatory model of speech production using computer simulation. PhD dissertation, Massachusetts Institute of Technology

Kaplan E, Kaplan G (1971) The prelinguistic child. In: Eliot J (ed) Human development and cognitive processes. Holt, Rinehart, and Winston, New York, pp 358–381

Keating PA (1990) The window model of coarticulation: articulatory evidence. In: Kingston J, Beckman ME (eds) Papers in laboratory phonology. I. Between the grammar and physics of speech. Cambridge University Press, Cambridge, UK, pp 451–570

Kelso JAS, Tuller B, Vatikiotis-Bateson E, Fowler CA (1984) Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. J Exp Psychol Hum Percep Perform 10:812–832

Kent RD, Minifie FD (1977) Coarticulation in recent speech production models. J Phonetics 5:115–133

Kent RD, Carney P, Severeid L (1974) Velar movement and timing: evaluation of a model for binary control. J Speech Hearing Res 17:470–488

Kozhevnikov VA, Chistovich LA (1965) Speech: articulation and perception. Translation by Joint Publications Research Service. Washington DC (JPRS 30543)

Kuhl PK (1979) Speech perception in early infancy: perceptual constancy for spectrally dissimilar vowel categories. J Acoust Soc Am 66:1668–1679

Lindblom B (1983) Economy of speech gestures. In: MacNeilage PF (ed) The production of speech. Springer, Berlin Heidelberg New York, pp 217–245

Lindblom B, Lubker J, Gay T (1979) Format frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. J Phonetics 7:147–161

MacNeilage PF (1970) Motor control of serial ordering in speech. Psychol Rev 77:182–196

MacNeilage PF, Davis B (1990) Acquisition of speech production: frames, then content. In: Jeannerod M (ed) Attention and performance. XIII. Motor representation and control. Erlbaum, Hillsdale, pp 453–576

Miyawaki K, Strange W, Verbrugge R, Liberman AM, Jenkins JJ, Fujimura O (1975) An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English. Percept Psychophys 18:331–340

Munhall KG, Ostry DJ, Flanagan JR (1991) Coordinate spaces in speech planning. J Phonetics 19:293–307

Oller DK (1980) The emergence of the sounds of speech in infancy. In: Yeni-Komshian GH, Kavanagh JF, Ferguson CA (eds) Child phonology, Vol. 1. production. Academic Press, New York, pp 93–112

Penfield W, Rasmussen T (1950) The cerebral cortex of man: a clinical study of localization and function. MacMillan, New York

Perkell JS (1980) Phonetic features and the physiology of speech production. In: Butterworth B (ed) Language production, Vol 1. Speech and talk. Academic Press, New York, pp 337–372

Perkell JS, Nelson WL (1985) Variability in production of the vowels /i/ and /a/. J Acoust Soc Am 77:1889–1895

Sachs J (1976) The developments of speech. In: Carterette EC, Friedman MP (eds) Handbook of perception, Vol VIL. Languange and speech. Academic Press, New York, pp 145–172

Sakata H, Shibutani H, Kawano K (1980) Spatial properties of visual fixation neurons in posterior parietal association cortex of the monkey. J Neurophysiol 43:1654–1672

Saltzman EL, Kelso JAS (1987) Skilled actions: a task-dynamic approach. Psychol Rev 94:84–106

Saltzman EL, Munhall KG (1989) A dynamical approach to gestural patterning in speech production. Ecol Psychol 1:333–382

Stark RE (1980) Stages of speech development in the first year of life. In: Yeni-Komshian GH, Kavanagh JF, Ferguson CA (eds) Child phonology, Vol 1. production. Academic Press, New York, pp 73–92

Sussman HM, Smith JU (1971) Jaw movements under delayed auditory feedback. J Acoust Soc Am 50:685–691

Werker JF, Tees RC (1984) Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant Behav Develop 7:49–63

Wood SAJ (1991) X-ray data on the temporal coordination of speech gestures. J Phonetics 19:281–292